APPLICATIONS OF AI AND MACHINE LEARNING

Editor-in-Chief

Dr. Raman Maini

Editors

Dr. Nirvair Neeru, Dr. Navdeep Kanwal, Dr. Abhinav Bhandari
Er. Gaurav Deep, Dr. Dhavleesh Rattan, Dr. Williamjit Singh, Dr. Navjot Kaur

ISBN: 978-93-93579-09-6

CONTENTS

VIDEO SEGMENTATION TECHNIQUES-A REVIEW ON ITS APPLICATION AND RECENT ADVANCEMENTS

Yadwinder Singh[*1], Lakhwinder Kaur[*2], Nirvair Neeru[*3]
[*]*Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India - 147001*
[1]yadwinder.singh.shergill@gmail.com
[2]mahal2k8@yahoomail.com
[3]nirvair.ce@pbi.ac.in

ABSTRACT- Digital Video Processing have been found in different realistic applications such as military, remote sensing, criminology and so on. Locating and detecting the objects is a tedious task. Video segmentation is one of the core steps in video processing system. A large number of models and recognition patterns are suggested by other researchers. However, there is a need for an accurate segmentation process. This paper is a review of different segmentation techniques on video data stated by researchers. Video entities needs to be modelled by using spatio-temporal relations. In general, segmentation operation done in regions as well as edges, accurate information has to be presented. Features combination and its representation defines the success rate of segmentation techniques. This review analysis will help the upcoming researchers to gain insight about the deep learning concepts in video segmentation process.

INDEX TERMS- Digital video, Segmentation, Feature representation, Video entities and Deep learning

I. INTRODUCTION

In recent days, the growth of digital video is increasing due to its important role in entertainment, multimedia and education systems. There is a demand for an efficient analysis of the browsing and retrieving the video data. In general, video is the composition of different units like shots, scenes and sequences which are aligned in some logical structure [1]. Extracting relevant and required information from an organized video data is always a cumbersome tasks. Henceforth, digital video analysis is one of the recent research area that deals with the extraction of knowledgeable information via video frames. It composes of both low-level and high-level information processing units. Video analysis is classified into two aspects, a) Extracting the regions information using statistical approaches i.e. to acquire insights about pixels and its intensities in a frames and b) Extracting the structural information i.e. to find the relationships between pixels and regions. Most of the researchers has studied more about the low-level image analysis [2]. Likewise, regions segmentation offers numerous benefits than the edge segmentation process.

Segmentation [3] is one of the core steps that presents the numerous research scope in low-level image analysis. In application wise, it is greatly helpful for decision making systems such as medical, robotics, remote sensing, satellite communication and animations. Segmentation is the process of splitting the regions into intersecting and non-intersecting, so as to yield the right information at the right time based on defined research objectives. Several researchers have suggested different and precise segmentation solutions [4], yet, it somehow lacks to serve all kinds of video scenarios. In addition, techniques developed for a class can't serve for other sets of classes. Extracting visual information mostly relies on semantic contents and domain knowledge. Then, modelling of spatial- temporal relations [5] from different regions is a tedious task which requires an efficient segmentation process.

The contributions of this study are:

a. Collects a well-reputed journals related to the review objectives.
b. Reviews the prior segmentation techniques by stating its merits and demerits.
c. Compares the suggested segmentation techniques in terms of objectives, datasets, and numerical analysis, so as to easily understand the current state of video segmentation techniques.
d. Presents the research challenges in this research domain.

The remaining sections of this paper is arranged as: Reviews of existing studies in Section II; Section III presents the comparative analysis and finally, concluded with research challenges in Section IV.

II. REVIEW OF EXISTING STUDIES

This section reviews the existing studies that have been developed using deep learning for segmenting the videos. In [6], they presented an iterative multi-path tracking with sparse point supervision for video segmentation process. Image annotations is a challenging task due to its higher time and monetary costs. Here, graph based optimization framework was used for aligning the locations during its segmentation process. Finally, k-shortest path algorithm was iteratively applied for classifying the objects, even under different modalities. Gaze dataset is used for experimental purpose. System has classified the objects with 40% outliers detection. System has drastically improved the morphological and edge costs transformation. In [7], sequential temporal video was taken for enhancing the segmentation accuracy, even in homogeneous consecutive frames. Boundary detection in shot scenes using conventional segmentation has lowered the edge deformation.

Semantic segmentation [8] is one of the research areas that intensely performed for temporal data. It has improved the spatial connectivities for each frame of videos. Then, a convolutional gated recurrent networks was employed for each recurrent part of the spatial associativities. Experimental analysis conducted on Camvid datasets, 0.812 (Synthia data) & 0.871 (ARDone). Here, optimizer takes higher storage space and thus training the hidden layer is difficult. Author in [9]

explored an end-to-end framework for better visual representation of object recognition systems. Initially, the fisher vectors were used for extracting the temporal models and then Gaussian mixtures has parsed the motion units into action units. Higher dimensions have not resolved the overfitting issues. Though, system accuracy has improved, then the labelling imbalance occurs for certain action units. Thus, author in [10] has suggested a deep bi-directional Long Short Term Memory (LSTM) with Convolutional Neural Networks (CNN) features for different video sequences. Here, every 6th frames in videos are used for taking deep features which has reduced the data redundancy. Then, LSTM network is formed for all long sequences of videos. Finally, three benchmark datasets including UCF-101, YouTube 11 Actions, and HMDB51 were used for experimental purposes. Recognition score on each dataset is presented as 92.84% (youtube), 87.64%(HMDB51) and 91.21% (UCF101). Learning long term sequences which are complex in nature is not handled properly.

Frame level CNN was suggested by [11] via LSTM for sequential videos analysis. In some cases, rich motion information were taken for extracting the salient features. It was experimented in UCF101 and HMDB -51 datasets and achieved an accuracy of 84.7% than prior algorithms. Some salient features failed to administer the long term relationships with other features. Video segmentation from multiple views can help for reduced data redundancy and the overlapping of data [12]. Deep neural networks is employed to combine the deep features in two tier frames. Then, the lookup operation is done for targeted classes using deep bi- directional long short term memory. The suggested techniques is applied on real time datasets, YouTube that has explored the precision (0.93), recall (0.860), f1-score (0.90) and events recall (0.87). Since cloud based architecture is used, the dominance of some information is not trained properly. In [13], objects are detected using tubelets -CNN which efficiently operated in segmentation tasks. Performing segmentation in temporal data pattern is highly risky process due to its dynamic nature. Here, boundaries are randomly perturbed for uniform distribution. It was implemented in ImageNet VID and YTO datasets that explored average accuracy of 76.8%.

In [14], the authors improved the background subtraction models using deep convolutional neural networks. Background objects detection is a complex task in video sequences. It was implemented in CDnet 2014, and wallflower datasets. F-measure is one of the performance parameters, in which, different datasets obtained, for CDNET 2014 (0.7548) and wallflower (0.7512). System has not focussed dynamic background objects due to time restrictions. Two stream-CNN was suggested to refine the video segmentation process, so as to achieve efficient global appearance. Initially, pixels levels based features are extracted and then foreground likelihood maps are computed. Finally, fusion networks are computed by least square regression methods. Experimental analysis on, Densely Annotated Video Segmentation (DAVIS) dataset achieved accuracy of 74.11% and 57.5% accuracy was given by SegTrack V2. Expressing the dilated foreground regions is not coded for certain color features which increased the error rate.

III. COMPARATIVE ANALYSIS

Finally, a comparative table is developed on the basis of objectives, techniques, datasets used, numerical results and drawbacks of the systems.

TABLE 1
COMPARISON OF DIFFERENT TECHNIQUES

Ref. No	Objectives	Techniques & datasets used	Numerical results	Drawbacks
[1]	To learn the video objects from static images.	Convolutional Neural Networks (CNN) Datasets: Extended Complex Scene Saliency Dataset (ECSSD); MSRA10K; Saliency Object Database (SOD); PASCAL-S.	Quality of the video frames were gradually increased. Different bounding box operations were used for improving the object quality. 10% annotated frames, 0.064 mIoU (Intersection of Union).	Low resolution objects are not applied.
[2]	To determine the objects from scenes of video data.	Frame differencing and frame intersection method.	System has increased the accuracy of the video segmentation process and also decreases the computational complexity.	Clustering on heterogeneous data sources are not discussed.
[3]	To explore high spatial-temporal video scenes.	Feed forward Convolutional Neural Networks. Dataset: DAVIS 2017; YouTube objects & SegTrack V2.	System has achieved object detection accuracy of 0.842(DAVIS); 0.784(YouTube) and 0.771 (SegTrackV2)	Some hidden layers avoided the complex patterns.
[5]	To develop a segmentation models on temporal data.	Convolutional Neural Networks & Recurrent Neural Networks. Dataset: Segtrack V2 and Davis	System has achieved 0.96(precision); 0.877 (recall) & 0.916 (f-measure)	Multiple layers are not used in recurrent units and also degraded temporal data representation.

[16]	To detect the saliency objects from video footages by improving visual qualities in the aspects of Spatial Dynamic Attention (SDA) and Temporal Dynamic Attention (TDA).	3D Convolutional Neural networks; *Human eye gazing database:* VAGBA datasets, Lübeck INB Dataset, IVB dataset, Collaborative Research in Computational Neuroscience – Data Sharing (CRCNS) dataset and Dynamic Image and eye movement.	Receiver Operation Characteristics (ROC) with area-under-curve (AUC) and Precision versus Recall (P-R) Curve were analyzed. Earth Mover's distance analysis was done on all datasets, and our methods depicts better results than prior methods. The results stated as, IVB (0.4988); CRCNS (0.6508); DIEM (0.6203); INB (0.3751); VAGBA (0.4013). Likewise, histogram analysis has estimated as IVB (0.7534); CRCNS (0.5773); DIEM (0.6121); INB (0.8371); VAGBA (0.8132).	a) When input data increases, higher training time is also taken. High level features are ignored to train under Convolutional Neural Network (CNN). b) Some inaccurate motion features are eliminated to develop the reliable system.
[17]	a) To differentiate the detection area from clutter background by computing pixel level similarity among the salient objects. b) To reduce the computational load on memory with better representation of selected features.	Pixel level matching based CNN; *Datasets:* DAVIS, SegTrack, Jumpcut and Thermal Road.	DAVIS: Region similarity and contour accuracy was calculated. Post processing results with CNN denoted as PLM, accuracy as 0.70 with stability rate 0.16 and speed rate 0.3s. SegTrack dataset has achieved average, 0.73. JumpCut dataset has achieved average, 9.55. Thermalroad dataset has achieved average, 10.6. All these achieved values are the best values.	a) Some background objects has similar appearance that confused the target classes. b) Thin objects are not handled properly during pre-training process.
[18]	To effectively classify the YouTube videos by extracting the local spatio-temporal information patterns.	Feature connectivity based Convolutional Neural Network (CNNs). *Dataset:* Sports 1M-datasets that composes of Aquatic Sports, Team Sports, Winter Sports, Ball Sports, Combat Sports, Sports with Animals.	Different frames models were employed and the collective results obtained given as, in average taken from single, early, late and slow in CNN for inputs, clip hit @1 (41.4); video hit @1 (63.9) and Video hit @ 5(82.4).	While labelling training videos, invasion of incorrect annotations has degraded the feature connectivity.
[19]	To develop a computation tool that estimates the discarded fish catch via monitoring CCTV footages.	Here, segmentation is done using N^4 fields algorithm and CNN used for classification purpose. *Datasets* contain 52 videos from 12 conveyor belts.	Relative error (%) is computed, that yielded 7.82% for validation and 7.13% for testing.	Some pixels are mismatching even in probability maps.
[20]	To detect the salient objects in videos using spatial and temporal saliency information.	Novel data augmentation techniques i.e fully convoluted networks. *Datasets:* Freiburg-Berkeley Motion Segmentation (FBMS) dataset, and Densely Annotated Video Segmentation (DAVIS) dataset	FBMS has achieved MAE (7.65%) & DAVIS has achieved MAE (6.36%).	Direct saliency maps is not possible due to lack of dynamic information.

[21]	To present an efficient background subtraction in videos for fine tuning the parameters.	Convolutional Deep Net (CDNet) *Dataset:* ESI	Obtained performance results are 0.9609 recall, 0.9984 specificity, 0.9499 precision, 0.0016 false positive, 0.0391 false negative, and 0.9507 F-measure.	Though, it handled different scenes, some ROI features are distracted.
[22]	To find the target class of the segmented regions, so as to reduce the false positive rate.	Convolutional Neural Network & Recurrent neural networks. *Dataset:* NYU Depth v2.	System has achieved 56.2% class accuracies for 52% average pixels.	a) Ground truth information is ignored in high level pixels in training process. b) Denoising operations is also ignored during learning spatial and temporal relations.
[23]	To find the temporal oriented human actions in realtime applications.	Temp oral Convolutional Networks and Long Short Term Memory (LSTM) based Recurrent Neural Networks. *Datasets:* University of Dundee 50 Salads, MERL shopping, and Georgia Tech Egocentric Activities (GTEA)	System has achieved 73.4 % accuracy on 50 salads (higher) and 64.7% accuracy on 50 salads (mid).	It has yielded better accuracy but the segmentation errors has not focussed.
[24]	To enhance the segmenting the objects for label inconsistencies.	Deep Convolutional neural Networks *Dataset:* Youtube-Object	System has achieved average accuracy of 0.741 (Batch Conditional Random Fields (CRF) and 0. 744 (Comb CRF)	a) Less accuracy found during object boundary detection. b) Some morphological operations are not utilized, when the color and positions of pixels.
[25]	To deeply understand the motions using object segmentation methods.	MoNet that exploited the deep motion.	Semi-supervised system has achieved 84.7% Mean, 96.8% recall and 6.4% decay.	Some adjacent frames are not arranged properly that lowered generalization capability.
[26]	To develop a deep neural networks from temporal and contextual information.	Tubelets Convolutional Neural Networks (T- CNN) *Dataset:* ImageNet VID Dataset, YouTubeObjects (YTO) Dataset	Models has achieved 72.3%(still-image) and 74.5% (+MGR)	a) False negatives rate is high in motion guided propagation and tracking. b) Complex scenes are not considered in false positive.
[27]	To leverage the unlabelled data via improving segmentation techniques.	SeSE-Net, a self-supervised deep learning for segmentation. *Dataset:* CAR, PET & CMR	System has achieved accuracies, CAR – 0.9732 CMR - 0.9852 PET- 0.8713	It is time consuming during unlabelled data analysis.
[28]	To efficiently detect the objects in streaming fashion.	Objects in Video Enabler through Label Propagation (OVERLAP) and Per-frame Convolutional Neural Networks	Per-class detection accuracy and mean average precision are the computed performance metrics. Results obtained are 20.57 (R-CNN) and 37.413 (Fine tune VOP)	Similar appearances of the background objects confused the multiple classes instances.
[29]	To improve the uncertainty of the features via different mask forms.	Unsupervised active learning. *Datasets*: DAVIS16 & Shining 3D dental datasets	System has achieved accuracy of 86.4 % (DAVIS) & 93.5% (Shining 3D)	Since sampling random process were employed, feature diversities are not handled for certain boundaries.
[30]	To improve the coding efficiency of video codec systems for better visual quality.	Convolutional Neural networks.	Bjontegaard Metric (BD)- Data Rate and BD- Peak to Signal Noise Ratio (PSNR) are the performance parameters studied. Flower garden data has achieved 1.25dB (PSNR) and 0.42dB for football.	a) Some blocks edges are distorted in texture motion parameters. b) Different textual contents are not classified properly.

IV. CONCLUSION AND FUTURE RESEARCH DIRECTION

With the recent developments made in the Information Processing Units (IPU), the information is being exponentially increasing from every aspect of the digital world. Along with the technical advancements in image analysis, there is a pressing need for video analysis. Hence, this paper is a review of different video segmentation techniques explored by several researchers. There is an urgent need to access and retrieve the data presented in video in an efficient way, which is known as 'video content analysis'. It composes of two tasks, namely, video segmentation and video retrieval process. Former one, video segmentation' is focussed in our study. The process of splitting the video into knowledgeable data by extracting its spatial and temporal patterns from its video sequences is coined as video segmentation. From the reviews conducted, the research challenges still pertains in this study are discussed. Spatial-temporal discontinuities often lowers the edge information during segmentation process. Combination of features taken for training the classifiers has to be precisely selected due to the issues like noise, motion of objects that distorts the representation of features. While doing frame based information analysis, invasion of foreground and background objects reduces the segmentation accuracy. Recently, deep learning is adopted, so as to devise the above mentioned challenges, yet the spatial-temporal discontinuities is not resolved due to higher time taken for pre-training the classifiers.

REFERENCES

[1]. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B. and Sorkine-Hornung, A., 2017. Learning video object segmentation from static images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2663-2672).

[2]. Kalith, A.S., Mohanapriya, D. and Mahesh, K., 2018. Video Scene Segmentation: A Novel Method to Determine Objects. Int. J. Sci. Res. Sci. Technol, 4, pp.90-94.

[3]. Bao, L., Wu, B. and Liu, W., 2018. CNN in MRF: Video object segmentation via inference in a CNN-based higher-order spatio-temporal MRF. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5977-5986).

[4]. Suresh, K., 2019. Various image segmentation algorithms: A survey. In Smart Intelligent Computing and Applications (pp. 233-239). Springer, Singapore.

[5]. Valipour, S., Siam, M., Jagersand, M. and Ray, N., 2017, March. Recurrent fully convolutional networks for video segmentation. In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 29-36). IEEE.

[6]. Laurent Lejeune, Jan Grossrieder, Raphael Sznitman, 2018. Iterative multi-path tracking for video and volume segmentation with sparse point supervision. Medical Image Analysis.doi: https://doi.org/10.1016/j.media.2018.08.007

[7]. Mashtalir, S. and Mashtalir, V., 2016, August. Sequential temporal video segmentation via spatial image partitions. In 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP) (pp. 239-242). IEEE.

[8]. Siam, M., Valipour, S., Jagersand, M. and Ray, N., 2017, September. Convolutional gated recurrent networks for video segmentation. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 3090-3094). IEEE.

[9]. Kuehne, H., Gall, J. and Serre, T., 2016, March. An end-to-end generative framework for video segmentation and recognition. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1-8). IEEE.

[10]. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M. and Baik, S.W., 2017. Action recognition in video sequences using deep bi-directional LSTM with CNN features. IEEE Access, 6, pp.1155-1166.

[11]. Wang, X., Gao, L., Song, J. and Shen, H., 2016. Beyond frame-level CNN: saliency-aware 3-D CNN with LSTM for video action recognition. IEEE Signal Processing Letters, 24(4), pp.510-514.

[12]. Hussain, T., Muhammad, K., Ullah, A., Cao, Z., Baik, S.W. and de Albuquerque, V.H.C., 2019. Cloud-assisted multi-view video summarization using CNN and bi-directional LSTM. IEEE Transactions on Industrial Informatics.

[13]. Kang, K., Ouyang, W., Li, H. and Wang, X., 2016. Object detection from video tubelets with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 817-825).

[14]. Babaee, M., Dinh, D.T. and Rigoll, G., 2018. A deep convolutional neural network for video sequence background subtraction. Pattern Recognition, 76, pp.635-649.

[15]. Caelles, S., Montes, A., Maninis, K.K., Chen, Y., Van Gool, L., Perazzi, F. and Pont-Tuset, J., 2018. The 2018 davis challenge on video object segmentation. arXiv preprint arXiv:1803.00557.

[16]. Wang, Z., Ren, J., Zhang, D., Sun, M. and Jiang, J., 2018. A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing, 287, pp.68-83.

[17]. Shin Yoon, J., Rameau, F., Kim, J., Lee, S., Shin, S. and So Kweon, I., 2017. Pixel-level matching for video object segmentation using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2167-2176).

[18]. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).

[19]. French, G., Fisher, M.H., Mackiewicz, M. and Needle, C., 2015. Convolutional neural networks for counting fish in fisheries surveillance video. Proceedings of the machine vision of animals and their behaviour (MVAB), pp.7-1.

[20]. Wang, W., Shen, J. and Shao, L., 2017. Video salient object detection via fully convolutional networks. IEEE Transactions on Image Processing, 27(1), pp.38-49.

[21]. Sakkos, D., Liu, H., Han, J. and Shao, L., 2018. End-to-end video background subtraction with 3d convolutional neural networks. Multimedia Tools and Applications, 77(17), pp.23023-23041.

[22]. Pavel, Mircea Serban, Hannes Schulz, and Sven Behnke. "Object class segmentation of RGB-D video using recurrent convolutional neural networks." Neural Networks 88 (2017): 105-113.

[23]. Lea, C., Flynn, M.D., Vidal, R., Reiter, A. and Hager, G.D., 2017. Temporal convolutional networks for action segmentation and detection. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 156-165).

[24]. Seong-Jin Park, Ki-Sang Hong, 2018. Video Semantic Object Segmentation by Self-Adaptation of DCNN. Pattern recognition letters (pp.1-15)

[25]. Xiao, H., Feng, J., Lin, G., Liu, Y. and Zhang, M., 2018. Monet: Deep motion exploitation for video object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1140-1148).

[26]. Kang, K., Li, H., Yan, J., Zeng, X., Yang, B., Xiao, T., Zhang, C., Wang, Z., Wang, R., Wang, X. and Ouyang, W., 2017. T-cnn: Tubelets with convolutional neural networks for object detection from videos. IEEE Transactions on Circuits and Systems for Video Technology, 28(10), pp.2896-2907.

[27]. Oh, S.W., Lee, J.Y., Xu, N. and Kim, S.J., 2018. Fast user-guided video object segmentation by deep networks. In The 2018 DAVIS Challenge on Video Object Segmentation-CVPR Workshops (Vol. 3).

[28]. Tripathi, S., Belongie, S., Hwang, Y. and Nguyen, T., 2016, March. Detecting temporally consistent objects in videos through object class label propagation. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1-9).

[29]. Yan Tian, Guohua Cheng, Judith Gelernter, Shihao Yu, Chao Song, Bailin Yang., 2019. Joint temporal context exploitation and active learning for video segmentation. Pattern recognition. (pp. 2-8)

[30]. Fu, C., Chen, D., Delp, E., Liu, Z. and Zhu, F., 2018. Texture segmentation based video compression using convolutional neural networks. Electronic Imaging, vol. 2, pp.16.

ACCESSIBILITY EVALUATION OF MOBILE PHONE FOR HIGHER EDUCATION APPLICATIONS

Vishal Gupta[#1], Hardeep Singh[*2]
Department of Computer Science, Guru Nanak Dev University
[1]vishalgupta@khalsacollege.edu.in
[2]hardeepcse.singh@gmail.com

ABSTRACT— Due to the high ratio of population education is a major concern for every individual. E-learning is another mode of education that provides zero-cost education to each individual. More than 15% population is disabled in any way, which includes 93 million children and 720 million adults. Everyone has the right to free education. Nowadays there exist a high number of free applications for mobile devices that provide information topic wise and course wise. Unfortunately, majority of mobile phone applications are inaccessible. This research evaluates the mobile applications accessibility using Accessibility scanner tool by Google. In this study, most used 10 e-learning mobile applications were evaluated and checked weather they comply with mobile accessibility guidelines of WCAG 2.1.

KEYWORDS— Evaluation; Accessibility; Mobile Applications; Higher Education; WCAG 2.1

INTRODUCTION

This According to statistics the present number of smartphone device users in the world today is more than 3.8 billion, and this means 48.53% of worldwide population owns a smart tablet or a smartphone [1]. These days online platform for study turns your dreams into reality and helps you to find a better job in each and every sector. Every adult and child is unique, and has specific learning needs. ICT in education brings a lot of advantage in our social and education life. The computer, laptops, and smartphones will enhance autonomous access of students to their education. ICT also encourages learners who seek special education to carry out work at their own speed. Many mobile apps are becoming the valuable part of our daily lives; the various types of valuable apps are for health, travel, map, finance, retail, music, productivity, entertainment, education etc. There are several smartphone apps that help us with our everyday activities, but not many of them are accessible, this indicates that a high number of adults and children with some special needs cannot easily interact and access to these mobile applications.

In this study, we present accessibility evaluation of online learning education smartphone applications. Online Education provides learners with flexibility to improve and update their skill set at flexible schedule that fits to him. Learners with disability or without disability have the right to free public online education. The objectives of this study are: To check accessibility status of online learning mobile applications of India under evaluation guidelines. To give useful suggestions and alternate solutions, issues among online learning mobile applications. WCAG 2.1 (Web Content Accessibility Guidelines) must be followed to make these applications accessible [2]. In this study, Accessibility Scanner Tool was used. It is a tool for automatically scanning and validating Android apps that can be installed on mobile phones. It's a free mobile application tool for android offered by Google Inc. [4]. Based on content labels, touch target, clickable items, text and image contrast, the accessibility scanner scans the screen and give recommendations to increase accessibility of your app. As the population increases the number of smartphone users and smartphone applications are also increasing day by day, this makes application development one of the most exciting area of software engineering. As a result, the development faces a significant challenge. As increase in smartphone devices the accessibility is neglected and rarely tested, because access barriers for people with disabilities are not expressly recognized [5].

LITERATURE REVIEW

In this research, the Google Accessibility Scanner, which is a mobile app, was used to conduct the accessibility study. The study examined different types of mobile apps using multiple performance metrics from WCAG 2.0 [12]. The report identified many commonly occurring accessibility failures and later proposed solutions to improve the errors [11]. Ten of the most popular smartphone apps, according to PCMAG, were evaluated using the Google Play Store's Accessibility Scanner. The test result shows that they do not achieve the WCAG 2.1 minimal acceptable standard [9]. The accessibility of 27 university educational websites evaluated with TAW and Wave tool to achieve WCAG 2.1 standard. The results demand further improvements [13]. The Accessibility Scanner tool was used to determine the accessibility of mobile apps for air quality, but not all of those mobile applications were available [10]. Examined 5,753 mobile applications, on the basis of tags, the study indicates that accessibility constraints are common issues, and it recommends that large-scale data collection and research be continued in order to enhance accessibility violations in mobile apps [6]. The findings demonstrated that prototypes developed using the WCAG 2.1 accessibility standards and the AT's (assistive technology) screen readers were more consistent with requirements [8]. The study indicated that smartphone apps would be available to every individual. The researcher suggests some modules that describe the issues and model regarding accessibility constraints [7].

METHODOLOGY

In January 2021, we have collected data regarding mobile apps. We have evaluated 10 e-learning mobile applications mostly used in India; the evaluated applications are encapsulated in Table 1. The automatic evaluation method is used for

mobile apps for the most popular Android operating system. From Play Store, Accessibility Scanner tool was chosen for evaluation. This tool follows some of the WCAG 2.1 principles [4]. The methodology uses five different steps for evaluating mobile apps, as shown in Fig. 1.

Step 1: Select & Install Applications from Google Play Store. In this step, we select mostly used 10 applications of online education for higher education in India. The apps were chosen from Play Store by Google; reference as the current version, size, updated on, and maximum number of downloads.

Step 2: Install and Activate Accessibility Scanner. In this step, a Google tool name as Accessibility Scanner was installed from play store. WCAG 2.1 standard applies for this tool. Activating this tool, a blue button appears on mobile home screen and requests for permission to capture screenshots and recording for accessibility test.

Step 3: In this step, we open individual app and activate the Accessibility scanner tool and give permission to capture screenshots and start the accessibility test, as an example, we select "Vidyakul" app and then we press tick mark blue icon button, present on screen. Afterwards, scanner tool will be ready to generate a report with the failures elements regarding evaluated mobile app as a result.

Step 4: In this step, the accessibility test report of individual application by Scanner tool is added in a sheet.

Step 5: Finally, the results were analyzed by Microsoft excel tool.

TABLE I
Tools for online education by using Google Play Store

Tool	Logo	Current Version	Download	Offered By	Size	Updated On
Vidyakul		4.036	100,000+	Vidyakul	22 MB	23 Dec,2020
Vedantu		1.8.0	10,000,000+	Vedantu	35 MB	18 Dec,2020
Unacademy		5.37.61	10,000,000+	SortingHat	31 MB	27 Jan,2021
Doubtnut		7.8.187	10,000,000+	Doubtnut	19 MB	23 Jan,2021
MyCBSEGuide		3.2.1	5,000,000+	ElpisTech.	10 MB	22 Oct,2020
Toppr		6.5.63	10,000,000+	Toppr	23 MB	7 Jan,2021
Meritnation		8.5.120	5,000,000+	Meritnation	16 MB	8 Jan,2021
BYJU'S		7.5.0.97	50,000,000+	BYJU'S	45 MB	18 Dec,2020
Adda247		9.5.4	5,000,000+	Adda247	12 MB	18 Jan,2021
Gradeup		10.3	10,000,000+	Gradeup	19 MB	21 Jan,2021

Fig. 1 Methods to examine accessibility in mobile apps

RESULTS AND DISCUSSIONS

WCAG 2.1guidelines [2] and mobile accessibility most related guidelines [3] applied to mobile applications shows in Table 2.

TABLE III
WCAG 2.1 PRINCIPLES AND GUIDELINES MOST RELATED APPLY TO MOBILE

WCAG 2.1 Principle (POUR)	Guidelines WCAG 2.1	Checkpoints WCAG 2.1 most related	Success Criteria for Mobile Apps	Level
(P) Perceivable	Guideline 1.4 Distinguishable	1.4.3 Contrast (minimum)	Requires a contrast of at least 4.5:1 (or 3:1 for large-scale text)	AA
		1.4.4 Resize Text	Text must be resizable up to 200%	AA
		1.4.6 Contrast Enhanced	Requires a contrast of at least 7:1 (or 4.5:1 for large-scale text)	AAA
(O) Operable	Guideline 2.1 Keyboard Accessible	2.1.1 Keyboard	Device manipulation gestures/Keyboard control for touch screen devices	A
		2.1.2 No Keyboard Trap	Keyboard control for touch screen devices	A
	Guideline 2.4 Navigable	2.4.3 Focus Order	Keyboard control for touch-screen devices	A
		2.4.4 Link Purpose (In Context)	Grouping operable elements that perform the same action	A
		2.4.7 Focus Visible	Keyboard control for touch-screen devices	AA
		2.4.9 Link Purpose (Link Only)	Grouping operable elements that perform the same action	AAA
	Guideline 2.5 Input Modalities	2.5.5 Target Size	Touch target size and spacing	AAA
(U) Understandable	Guideline 3.2 Predictable	3.2.3 Consistent Navigation	Provide instructions for custom touchscreen and device manipulation gestures	AA
		3.2.4 Consistent Identification		AA
(R) Robust	Guideline 4.1 Compatible	4.1.3 Status Message	Assigning appropriate error or success message	AA

Table 3 presents evaluation of multiple pages of mobile app, with the Google scanner tool that capture and scan the individual page of mobile app based on following: Content Label which further consist of (Item Label, Link Purpose unclear, and Item Description), Low Contrast which further consist of (Image Contrast, Text Contrast ratio), Touch Target Size, and Implementations which inherit further (Item type not supported and Clickable Items).

TABLE IIIII
MOBILE APPS EVALUATED BY ACCESSIBILITY SCANNER

Tools	Item Label	Touch Target	Item Description	Text Contrast	Image Contrast	Clickable Items
Meritnation	40	70	10	80	13	0
Toppr	15	39	6	98	0	0
Doubtnut	30	35	10	70	5	0
Gradeup	23	28	3	65	15	2
Unacademy	10	30	2	68	0	0
Vedantu	11	19	71	0	0	0
Adda247	30	50	3	14	0	0
MyCBSEGuide	3	16	1	45	16	0
Vidyakul	39	37	2	0	0	0
BYJU'S	32	33	2	10	0	0

Item label refers to the services such as screen reader which is a type of assistive technologies; rely on item labels to understand the meaning of elements in an interface. Screen readers like Windows Jaws, NVDA, and Google ChromeVox are essential to disabled person, who is blind, and people who are visually impaired. Missing of item label can help the disabled person who has any type of vision problem, from studying the content information on the display device screen; as a result, users with vision problems can find it difficult to access the content. It is suggested that to include the proper information about the type and state of the label, if it includes button for "save form" then the text in label would be "save form button". For touch target, it is suggested that, element of the screen must be large enough; it should be CSS pixels 48 by 48. Color contrast between images and text is therefore measured according to WCAG 2.1 recommendations, which state that small text must have a contrast of at least 4.5:1(below 18 points) and (18 points & above) for large text and a contrast of at least 3.0:1. Sufficient color contrast make application more attractive and makes text and images easier to read and understand.

Table 4 presents, a large number of errors corresponds to "Meritnation ", with 213 errors corresponding to 17.7%; followed by "Toppr" which 158 errors and corresponds to 13.2%. The tools with smallest number of issues are "Vidyakul" with 78 errors corresponding to 6.5%, and "BYJU'S" with 77 errors to 6.4%. The data obtained in reference to elements with faults show that the average value is 120.1, 105.5 as a median, 77 minimum, and 213 as a maximum, and 43.9 as a standard deviation value.

TABLE V
TOTAL ERRORS BY ACCESSIBILITY SCANNER TOOL

Tools	Elements	Per (%)
Meritnation	213	17.7%
Toppr	158	13.2%
Doubtnut	150	12.5%
Gradeup	136	11.3%
Unacademy	110	9.2%
Vedantu	101	8.4%
Adda247	97	8.1%
MyCBSEGuide	81	6.7%
Vidyakul	78	6.5%
BYJU'S	77	6.4%

The most common errors in the accessibility assessment of mobile apps are depicted in Fig. 2. The most common failures lead to "Text contrast" with 450 errors (37.5%), "Touch target" with 357 errors (29.7%), "Item label" with 233 errors (19.4%), "Item descriptions" with 110 errors (9.2%), "Image contrast" with 49 errors (4.1%), and "Clickable items" with 2 errors (0.17%). As a result, the more frequent errors are associated with Text Contrast, accompanied by the Touch Target.

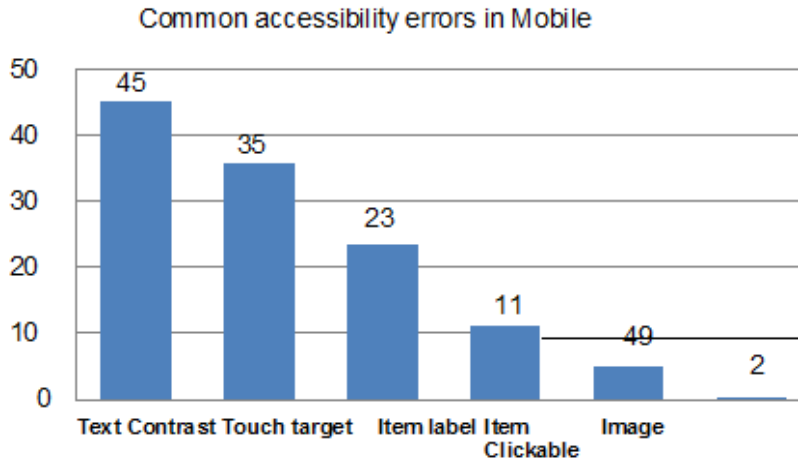


Fig. 2 Mobile App common accessibility errors

CONCLUSIONS

According to the findings, the mobile apps evaluated did not accomplish adequate levels of accessibility. As a result, it is important to correct the errors to comply with W3C (World Wide Web Consortium) suggested standard of accessibility. Furthermore, the assessment results can be used as a standard reference point for the creation of more accessible mobile applications for higher education. The application developers and designers need to follow the mobile accessibility guidelines throughout mobile application development cycle. As a future step, accessibility evaluation can be carried out for different types of mobile applications, and evaluation can also be performed on different operation systems.

REFERENCES

[1] https://www.who.int/news-room/facts-in-pictures/detail/disabilities.
[2] World Wide Web Consortium (W3C): Web Content Accessibility Guidelines (WCAG) 2.1. https://www.w3.org/TR/WCAG21/.
[3] World Wide Web Consortium: Mobile Accessibility: How WCAG 2.0 and Other W3C/WAI Guidelines Apply to Mobile. https://www.w3.org/TR/mobile-accessibility-mapping/.
[4] https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.auditor&hl=en_US&gl=US.
[5] Google: Make apps more accessible. https://developer.android.com/guide/topics/ui/ accessibility/apps.
[6] Ross, A.S., Zhang, X., Wobbrock, J.O., Fogarty, J.: Examining image-based button labeling for accessibility in Android apps through large-scale analysis. In: ACM SIGACCESS Conference on Computers and Accessibility (ASSETS 2018) (2018).
[7] El-Glaly, Y.N., Peruma, A., Krutz, D.E., Hawker, J.S.: Apps for everyone: mobile accessibility learning modules. ACM Inroads 9, 30–33 (2018). https://doi.org/10.1145/ 3182184.
[8] Carvalho, L.P., Freire, A.P.: Native or web-hybrid apps? An analysis of the adequacy for accessibility of Android interface components used with screen readers. In: Proceedings of 16th Brazilian Symposium on Human Factors in Computing Systems, pp. 362–371 (2017). https://doi.org/10.1145/3160504.3160511.
[9] Acosta-Vargas, P., Salvador-Ullauri, L., Jadán-Guerrero, J., Guevara, C., Sanchez-Gordon, S., Calle-Jimenez, T., Lara-Alvarez, P., Medina, A., Nunes, I.L.: Accessibility assessment in mobile applications for android. In: International Conference on Applied Human Factors and Ergonomics, pp. 279–288 (2019).
[10] Acosta-Vargas, P., Zalakeviciute, R., Luján-Mora, S., Hernandez, W.: Accessibility evaluation of mobile applications for monitoring air quality, pp. 1–11. Springer, Switzerland (2019).
[11] Zein, S., Salleh, N., Grundy, J.: A systematic mapping study of mobile application testing techniques. J. Syst. Softw. 117, 334–356 (2016). https://doi.org/10.1016/j.jss.2016.03.065.
[12] V.Balaji and K.S.Kuppusamy,“Accessibility Analysis of egovernance Oriented Mobile Applications”,in IEEE Explore – International Conference on Accessibility to Digital World (ICADW), 2016.
[13] V. Gupta and H. Singh, "Web Content Accessibility Evaluation of Universities' Websites - A Case Study for Universities of Punjab State in India," 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), 2021, pp. 546-550, doi: 10.1109/INDIACom51348.2021.00097.

VARIOUS CLUSTERING TECHNIQUES USED IN BIG DATA-A REVIEW

Prabhjot Kaur[1], Madan Lal[2], KanwalPreet Singh Attwal[3]
[123] *Computer Engineering, Punjabi University Patiala*
[1]prabjot351@gmail.com
[2]mlgtb@rediffmail.com
[3]kanwalp78@yahoo.com

ABSTRACT: Analysts classify and define big data generally with characteristics like volume, velocity, value, variety, veracity and variability. Big Data refers to the extreme data which is generated from millions of sources of data every second, which needs to be managed and dealt with in a smart and outstanding/effective way. Big data analysis pulls intelligence from a vast amount of complicated and dynamic data. The most crucial phase in big data analytics is data cleaning, which allows for easier prediction, decision- making, and clustering of data utilising data organising tools. There comes the idea of clustering. Clustering is the process of combining similar data from a large data points such that data points in the same group are more similar than data points from other groups. It has variety applications such as designing spam filters, identification of fraud or any criminal activity, performing analysis of documents, classifying traffic in network and helping in marketing analysis. The paper presents review on different clustering techniques which are applied on Big data.

KEYWORDS—Big Data, Clustering, k-Mean, Hadoop, Performance

1. INTRODUCTION

Since the amount of data is increasing at an alarming rate, analysis of data has become difficult. Data not only need to be collected and managed but also to gain meaningful information value from it. For this purpose different clustering techniques are employed on large data sets in order to gain informational and meaningful values. (SINANC, 2013)

Cluster analysis is the analysis of the structure of a dataset by dividing the data into groups so that the elements similar to one another are together in one group and dissimilar elements are assigned to different ones. Recently, along with the development of big data, cluster analysis has been extensively studied and widely applied in various fields, such as physics, biology, economics, engineering, sociology, and data mining. (Joarder & Ahmad). For solving the problem of clustering, several approaches have been proposed in the literature, which includes: non-hierarchical clustering (k-means, k-means ++, etc.), (Anoop & Sripriya, 2020) (Nair, Elayidom, & Gopalan, 2019) hierarchical clustering (T. Sajana, 2016), clustering for probability functions, or fuzzy clustering (Saeed, Aghbari, & Alsharidah, 2020). Among the above mentioned approaches, k-means clustering is the most well-known and widely applied in various fields. However, the k-means algorithm and its extensions usually require a pre user-defined number of clusters that is often unknown in practice. Furthermore, spherical clusters are obtained by k-means algorithm, which means it is not suitable for arbitrary-shaped clusters. These problems have been the major drawbacks of clustering so far which lead to many difficulties and challenges in solving this problem (Joarder & Ahmad).

2. RELATED STUDY

A hybrid of two major clustering algorithm had been proposed by Kumar, S.; Singh, M; to overcome the drawbacks of existing clustering schemes. Traditional approaches for clustering did not perform well on huge data due to complexity of data. Data set undertaken is publically available, National Climatic Data Center. The hybrid algorithm produces clusters of high quality and with less number of iterations which shows higher precision, recall and F measure. (Kumar & Singh, 2019)

Heidari, S; introduced MR- VDBSCAN, a new technique for clustering huge data with varying densities using a Hadoop framework running on Map Reduce, was introduced. The density of each point is calculated using local density. This can avoid the problem of combining clusters with varying densities. It consists of 3 layers, data partitioning layer, map reduce layer and merging and reliability layer. It is compared with other algorithms and proved to be more precise and accurate and showed good varying density clustering scalability and capability. (Heidari, Alborzi, Radfar, Afsharkazemi, & Ghatari, 2019)

Kumar, G.A.; did analysis of various clustering algorithms to classify the data points that are generated in a random manner. The investigation of various clustering strategies is done to extract information which helps researchers to choose better and effective algorithms to work depending on the prerequisites. In this scenario analysts classified the big data on various factors like Volume, Velocity and Variety. Various comparisons are made on the basis of different methods like data set classification, dimensionality, avoidance of outliers, shape of clusters and computational complexity. (Kumar G. , 2020)

Bangui, H. et al. aimed to identify drawbacks and to gap these drawbacks using the clustering approach. This research bridges the gap between clustering algorithms along with IoT. By tackling this specific aspect - clustering algorithm in Big Data, this paper examines big data technologies, related data clustering algorithms and possible usages in IoT. (Bangui & Buhnova, 2019)

Xiao, W.; Hu, J.; provided a classification for existing clustering algorithms based upon few factors of Spark and parallel clustering approaches. All the clustering algorithms has common features, they all are based on classical clustering algorithms, they all used RDD (Resilient distributed dataset) provided by Spark to store data points and the complete use of functions provided by spark to understand parallel operation of data partitions. All the characteristics of spark are fully used on memory computation which improved the efficiency of multiple iterations. (Xiao & Hu, 2020)

Joarder, Y.A.; Ahmad, M.; defined an algorithm that is Extended Generation k-Means algorithm that solved multiple issues like unhealthy initialization, dynamic centroid selection or empty clustering. This algorithm provides effective way of avoiding these defined drawbacks. Ten publically available data sets are taken from University of California for experimentation and it is proved that there is no lack of performance due to any modification and the suggested algorithm is semantically equivalent to traditional k- means. (Joarder & Ahmad)

Anoop, M.; Sripriya, P.; generated an algorithm FIC-PIXDCDC, that is based upon the Focused Information Criterion and Partitioned Iterative X-means Dice Correlation Data Clustering (FIC-PIXDCDC) Method. This algorithm aims at increasing the clustering performance of big geo spatial data in finding the frequently visited location of the user in social network platform when the geo spatial dataset is taken as an input. For experimentation Weeplaces dataset, obtained from location based social network services is considered. In this algorithm, the dataset is taken as input and number of clusters and centroids are randomly chosen. The dice correlation between each input data and cluster centroid is calculated. Focused Information criterion is applied to construct optimal number of clusters, process is repeated until no deviations are found in cluster centroid. Lastly, this method groups the interrelated geo-spatial data with higher accuracy and lower time to find the location information of frequently visited users in social network in a precise way. (Anoop & Sripriya,2020)

Nair, S.C.; Elayidon, M.S.; Gopalan, S.; defined a strategy that is parallel in nature and then applied in the Hadoop Distributed File System (HDFS) for reduction of execution time. They employed Bacterial Foraging Optimization algorithm (BFO) to handle the local optimality problem in k- means technique. Secondly, they designed parallel processing scheme for KM- BFO mechanism and is implemented in hadoop to reduce the execution time and is proved to be superior to existing approaches. This experiment was conducted using KDD-99 training data set. The Modified Bacterial Foraging Optimization (MBFO) algorithm calculated the wellness of population to choose the k values in clustering. (Nair, Elayidom, & Gopalan, 2019)

Saeed, M.M. et al.; investigated the Spark-based clustering methods by parameters of their support to the characteristics Big Data and propose a new taxonomy for the Spark-based clustering methods. (Saeed, Aghbari, & Alsharidah, 2020)

Zhang, X.; proposed an efficient 2 staged k-Mean clustering algorithm that is based upon observation point mechanism, which in turns find out the center of cluster that is to be created based upon non-disturbance of outliers in raw data. (Zhang, He, Jin, Qin, Azhar, & Huang, 2020)

Shen, Duan. discovered the importance and methods for data mining along with the concept of clustering in data mining, its basics, characteristics all on the basis of k- means. The data from teaching satisfaction based survey is collected and is modeled on SPP Modeler, which isa data mining platform. (Shen & Duan, 2020)

Tiwari, V.; provided a way out in terms of cloud computing to tackle the limitations of clustering algorithms in Big Data. By using the cloud computing framework data can be organized in a better way by using the traditional clustering algorithms. The conducted survey is on K means++ and Mini Batch K means clustering algorithm using Map reduce in cloud computing to overcome the performance degradation problem because of the sequential processing approaches. Here, cloud computing along with parallel processing is introduced which is the solution of the defined problem. It also helps in reducing the cost. (Tiwari & Waoo, 2019)

Suryawanshi, R.; proposed an improved k-Means clustering algorithm which has given a way out to tackle the drawbacks of traditional k-Means algorithm like outliers and execution time management and problems arose during the large sized dataset. (Suryawanshi & Puthran, 2016)

Ding, Sun, & Zeng,; proposed a big data clustering algorithm which is completely based on cloud computing. To process the numerical attribute, the cloud extended distributed feature fitting method is used. The research was done using a combination of fuzzy C- scores and linear regression analysis. Cloud computing and adaptive quantitative recurrent classifiers are used to classify data.. The results showed better information fusion performance and retrieval ability of numerical data. (Ding, Sun, & Zeng, 2020)

Sajana, T et al.; focused on to study of various clustering algorithms and also extracted outthe characteristics along with the advantages and drawbacks of clustering algorithms. The result concluded that to identify outliers in vast datasets, BIRCH, CLIQUE and ORCULUS algorithms should be used. On spatial data to obtain clusters of arbitrary shape STING, ORCULUS, PROCULUS, for categorical data, CURE, ROCK and on numerical data non convex shaped clusters can be formed by COBWEB and CLASSIT algorithms. (T. Sajana, 2016)

Tao, Q. et al.; proposed an intelligent clustering algorithm to tackle the problem of carrying out clustering in high dimensional datasets. In this algorithm k-Means is modified to produce novel intelligent weighting k-means clustering. It firstly increase the distance in clusters and secondly eliminate the sensitivity of initial cluster centres.

The convex clustering model and compositional convex clustering with Sparse Group Lasso are extended for clustering

high-dimensional sparse compositional data. This information comes from the field of microbiology.. (Tao, Gu, Wang, & Jiang, 2019)

Wu, C.; also obtained an optimized solution to deal with large scale data sets that are present in Big Data. It also tackles the problem of high computational complexity in traditional clustering algorithms. (WU, 2019)

Dafir, Z.; provided an overview of parallel computing technique in clustering and categorize these based upon platform and efficiency of techniques for working in Big Data. This review found out the CURF technique to be better for parallel computing for clustering in Big Data. (Dafir, Lamari, & Slaoui, 2020)the computing

M.V.S Prasad, O.Naga Raju; proposed an optimized repartitioned k- means clustering algorithm which when performed on pollution data set with 449 files with 17500 observations each using Map Reduce hadoop framework provides high performance , measured on the basis of inter and intra cluster similarity feature in Big data analysis. An improved repartitioned k-means algorithm is exercised to large set of data for high computational processing over group of connected systems on top of Map Reduce framework which provides high performance. K means is implemented utilising map reduction technology, and the notion of repartitioning is also introduced, which separates the acquired virtual data partitions into new partitions of known number in order to minimise work. Each virtual partitioning generated in the Map phase is recombined during the repartitioningprocedure. The results showed that the reduction step was completed faster, and that the workwas completed faster as well. (M.V.S. Prasad, 2020)

Agnivesh, Rajiv Pandey, Amarjeet Singh; proposed some of the improvements in k means clustering algorithm to overcome its limitations .They presented 2 algorithms, First, computesthe location of the initial points or the seeds which are huge in number. Second, merges these seeds on the basis of an edge value. Experimentation is done on UCI real data sets and resultsproved that this work excels traditional k means and k means Map reduce as the performance is high due to less computation time and also is cost effective (Agnivesh, 2019)

3. BIG DATA

Big data is the term used for gigantic sets of data that have large, different and complex structure which makes the data difficult to store, analyse and visualize for further processes and results using traditional tools and techniques. The extent to which information is generated and made available, or digitization, is the most important reason for the creation of Big data. It's the process of converting continuous and analogue data into a digital and machine-readable format. The data in Big data is of three type- Structured data, Semi Structures and Unstructured data.

Discovering and extracting usable information and knowledge from such a large volume of data is difficult, and traditional relational databases cannot match the needs of users, new technologies are required to preserve and extract meaningful information from it. Large datais typically defined as a collection of data that exceeds the capacity of normal management tools and databases to extract, refine, manage, and analyse it. Since, it is not possible to manage them with classical tools, it is difficult to extract hidden knowledge and knowledge atpredetermined times (Heidari, Alborzi, Radfar, Afsharkazemi, & Ghatari, 2019). Furthermore, the concept of big data analytics arises; it is application of the analytics technique. It is useful in providing valuable insights from this varied and rapidly differing data.

Big data is defined by its characteristics of the heterogeneous data. They are Volume, Variety, Velocity, Veracity, Variability and Value. Volume is the magnitude of the data in terabytes and petabytes. This high volume outcast the traditional storage and analysis techniques. Variety defines the structural heterogeneity of the data. It comes from variety of different sources such as structured, unstructured and semi structured data. The rate at which data is generated, as well as the pace at which it is analysed and dealt with, is referred to as velocity. The rapid proliferation of data necessitated real-time analytics and evidence planning. Because standard data management systems were unable to handle the massive amounts of data, big data technologies were implemented. Other characteristics were also added later on to clearly define and understand what Big data really is, they were:

Veracity, it was coined by IBM and refers to unreliability and uncertainty in the data sources. Variability also known as complexity is another characteristic which describes the variation in data rate as it is generated through number of resources. Lastly, Value, introduced by Oracle defines the big data attribute. These dimensions or characteristics are dependent on each other, if one dimension changes there is high possibility that other dimension will change too.

3.1 Challenges

The highlighted challenges linked with Big data were capturing data, storage, searching, sharing, transfer, analysis and visualization. Organizations usually take aid of enterprise servers to face these challenges. Google solved these issues using an algorithm known asMap Reduce which runs on Hadoop, it divides the task or job into smaller sub jobs and assignevery computer to particular sub job and look for results when all jobs get executed at the same time. It follows divide and conquer approach by breaking big problem into smaller unitsand processing them in parallel.

Map Reduce consists of two stages; Map and Reduce Step.

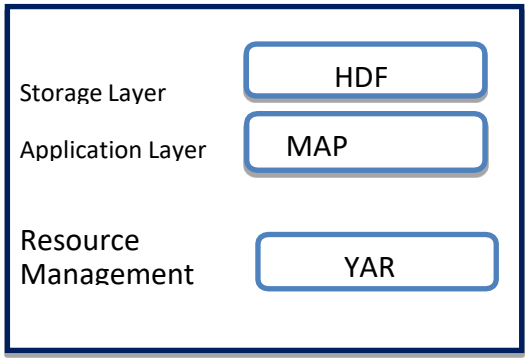
- Map –Each element is broken down into tuples, which are key value pairs, and the set of datais turned into another set of data.
- Reduce –It takes the Map stage's output as input and combines the tuples into a smallercollection of tuples.

3.2 Hadoop

It is created after getting inspired by the solution provided by Google. It is open source platform and have java based framework. It works in an environment which provides distributed storage and computation across computer clusters.

Hadoop Architecture: it comprises of three major layers.

- HDFS
- Yarn
- Map Reduce



4. CLUSTERING

Clustering a data mining technique used to analyze the data stored in various fields. Basically it is applied on big data to manage and analyze it effectively. (Swarndeep Saket J, 2016)

It is unsupervised strategy that comes handy in dealing with vast data, it is used to construct models that are descriptive in nature. Unsupervised algorithms look for the patterns and certain structures in the variables because there is no target variable (Attwal & Dhiman, 2020). Here, the data is divided into groups or clusters of objects that show more likeliness to each other instead to the objects of other groups. Clustering divides the data into clusters on the basis of data homogeneity. The division takes place according to the inter class and intra class similarities of the data objects in the data set to make the final clusters. The good cluster, must have high intra class and less interclass similarity (Kumar G. , 2020). According to (Attwal & Dhiman, 2020), clustering parts the objects or the records of the database into different groups which depends on the criteria defined on the characteristics of the objects.

These clusters are formed only after the execution of the clustering algorithms. There are different types of clustering algorithms used depending upon the type of data. The next section provides with detailed information.

4.1 CLUSTERING ALGORITHM IN BIG DATA

This section provides diverse clustering algorithms after taking all the properties of Big Data under consideration like, noise, dimensionality, algorithm computations, cluster shape and size, etc.

4.1.1 Partitioning based Clustering algorithms

In the beginning all objects are assumed to be in one big cluster. By repeatedly locating the points between the divisions, the objects are separated into a number of partitions. The initial data set contains n objects, and this technique separates the accessible data into k parts. Also each object must belong to exactly one group and each group must contain at least one object.Each partition must be considered as one group or cluster

The partitioning algorithm includes K-means, K- medoids (PAM, CLARA, CLARANS,). Clusters of Non convex shapes are obtained.

4.1.2 Hierarchical Clustering algorithm

The hierarchical decay of the data set or anything under consideration takes place. Hierarchical clustering is performed in 2 ways, Agglomerative and Divisive. In the Agglomerative technique, a single object or data point is first considered a cluster, and then numerous merge operations based on the minimal distance are done. The process is repeated until the desired cluster number is reached. The Divisive method works by fragmenting a cluster of objects into little clusters until the required clusters are generated. Hierarchical algorithms in which clusters of non-convex and arbitrary hyper rectangular shape formed are BIRCH, CURE, ROCK, Chameleon, SNN, GRIDCLUST, CACTUS.

4.1.3 Density based Clustering algorithm

The points belonging to each cluster are supposed to be drawn from a certain probability distribution in the density-based method. Only spherical clusters can be used with this technique. The advantage of such clustering is that the density of points inside the cluster is much higher than outside. Core points, boundary points, and noise points are the three types of data items. All of the core points are joined together to form a cluster based on their densities. Various clustering

methods, such as DBSCAN, OPTICS, DBCLASD, GDBSCAN, and DENCLU, produce arbitrary shaped clusters.

4.1.4 Grid based Clustering algorithm

The data set is partitioned into a number of cells using a grid-based technique. The grid layout is used to create clusters. To construct clusters, the Grid algorithm employs subspace and hierarchical algorithms. STING, CLIQUE, Wave cluster, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS are among the approaches. Grid algorithms are exceptionally speedy atprocessing when compared to all Clustering algorithms. Adaptive grid techniques such as MAFIA and AMR have been developed to address these issues. The grid cells create arbitrary shaped clusters.

4.1.5 Model based Clustering algorithm

This algorithm is based on hypothesizing a model for every cluster to find best fit of the data according to the mathematical model. It can automatically determine the number of cluster onthe basis of standard statistics. The method may locate clusters by constructing a density function that reflects the spatial distribution of the data points. The number of clusters can be automatically determined based on standard statistics taking outlier.Various tactics, including as statistical, conceptual, and resilient clustering methods, are used to connect a set of data points. Neural network approach and Statistical approach are the two approaches for performing model based algorithms. Well known model based clustering algorithms are EM, COBWEB, CLASSIT, SOM, and SLINK.

5. CONCLUSION

This research examines several clustering approaches and separates data clustering algorithms for managing large data sets. Generally, clustering algorithms must be improved by reducing their time and space difficulties in order to manage large amounts of data. In the presence of a large number of outliers, the CLIQUE, BIRCH, and ORCLUS algorithms perform better in data clustering for big data analytics, according to the study. The current study also reveals that by using the CURE and ROCK techniques on ordered data, effective clusters can be generated. Spatial information methods like OPTIGRID, STING, PROCLUS, and ORCLUS can be included into clustering algorithms to construct effective discretionary clusters.

6. REFERENCES

[1]. Kumar, S.; Singh, M.; ―A Novel Clustering Technique for Efficient Clustering of Big Data in Hadoop Ecosystem‖, Big Data Mining and Analytics, Vol. 2, No: 4, 2019, page no: 240-247

[2]. Heidari, S.; Alborzi, M.; Radfar, R.; Afsharkazemi, M.A.; Ghatari, A.R.; ―Big data clustering with varied density based on MapReduce‖, Vol. 6, Number: 77, 2019

[3]. Kumar, G.A.; ―The Study Of Various Clustering Algorithms In Big Data Clustering,‖ International Journal Of Scientific & Technology Research, Vol. 9, Issue: 04, 2020, page no: 6-12

[4]. Bangui, H., Ge, M.; Buhnova, B.; ―Exploring Big Data Clustering Algorithms for Internet of Things Applications‖, Exploring Big Data Clustering Algorithms for Internet of Things Applications, 2019, page no: 269-276

[5]. Xiao, W.; Hu, J.; ―A Survey of Parallel Clustering Algorithms Based on Spark‖, Hindawi Scientific Programming, Vol. 2020, Article Id: 8884926, 2020, page no: 1-12

[6]. Joarder, Y.A.; Ahmad, M.; ―A Hybrid Algorithm Based Robust Big Data Clustering for Solving Unhealthy Initialization, Dynamic Centroid Selection and Empty clustering Problems with Analysis‖,WorldUniversity of Bangladesh, https:// arxiv.org/ftp/arxiv/papers/2002/2002.09380.pdf, 2020, page no: 1-18

[7]. Anoop, M.; Sripriya, P.; ―Focused Information Criterion Based Partitioned Iterative X-Means Dice Correlation Clustering For Big Geo-Social Data‖, Journal of Critical Reviews, Vol. 7, Issue: 6, 2020, page no: 54-62

[8]. Nair, S.C.; Elayidom, M.S.;Gopalan, S.; ―KM-MBFO: A Hybrid Hadoop Map Reduce Access for Clustering Big Data by Adopting Modified Bacterial Foraging Optimization Algorithm‖, Vol. 8, Issue: 2S11, 2019, page no: 146-152

[9]. T. Sajana, C.M.S. Rani, and K.V. Narayana, ―A Survey on Clustering Techniques for Big Data Mining,‖ Indian Journal of Science and Technology, vol.Vol. 9, Issue: 3, 2016, page no: 1-12

[10]. Saeed, M.M.; Aghbari, Z.A.; Alsharidah, M.; ―Big data clustering techniques based on Spark: a literature review‖, PeerJ Computer Science, Vol. 6, 2020, page no: 1-28

[11]. Zhang, X.; He, Y.; Jin, Y.; Qin, H.; Azhar, M.; Huang, J.Z.; ―A Robust k-Means Clustering Algorithm Based on Observation Point Mechanism‖, Hindawi, Vol. 2020, Article ID: 3650926, 2020, page no: 1-11

[12]. Shen, H.; Duan, H.; ―Application Research of Clustering Algorithm Based on K-Means in Data Mining‖, International Conference on Computer Information and Big Data Applications (CIBDA), IEEE, 2020, page no: 232-241

[13]. Tiwari, V.; Waoo, A.A.; ―Comparatively Analysis on K-Means++ and Mini Batch K-Means Clustering Algorithm in Cloud Computing with Map Reduce‖, International Research Journal of Engineering and Technology, Vol. 6, Issue: 9, 2019, page no: 814-818

[14]. Suryawanshi, R.; Puthran, S.; ―A Novel Approach for Data Clustering using Improved Kmeans Algorithm‖, International Journal of Computer Applications, Vol. 142, No. 12, 2016, page no: 13-18

[15]. Abdel-Fattah, M.A.; Helmy, Y.M.; Mossad, S.M.; ―Improving the Efficiency of Implementing KMeans Algorithm on Different Big Data Platforms‖, International Journal of Scientific & Engineering Research, Vol. 11, Issue: 1, 2020, page no: 52-57

[16]. Ding, H.; Sun, C.; Zeng, J.; ―Fuzzy Weighted Clustering Method for Numerical Attributes of Communication Big Data Based on Cloud Computing‖, Symmetry, Vol. 12, Issue: 530, 2020, page no: 1-12

[17]. Tao, Q.; Gu, C.; Wang, Z.; Jiang, D.; ―An intelligent clustering algorithm for high- dimensional multiview data in big data applications‖, Neurocomputing, Vol. 393, 2020, page: 234-244

[18]. Wu, C.; ―Research on Clustering Algorithm Based on Big Data Background‖, Journal of Physics: Conference Series, 2019

[19]. Dafir, Z.; Lamari, Y.; Slaoui, S.C.; ―A survey on parallel clustering algorithms for Big Data‖, Springer, 2020

[20]. Amir Gandomi and Murtaza Haider, "Big Data Concepts, Methods, and Analytics," Elsevier, 2014.

[21]. M.V.S Prasad, Dr. O.Naga Raju ;‖Extended k-means clustering technique using Map reduced Segmentation clustering for Big Data‖,IJRAR,2020

[22]. Agnivesh, Rajiv Pandey, Amarjeet Singh; ―Enhancing K-means for Multidimensional Big Data Clustering using R on Cloud‖ International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-7, May 2019

[23]. Seref Sagiroglu ; Duygu Sinac.; :Big Data: A Review‖ Gazi University Department of Computer Engineering, Faculty of Engineering Ankara, Turkey. IEEE,2013

[24]. Swarndeep Saket J and Dr. Sharnil Pandya, ―An Overview of Partitioning Algorithms in Clustering Technique,‖ 2016.

[25]. Dheeraj Kumar, James C. Bezdek, Sutharshan Rajasegarar, Marimuthu Palaniswami, Christopher Leckie, and Timothy Craig Havens; ―A Hybrid Approach to Clustering in Big Data‖ , IEEE.

[26]. Mugdha Jain, Chakradhar Verma,; ―Adapting k-means for Clustering in Big Data‖, International Journal of Computer Applications,2014

[27]. Kanwal Preet Singh Attwal; Amardeep Singh Dhiman,; ―Investigation and comparative analysis of data mining techniques for the prediction of crop yield‖ Int. J. Sustainable Agricultural Management and Informatics, Vol. 6, No. 1, 2020

IMPLEMENTATION of BI-DIRECTIONAL HYBRID OPTICAL - WIRELESS ACCESS NETWORK AND ANALYSIS of MULTI PATH FADING on it

Harmanjot Singh[#1], Simranjit Singh[#2], Simranjit Singh Tiwana[#3]
Department Of Electronics & Communication, Punjabi University Patiala
Harman.dhaliwal.nba@gmail.com
sjsingh@pbi.ac.in
simranjit@live.com

ABSTRACT: Hybrid Optical Wireless Communication Access Network (HOWAN) has gained huge popularity, by offering attractive and seamless options for many human communication needs due to beneficial characteristics. For example, flexibility in movement, cost effective solution, better utilization of bandwidth and mobility option. With the huge demand of access, networks for various users are pushing the Free Space Optics (FSO) Communication to offer higher data rates and increase its traffic. FSO system also has one advantage that to low installation cost and easy to set up as compare with fiber installation is difficult. This paper aims to present a multi carrier generation by HOWAN UDWDM PON (Hybrid Optical Wireless-Access Ultra Dense, Wavelength Division Multiplexing Passive Optical Network) based on orthogonal frequency division multiplexing (OFDM) format that can achieve both wired and wireless access network and impact of fading on the signal during transmission. Symmetric compensation at 10Gbps with channel spacing 25 GHz, covered the distance up to 100 km for both upstream and downstream. At the Optical Line Terminal (OLT), multi carrier signal based multiplexed OFDM modulated data for all the 8channels, each having 10Gbps for downlink is transmitted thorough Single Mode Optical Fiber (SMOF) over Free Space Optics (FSO)communication and then check the impact of multi path Fading through IEEE 802.11 AC TGac channel. At the user end, a wavelength re-used technique is employed for better utilization of resources to carry the upstream data at the Optical Network Unit (ONU).The quality factor observed for both upstream and downstream lies between 6 to 6.7 range, the average bit error rate lies within $1.5e^{-10}$. Thepath loss and shadowing reported at the spectrum when signal passed through Free Space Optics is -50 dBm to -100 dBm.It is observed from the system that downstream transmission of signal within the network performs better than upstream data transmission due to losses practiced by upstream signal by using re-modulation of carrier signal. The system is evaluated based on the performance of the network with respect to Bit Error Rate, coverage area and obtain a network with an excellent access property.

KEYWORDS: - QAM sequence decoder; PON (Passive Optical Network), UDWDM (Ultra Dense Wavelength Division Multiplexing), SMF (Single Mode Fiber), DCF (Dispersion Compensation Fiber), OFDM (Orthogonal Frequency Division Multiplexing), FSO (Free Space Optics) and HOWAN (Hybrid Optical Wireless Access Network)

1. INTRODUCTION:

In the upcoming future, the human settlement is expectable to be available towards the urban areas that are considered to be smart as new urban development is expanding so progressively which will have practiced of providing the processing details in upcoming advanced era of world. The advanced information stages of development will provide benefit in term of economic hardship, social development, technological innovation, mental stability for the welfare in habitat of the human race so as to facilities that include management, e-learning, tele-health, safekeeping and privacy, landholdings, sustainable haulage, serviceableness [1].by including hybrid approach for both optical and wireless communication the requirement in technological development and infrastructure capacity, optical communication provide advantages in term of higher bandwidth and wireless technology could supplement portability needs for the end users. Accordingly, numerous research has been carried out and on the way for the smart cities requirements, the best solution is for introducing the hybrid architecture for optical as well as wireless communication, which mainly engrossed towards the capacity for planning and designing in networks [2, 3]. At the optical technology end, the new high capacity technology has been demanding the progressive needs of urban areas with smart technology such as WDM i.e Wireless Division Multiplexing requirement to boost rate of increase in traffic of data and capacity for high speed of data transmission [4, 5]. Wireless Division Multiplexing technology can be widely used in communication system that multiplex or combine multiple signals on a optical fibre used on single by combining different signals on different wavelengths simultaneously. Proposed criteria helps to increase the capacity of the system. There are two techniques for HOWAN, firstly, ROF i.e Radio over Fiber and secondly fibre as well as radio. Radio over Fiber (RoF) provide effective technology convergence with respect to system having access to both optical as well as wireless system [6]. Radio over Fiber is a technology for communication in which, light in the form of data modulation within radio signals and transmission through optical channel communication link towards the users in other end in the form of wireless signal [7]. Thus practice of Radio over Fiber (RoF) toward the smart areas in urban cities is new concept, however, it lags implementation on a higher scale [8]. digitization in Radio over Fiber RoF was realized and in practice of various fields of optical communication, long reach communication in satellites, as well as local area wireless systems [9, 10]. Numerous research has carried out towards recent times in relation to RoF and WDM focused entirely on a high capacity HOWAN system which could have improvement in the efficacy and reliability for high rate as well as low cost communication network systems [11, 12]. Second advancement in the HOWAN is the use of Radio and fibre (R&F) optical communication at back haul and Free Space Optics (FSO) at front user end. In this

technique, at the back haul network end optical transmission of signal via the fibre for optical systems and on front haul, signal is demodulated and transmitted in electrical form through wireless devices.

1.1 Free Space Optics (FSO):

Free Space Optics (FSO) considered as optical axis technology which employs lasers advantage for travel through air so as to achieve high connections towards bandwidth for optical systems benefits or it's an approach for the propagation of light towards LED's and LASER's through frees space (air, vacuum, outer space, etc.). Free Space Optics establish large speed of data as light provides higher bandwidth with no requirement for optical fibre cable or costly spectral media. It mostly operates within a range from 780 nm to1600 nm band window with using uses O to E and E to O conversion devices for interface between optical and electrical signals.

FSO consist of three stages during its operations: first of all, optical radiation transmission in air medium, free space, other wireless medium which exhibit factors like pollution, aerosol, fog, gases, smoke etc. and secondly, receiving end to receive the signal reception at user end. The range of FSO varies from 300 meter to 10 k meter, basically it is used for short range communication. the need of future demands replacement of optical fibre by Free Space Optics because it provides portability to user. Like the optical fiber communication, FSO also make use of laser devices for data transmission, however difference lies between the medium of communication in place of a fiber glass, it uses air as transmission medium.

1.2 Fading: When the signal travel through wireless medium, due to the presence of multiple objects within the path, the signal experience change in its characteristics due to various parameters like (reflection, refraction, diffraction, scattering, etc.). It is a random process which can be shown by communication channel that experience fading. This lead to attenuation of a signal with various variables like (time, position of user, frequency of the signal) and then further divided into small scale and large scale fading. There are many different types of fading like Rayleigh fading, rician fading, nakagami distribution

1.3 Hybrid Optical Wireless Communication (HOWC): The term explains how the communication in wireless interface within optical fibre technology. Generally, HOWC consist of IR i.e. Infrared Communication for shortest distance range and FSO i.e. Free-Space Optics Communication for large data range at user end. The VLC i.e. Visible Light Communication is a technology for data communication where LED and LASER are utilized for optical carrier data transmission on back bone network. Nowadays, LED i.e. light emitting diode at source using wavelength in the range of 380 nm to 780 nm can be vigorously established in the form of silicon photodiode that responding to be used as detector as well [14]. Here channel of transmission includes optical fiber, Indoor, Vacuum or Outdoor.

2. SYSTEM DESCRIPTION

The growing demand for commercial software for simulation and design of optical communication system has led to the availability of a number of different software solutions. With the advancement of technologies OptiSystem 17.0 is useful and innovative optical communication system simulation tool for designing, testing and optimization of optical link.

With the help of Optisystem software tool network achieves good and efficient results. The proposed 8×100 Gbps Hybrid optical wireless ultra-dense spectral efficient wavelength division multiplexing passive optical network as shown in Figure 1.1 is modeled by using OptiSystem 17.0 software and Matlab R2016a software. Network shows full duplex communication from OLT i.e. Optical Line Terminal Backhaul to ONU i.e. Optical Network Unit Front end or vice versa.

Figure 1.1 shows the basic network of optical OFDM i.e Orthogonal Frequency Division Multiplexing with UDPON i.e. Ultra-Dense Passive Optical Network employing free space optics at user end. Non optical carrier single sideband modulation is used to make the system more efficient, more tolerant to scattering and chromatic dispersion path loss impacts. Ultra dense wavelength multiplexing and provides high spectral efficiency by using 25 GHz channel spacing, which provide spectral efficiency by accommodating more number of user in narrow wavelength.

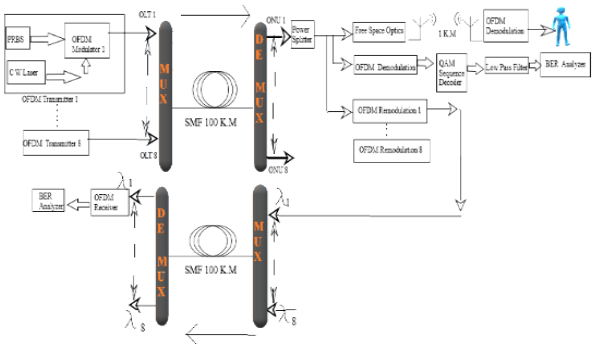


Figure 1: Hybrid Optical Wireless Ultra Dense Spectral Efficient Bidirectional Passive Optical Network

Data rate for the bidirectional HOWAN optical system is taken as 10 Gbps for both downstream and upstream. System parameter details are recorded in Table.1. In OFDM, multiple closely spaced orthogonal subcarrier signals are transmitted

with overlying spectral to carry data in parallel form. Modulation is centered on inverse fast Fourier transform at back end and Demodulation is based on Fast Fourier Transform algorithms at the user end.

Figure 2 represents the component involve for on-carrier single sideband modulation generation. In this generation of the signal, In the beginning pseudo random code bit generator produces the binary data bits in form of 1's and 0's. After Quadrature amplitude the signal is map into parallel bits. Then the signal is pass through IFFT for inverse fast Fourier transform to synthesize the signal in a discrete frequency domain, then again convert the signal into serial form by employing parallel to serial converter.

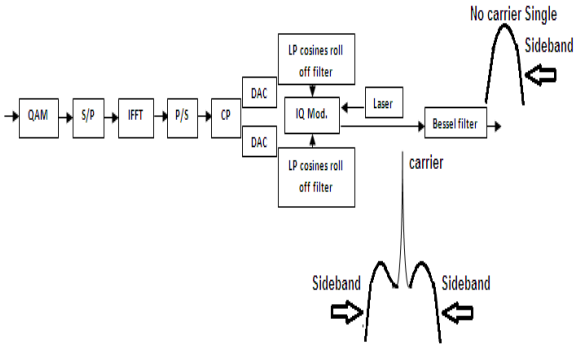


Figure 2: Design circuitry of non-carrier single sideband modulation signal

In this circuitry the main aim is to quash the effects of inter-carrier interference in the system, cyclic prefix is fused in network for this. Digital to analog conversion is essential for signal modulation and it is done by DAC, signal is up-converted to optical frequency by using the help of two intensity modulators (MZMs). A continuous wave laser preferred for operation is C-band (1530 nm-1570 nm), because of the minimum scattering effect in this frequency band. Non carrier single sideband signal is admitted after OFDM signal modulation generation through Bessel filter for all the wavelength in use.

Table 1: Parameters defined for HOWC-UDWDM bidirectional network

Downstream Data rate	10 Gbps
Upstream Data rate	10 Gbps
Downstream Expanse	100 Km
Upstream Expanse	100 Km
Modulation used	CO-OFDM
OFDM supported subcarriers	512
FFT/IFFT points	1024
No. of Channels link	8
Channel Spacing	25GHz
Distance of Free space Optics	1 kilometer
Bandwidth of MUX	60 GHz
Signal for transmission	Non-Carrier Single Sideband Modulated signal
Source frequency range of LED	193.1THz
Fading	IEEE 802.11ac

Multiplexed optical signals are transmitted over single mode fiber of 100 km length for downstream having attenuation in the range of 0.4 dB/km and dispersion effect of 16.75 ps/nm/km. Receiver section involves de-multiplexer after optical fiber and align particular wavelength to precise output port with the respect to multiplexer. Prior to OFDM demodulation of signal, each received wavelength at particular port is divided into three signals by using power splitter. One division of power splitter is send to the free space optics (FSO) and then to coherent optical OFDM demodulator. After than this demodulated signal is obtained in binary form and send to the IEEE 802.11 WLAN TGac fading channel for analysis of multipath fading on the network and another is send to wired receiver and the third one is bolstered to another side of the system so called upstream signal to re modulate the signal from downstream to upstream. Here, in the re modulation process, the system uses the power of the original signal to re modulate the data from downstream to upstream. Demodulator section consist of local oscillator for providing phase matching, photo-detectors and avalanche photo diode to get the real and imaginary part of the signals, to detect the multilevel signal m-array threshold detector is used followed by QAM decoder. An error vector magnitude (EVM)is calculated by constellation analyzer. It is to be noticed that no optical amplifier is used in this proposed communication network. At the last BER analyzer is used with the end goal that it gives results with respect to Q factor and BER.

3. RESULTS AND DISCUSSION

Keeping in mind the end goal of this work is to achieve a non-optical carrier single sideband modulation, hybrid optical wireless passive optical network, a comprehensive simulation tool Optiwave optisystem17.0 and matlabR2016a is used.

Optical spectrum analyzer is used at the receiver for visualize the carrier signal that illustrates the frequency as well as power for each carrier signal. Figure 5 (a) shows the double side band of the signal after OFDM modulation at transmitter side on optical spectrum analyzer and Figure 5 (b) illustrate the single side band representation of the signal with non-optical carrier after passing through the Bessel filter.

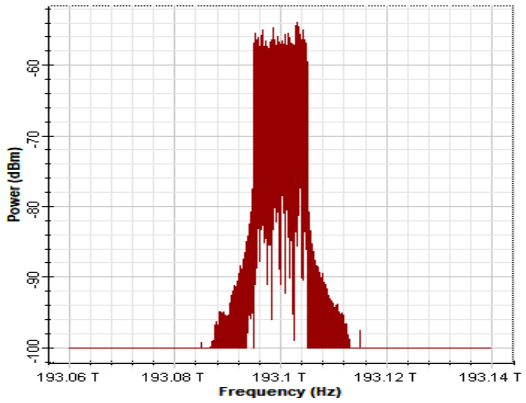


Figure 5 (a): Double sideband signal representation on optical spectrum

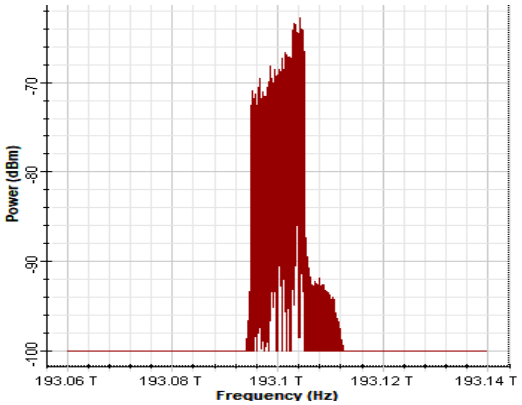


Figure 5 (b): Optical spectrum of Non-optical carrier single sideband modulation signal (NOC-SSB).

3.1 Downstream and Upstream Transmission Performance: The OFDM downstream and upstream transmission is evaluated at a distance of 100 km within acceptable range of Q factor. It is apparent that with the increase in the communication length beyond 100 Km, the Q Factor and Bit Error Rate are not within the permissible international defined standard for both downstream and upstream transmission.

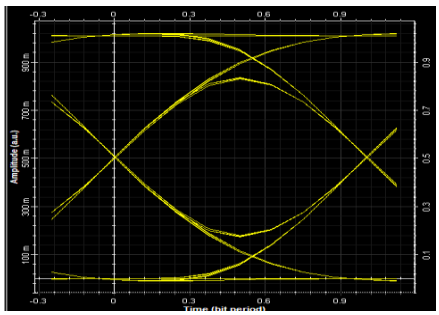


Figure 6 (a): The eye graph of downstream 10Gbps modulated signal after 100 km long for wireless access

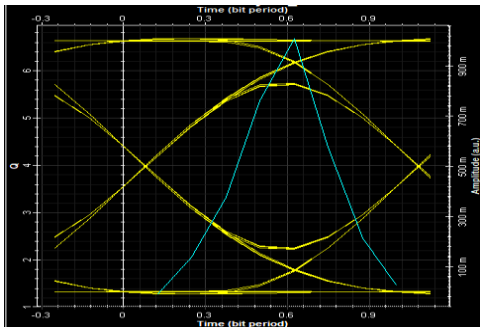


Figure 6 (b): The eye graph of Upstream 10Gbps modulated signal after 100km long for wired access

From the Figure 6, It is observed that Eye opening is more in case of the downstream than upstream. Eye opening is directly proportional to the Q factor. Figure 7 illustrate the radio frequency spectrum of the signal with respect to the power at the user end.

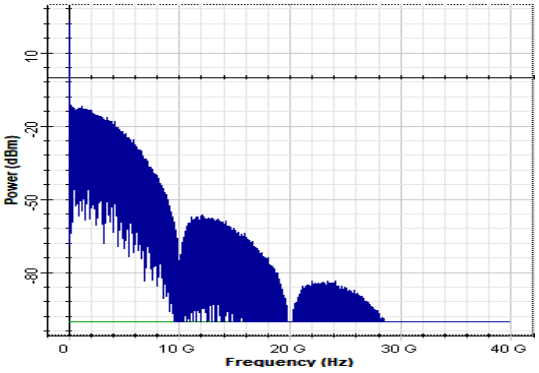


Figure 7: Radio Frequency Spectrum of Downstream Signal after Free Space Optics.

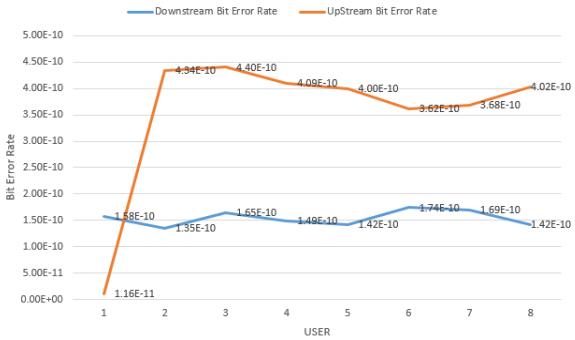


Figure 8: Graphical representation of various Users and Minimum Bit Error rate in HOWAN.

Figure 8 shows the various users in comparison to the Bit error rate in the downstream and upstream signal. Here from the figure it shows that bit error rate with respect to the user and is in permissible range up to 10 Gbps, above it the signal get distorted. From the investigation it reveals that the errors in the network increase in the upstream transmission with the increase in distance. However, it has been observed that the errors impact are less in case of downstream signal due to re modulation of carrier signal and it is notable that BER of 1.5 e^{-10} reported at the 100 km.

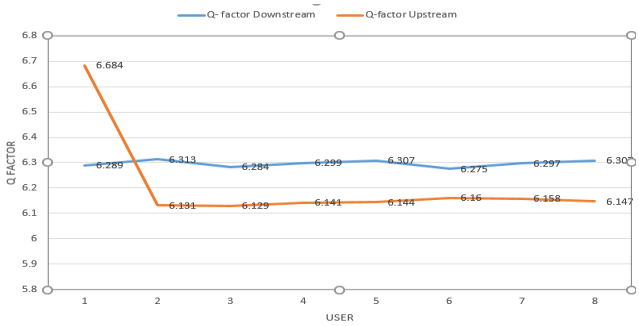


Figure 9: Graphical representation of Min BER and Power of HOWAN.

The graphical representation of minimum bit error rate and power of HOWAN is shown in Figure 9. From the Figure, it is very clear that the quality factor for both upstream and downstream lies between 6 to 6.7 ranges, which provide better results.

Table 2: Parameters of FADING Channel

Fading	IEEE 802.11 AC TGac
Number of Occupied Subcarriers (NST) **(N_{ST})**	242
Number of Data Subcarriers **(N_{SD})**	234
Number of Pilot Subcarriers **(N_{SP})**	8
Number of Transmitting Antenna	1
Number of Guard Interval	2048
Modulation	Binary Phase Shift Keying

The fading on the received bits at the user end is analyzed by passing the bits through the WLAN channel system object in Matlab software. The wlan TGac Channel perform the filtering of an input signal through an IEEE 802.11ac (TGac) multipath fading channel. First of all, Create the wlan TGac Channel object and set its parameters. Table 2 shows the parameter of fading channel used at the receiver side, which is used to analyze the fading in the channel.

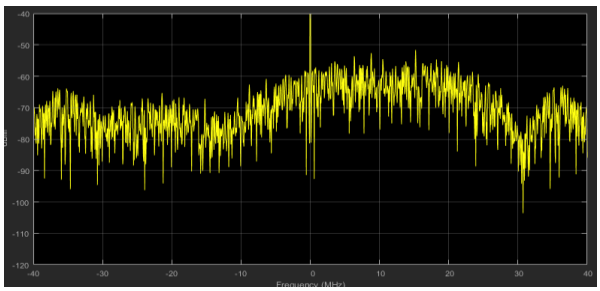


Figure 10: IEEE 802.11 AC multipath fading analysis before FSO.

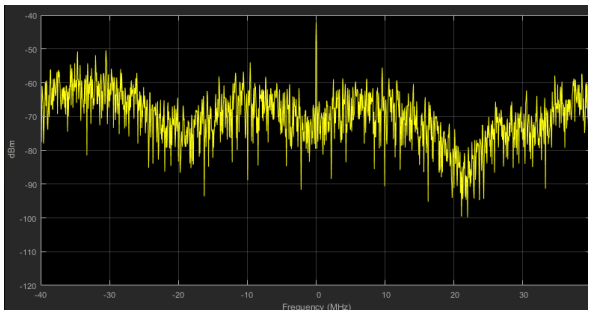


Figure 11: IEEE 802.11 AC multipath fading analysis after FSO.

Multipath fading analysis before Free Space Optics is shown in Figure 10, and the Figure 11 shows multipath fading analysis after Free Space Optics. Here in the Figure 10, it has been observed that as the path loss and shadowing applied on the signal, the average mean received power spread of signal is between – 50 dBm to -90 dBm and the average spread of power across the spectrum when signal passed through Free Space Optics is -50 dBm to -100 dBm. This is due to dispersion and attenuation in multipath due to out of phase signal multipath signal received at receiver.

4. CONCLUSION

In this research, we design and analyzed an integrated Hybrid Passive Optical Wireless Access network and ultra-dense wavelength division multiplexing by applying data transfer capacity efficient and scattering tolerant non optical single sideband carrier less modulation for both downstream and upstream transmission. System performance such as Q factor, Eye diagram and power are calculating and improved. The finding show adequate and improved results. A bidirectional Hybrid OFDM UDWDM PON system design detected by using APDs diodes. APD has very high sensitivity as compare to other diodes like PIN or other optical receivers. The designed network positively provide the coverage range for 100 km for both upstream and downstream at symmetrical data rate of 10Gbps. The quality factor observed for both upstream and downstream lies between 6 to 6.7 range, the average bit error rate lies within 1.5 e^{-10} and path loss , shadowing reported at the spectrum when signal passed through Free Space Optics is -50 dBm to -100 dBm. The reason for downstream data stream provide better results than upstream data stream due to losses added during re modulation of carrier signal. It has been observed that results improved at high power level, employing non optical carrier single side band modulation and at small coverage.

REFERENCES

[1] J. Jin, J. Gubbi, T. Luo, and M. Palaniswami, “Network architecture and QoS issues in the internet of things for a smart city,” in Proceedings of the 2012 International Symposium on Communications and Information Technologies, ISCIT 2012, pp. 956–961, IEEE, Gold Coast, QLD, Australia, October 2012.

[2] M. Chakkour, O. Aghzout, B. Ait Ahmed, F. Chaoui, and M. El Yakhloufi, “Chromatic Dispersion Compensation Effect Performance Enhancements Using FBG and EDFA-Wavelength Division Multiplexing Optical Transmission System,” International Journal of Optics, vol. 2017, 2017.

[3] A. Cimmino et al., “The Role of Small Cell Technology in Future Smart City Applications,” Transactions on Emerging Telecommunications Technologies, vol. 1, pp. 06–19, 2012.

[4] S. P. Singh, S. Iyer, S. Kar, and V. K. Jain, “Study on Mitigation of Transmission Impairments and Issues and Challenges with PLIA-RWA in Optical WDM Networks,” Journal of Optical Communication, De Gruyter, vol. 33, no. 2, pp. 83–101, 2012.

[5] S. Iyer and S. P. Singh, “Spectral and power efficiency investigation in single- and multi-line-rate optical wavelength division multiplexed (WDM) networks,” Photonic Network Communications, vol. 33, no. 1, pp. 39–51, 2017.

[6] V. Reddy and L. Jolly, "Radio over Fiber (RoF) Technology an Integration of Microwave and Optical Network for Wireless Access," in Proceedings of the International Conference and Workshop on Emerging Trends in Technology (ICWET 2015), Bali, Indonesia, 2015.

[7] A. Sharma and S. Rana, "Comprehensive Study of Radio over Fiber with different Modulation Techniques – A Review," International Journal of Computer Applications, vol. 170, no. 4, pp. 22–25, 2017.

[8] N. Singh and H. Kaur, "A Review on Radio over Fiber Technology with Its Benefits and Limitations," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, no. 7, 2016.

[9] N. M. Kassim, "Recent trends in radio over fiber technology," Includes index ISBN 978-983-52-0671-9 First Edition, 2008.

[10] D. Novak, R. B. Waterhouse, A. Nirmalathas et al., "Radio Over-Fiber Technologies for Emerging Wireless Systems," IEEE Journal of Quantum Electronics, vol. 52, no. 1, pp. 1–11, 2016.

[11] S. Singh et al., "Optimization and simulation of WDM-RoF Link," International journal of scientific and research publications, vol. 2, no. 1, pp. 2250–3153, 2012.

[12] S. Jain and B. Therese A, "Four Wave Mixing Nonlinearity Effect in WDM Radio over Fiber System," International Journal of Scientific Engineering and Technology, vol. 4, no. 3, pp. 154–158, 2015.

[13] M. Khatib, "Contemporary Issues in Wireless Communication", e- book (Intech), 2014.

[14] N. Sklavos, M. Hübner and P.G. Kitsos, "System-Level Design Methodologies for Telecommunication", Springer, 2013.

FAKE REVIEW DETECTION ON AMAZON DATASET USING CLASSIFICATION TECHNIQUES IN MACHINE LEARNING

Parminder Kaur[#1], Navroz Kaur Kahlon[*2], Priyanka Jarial[#3]
[#]*Department of computer science and engineering, Punjabi university Patiala, India*
[1]First.parmjassi35@gmail.com
[2]Second.kahlon.navroz3@gmail.com
[3]third.jarial.priyanka@gmail.com

ABSTRACT-With the growth of the internet, e-commerce has come up as a huge platform for marketers. It has proved very successful also, as many people want to buy products online. Before buying online products buyers check the quality of products by reading reviews on products. Amazon.com is one of the largest electronic commerce websites in the world from which people purchase different products and give their reviews on them. These reviews ensure the quality of the product for buyers and decide whether to buy it or not. But some customers write fake reviews to promote or demote the product. These fake reviews mislead customers and degrade the reputation of the product. These fake reviews must be detected to grade the reputation of the product in e-commerce. In this paper, we use supervised learning techniques to detect fake reviews.

KEYWORD: - E-commerce, fake reviews, Weka tool, Supervised learning

I. INTRODUCTION

With the innovation of the web and availability of wireless network access higher speed, increase to developing countries to use e-commerce for getting product and services (karim, 2020). With the growth of internet technology, people prefer the online mode of marketing. E-commerce is a rapid growth area. Generally, e-commerce websites provides a choice to its customers to share their experience about particular service or product in the form of reviews (Hatwar et al., 2019). These reviews provided by the customers are a good source of information because future customers can check these reviews before buying a product. Reviews shows a great effect on the decision-making process. For example, companies make decisions about which product is best for customers. As a result, online reviews on customers' decision-making process in e-commerce (Dowari et al., 2020).

Before buying a product, reading product reviews become a habit for customers. If the review is positive they want to buy that product, otherwise, if the review is negative, they tend to buy other products. Positive reviews become financial benefits for their business growth that improve the accuracy of their product (Elmogy et al., 2021). Many of the companies hire professionals to give positive reviews to promote their products and negative reviews to demote products of their competitors (AshwiniMC & PadmaMC, 2020).

A fake review can be written by bots, it can be written by someone who has not used the product or service. Fake reviews give a large impact on the product or service (Wahyuni & Djunaidy, 2016). It called spam, deceptive reviews. Fake reviews can be detected by sentiment analysis. Fake review classified into three types. Type1 (untruthful opinions): opinions that give undeserving positive reviews or giving harmful negative reviews. Type2 (reviews on brand only): opinion that given not on the product but promote only the brands, manufacturer. Type 3 (Non-reviews): Having two sup-parts one is advertisement and the other is irrelevant reviews that have no opinions. Fake reviews are also called spam that is classified as single spam or group spam (P. Jain et al., 2019).

II. RELATED WORK

(Wahyuni & Djunaidy, 2016)iterative computation framework (ICF) method for measuring the value of review using text mining and opinion mining. The limitation of this method is that the same process needs to be optimized because it detects fake reviews in a short amount of time. (Tadelis, 2016), reputation and feedback played an important role in the online marketplace from recent research, eBay data proposes effective percent positive (EPP) that measure the seller's true quality. In future research will be applied with amazon and yelp dataset showing the distribution of rating because Uber and eBay only show the average score of reviews. (Narayan et al., 2018)mentioned three sets of features, LIWC, POS, and n-gram from different approaches. Online fake review detection is an open research area to be explored with more features, linguistic inquiry, and word count, and uni-gram along with sentiment score as features are combined in this work. Logistic regression classifier recorded with an accuracy of 86.25%.(A. G. E. Elmurngi, 2017) this statistical technique is used to evaluate the performance for fake reviews detection which increases the vulnerability in the reputation systems. Analyzes the dataset of a movie review by text classification using the weka tool. They use five classification algorithms like Decision Tree, naïve Bayes, SVM, logistic regression, KNN, etc. Uses parameters without stop words and with stop words to detect fake reviews. Supervised learning techniques have the best accuracy of 81.4% with the SVM algorithm. (Shashank Kumarr Chauhan, 2017)sentiment analysis of review technique for fake review detection. This method calculates the sentiment score of the review. (G. Jain et al., 2017) sentiment analysis plays an important role in making business decisions about products/services. This study proposes two machine learning algorithms for detecting fake reviews on amazon product reviews is naïve Bayes and SVM classifier. The drawback of research work is feature weighting in classifiers. (Dowari et al., 2020) Decision Tree-j48, Logistic regression, Naïve Bayes, and SVM algorithms for fake review detection. It is tested using the datasets of reviews for Amazon products. They detect number of unfair reviews which include unfair negative reviews and unfair neutral reviews during detection process using WEKA tool. (E. I. Elmurngi & Gherbi, 2018)a framework that deals with both labeled and unlabeled data. This framework obtained an accuracy of 90.19% with supervised learning and 83.70%. Big data techniques need to be explored. (Rout et al.,

2018)presents issue in the context of product reviews. Analysis of amazon product data from which reviews are detected which are duplicate or near-duplicate identifying three types of a fake review. (Jindal & Liu, 2008) supervised learning method on a large scale amazon used to polarize those reviews which were unlabeled. Support vector machine of 10 fold provides better accuracy. (Tanjim UI Haque, Nudrat Nawal Saber, 2018)various techniques used for identifying fake review detection. They identify characteristics and challenges in fake review detection. Current challenges need to be addressed briefly. (Holla, 2018)surveyed that consumer reviews are important in the field of e-commerce. They analyze current research on fake review detection and identify the characteristics, strengths for further improvements. They used different datasets like Amazon, Yelp.com, and Tripadvisor.com, etc. The major drawback is the lack of labeled data for supervised learning techniques. (Vidanagama et al., 2020)(Marriwala et al., 2020)summarize the existing dataset and methods, firstly described fake review and its three types, Type1, Type2, and Type3. After that analyzed from two methods: traditional statistical methods and neural network models. They used different features like semantic, linguistic features, meta-data, behavior features, text features, etc. For fake review detection, different methods are used like supervised learning, semi-supervised learning, neural network models, etc. The major drawback in this research area is the lack of a golden dataset. Large-scale data are difficult to analyze. Mostly the task is focused on traditional statistical methods, which should be explored with neural network models. (Ren & Ji, 2019) semi-supervised and supervised text mining models to detect fake online reviews. Gold standard dataset developed by Ott et al from the Chicago area. The fake reviews are generated from AMT and the remaining from online review sites like Tripadvisor.com, Yelp.com, Expedia and Hotels.com, etc. (Islam, 2019) a model for fake review detection i.e. IP address, account, and negative word dictionary using senti-strength. Tracking IP address to detect fake reviews from the dataset. The purpose of a new algorithm to detect fake reviews to gain more accurate results and to get filtered data. (P. Jain et al., 2019) the study of fake reviews, spotting fake reviews from group reviews is easier than spotting a single fake review. They proposed a methodology with the scoring algorithm by analyzing a dataset of groups using SVM classifiers. They analyze behavioral features for finding fake reviews. (Hatwar et al., 2019)a feature framework for analysis of a dataset from social sites. They present two types of features, review-centric and user-centric features. They detected fake reviews of consumer electronics using features. (Barbado et al., 2019)text analytics to analysis and depth in the study of data with opinion relevant to the subject from which the data is obtained. A baseline model is used to evaluate the data of text review.(Kaur, 2019)mentioned four machine learning methods for finding fake yelp reviews. They implement algorithms like Logistic Regression, Support Vector Machine, Gaussian Naïve Bayes, and XGBoost. XGBoost showed a high score in prediction. There is a limitation in the dataset. (Fong, 2019) surveyed about fake reviews, types of fake reviews. They inform the current research situations, implications, and limitations of fake reviews. They propose an antecedent, consequence, intervention. Conceptual framework to detecting fake reviews.(Wu et al., 2020)mentioned product reviews and identifying fake reviews, types of fake reviews. By using opinion mining fake reviews can be analyzed and detected efficiently. The proposed system uses three different classifier Decision Tree, support vector machine, Naïve Bayes, etc. from this SVM performs best as compared to others on the amazon dataset. (AshwiniMC & PadmaMC, 2020)survey machine learning and deep learning techniques for fake review detection. Machine learning techniques which are used most were NB, SVM classifier and Naïve Bayes model has the highest accuracy 90.423% used by Satuluri Vanaja. Deep learning techniques are of RNN AS LSTM, Bi-LSTM, and GRNN. LSTM model has the highest accuracy of 98.9% used by Y, Wang, and J. Zhay. (Rodrigues et al., 2020) presents three machine learning algorithms Naïve Bayes, SVM, and KNN. Twitter data need to detect as true or fake. The author evaluates twitter data with Naïve Bayes by giving 97% accuracy, SVM is 98%, and KNN is 90% of accuracy. SVM is the best algorithm for fake review detection. (Huy le, 2020) make use of ML models like Naïve Bayes, SVM, and decision tree for opinion mining. They use hotel reviews. Data is divided into 80% of training data and 20% testing data. They present an opinion mining algorithm for supervised learning of hotel reviews. (karim, 2020)Mentioned the importance of reviews. Machine learning techniques are used for fake review detection. They propose features of reviews like behavioral features. They study yelp dataset using logistic regression, NB, RF, KNN, SVM classifiers for detecting fake reviews. (Huy le, 2020) mentioned the goal of fake review detection. They used a classification method to classify fake reviews with three classification algorithms, support vector machine, ANN, Random forest. A clustering model is used to identify the hidden patterns of a fake reviews. They use non-text features to solve fake review problems. In the Table 1 we make a comparison of different studies.

Table 1: Parametric Comparison of Previous Study on ML techniques

Author	Dataset	Strength	Limitations	Highest Accuracy
(A. G. E. Elmurngi, 2017)	Movie reviews	Improved classification accuracy with two methods	Should be focus on incompatibility and manipulation issues	SVM-81%
(Rout et al., 2018)	Different review sites	Dealing with huge amount of data	Big data technique need to be explored	NB-90.19%
(Islam, 2019)	Hotel reviews	Improving efficiency with text mining techniques	Multilingual analysis need to be explored	NB-86.32%

(Hatwar et al., 2019)	Amazon.com	Automatically classifiers input text data into fake or not fake review	To investigate different kinds of feature selection methods	SVM-81.92%
(Fong, 2019)	Yelp.com	Get filtered reviews from yelp	Lack of some dataset as trust factor, user profile	XGBoost-99%
(AshwiniMC & PadmaMC, 2020)	Amazon.com	Automatic system to classify fake or genuine reviews.	Content aware classification require to identify sarcasm	NB-90%
(karim, 2020)	Hotel reviews	Classify two-class problem like positive and negative	User different feature selection methods	RF-92%
(Elmogy et al., 2021)	Hotel reviews	Extract feature behavior like TF-IDF using bigram and tri-gram	Should be focus on other behavioral features like frequent times, time reviewer etc.	KNN-83%

III. METHODOLOGY

We analyze amazon reviews from a dataset on kaggle.com. The dataset of amazon product reviews is labelled as label1 and label2 as fake review and genuine review respectively. The dataset contains a total of 1600 reviews of different product reviews of which 50% are fake reviews and 50% are real reviews. The dataset contains different attribute like doc-id, label, rating, verified purchase, product id, product category, product title, review title, review text. To carry out this study a predefined process is followed. The process followed is shown in figure 1:

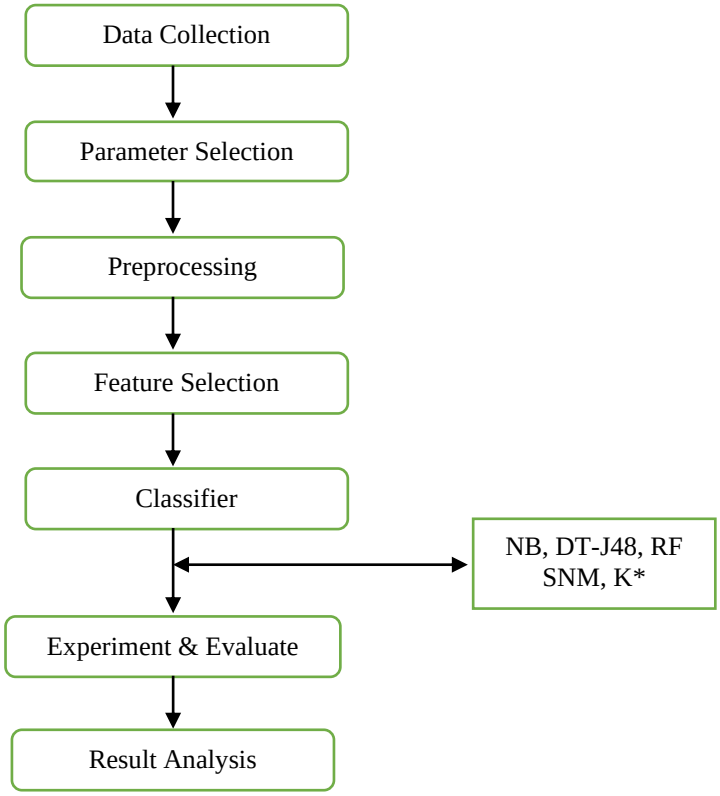


Fig 1: Proposed methodology

3.1Data collection

The first is to collect data for fake review detection process. Recent researchers used online data from different review sites. It is more difficult to find a labelled dataset. We collect the labelled amazon dataset kaggle.com. You can find the full dataset of amazon review from kaggle.com.

3.2 Parameter selection

After collecting the data the next step is to select parameters from collected dataset. We used amazon reviews dataset publically available from kaggle.com. This dataset contains 1600 instances and each instance consists of 9 attributes. This dataset is collected in .CSV file. We describe parameter, description and value as shown in Table2.

Table 2: Parameters selection for fake review dataset

Parameter	Description	Value
Doc Id	Document id number	Numeric
Label	Label of data	{label1, label2}
Rating	Range from 1to 5	{1, 2, 3, 4, 5}
Verified Purchase	To verify a product	{Y/N}
Product Id	Id of product	String
Product Category	Category of product	String
Product Title	Title for the product	String
Review Title	Title for the review	String
Review Text	Text to be written	String

3.3 Preprocessing

This is the next step, after selecting the data from available dataset. Preprocessing means removing stop words and punctuations, removing commas etc. We use WEKA tool for implementing the classifier.

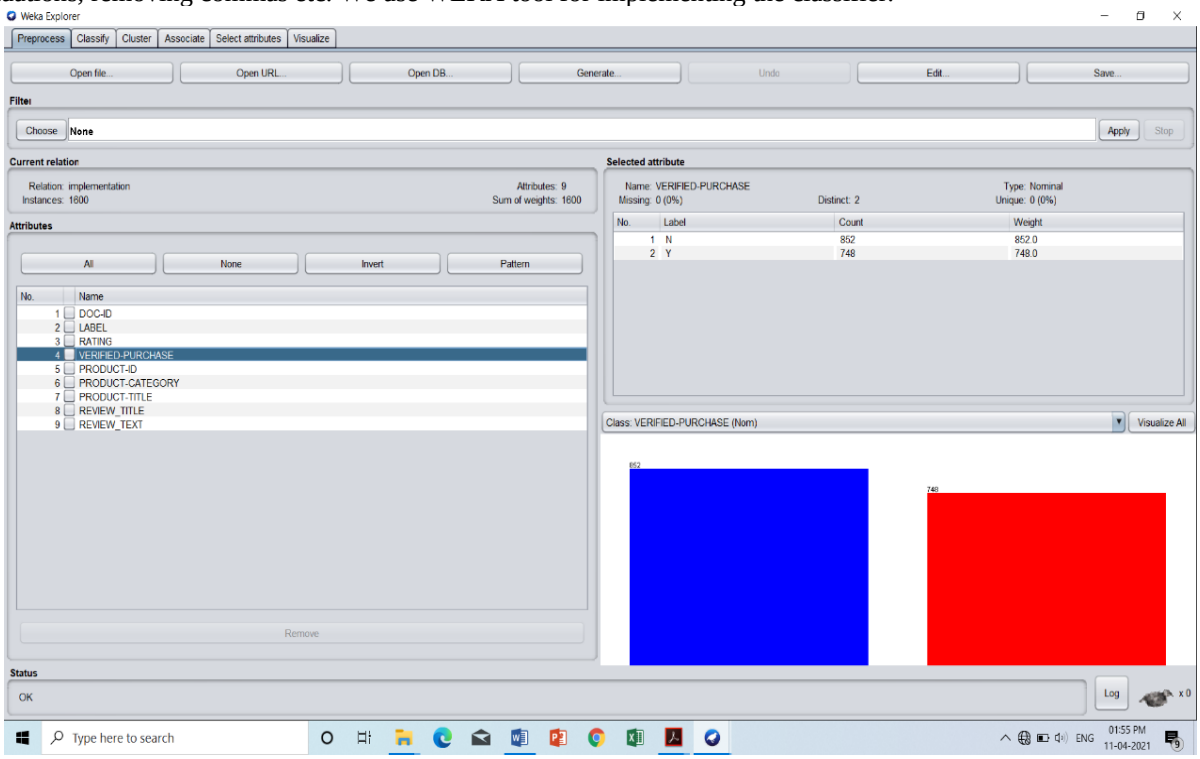


Fig 3: Output of preprocessing

3.4 Feature extraction

From the recent researches, researchers use review centric features and reviewer centric features. We take some features to identify fake reviews like rating, verified purchase.

3.5 Classification

In this step, we apply classification techniques to detect fake reviews from amazon dataset. In this paper, we will apply four supervised classification techniques: NB, DT-J48, and RF. Fig 4, 5, 6, 7, and 8 shows classifier below.

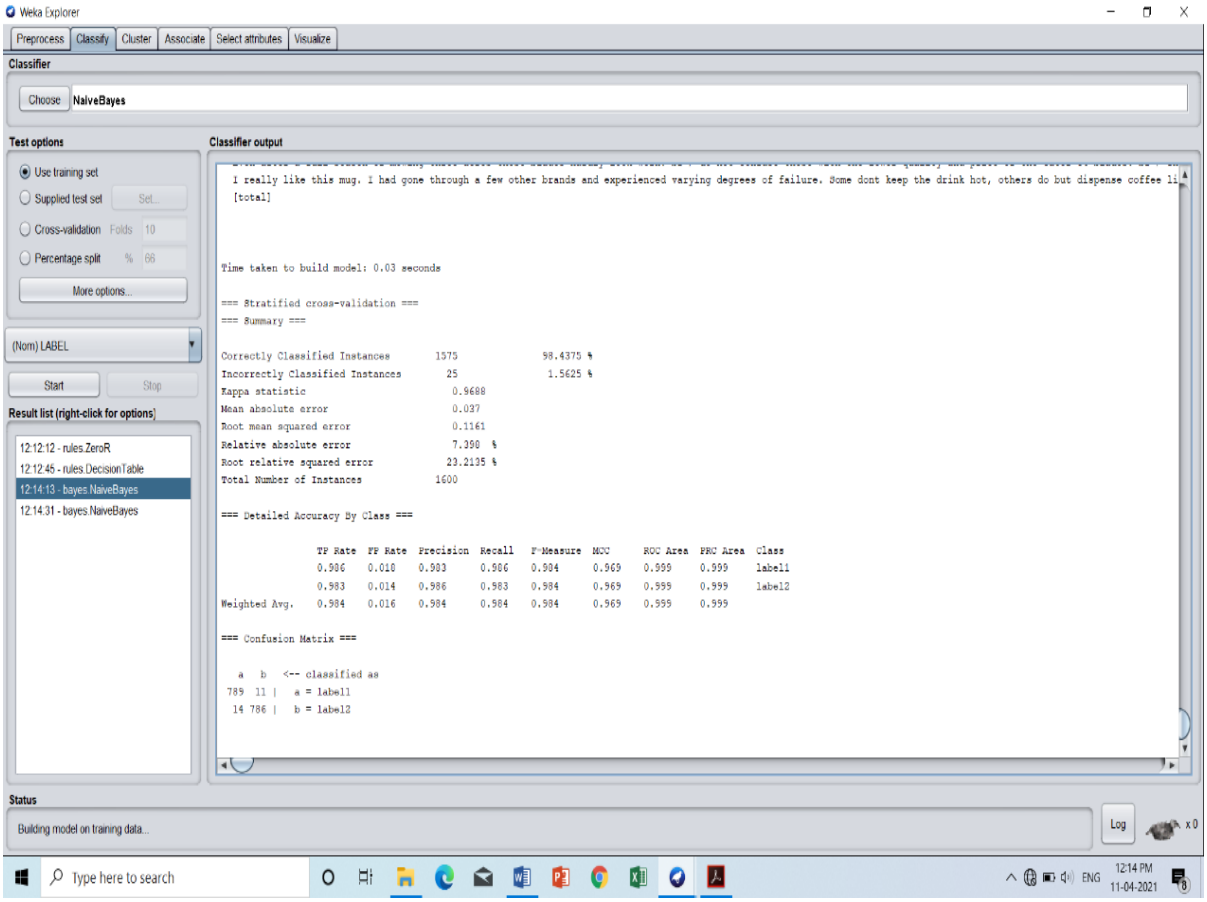


Fig 4: Output of naïve Bayes

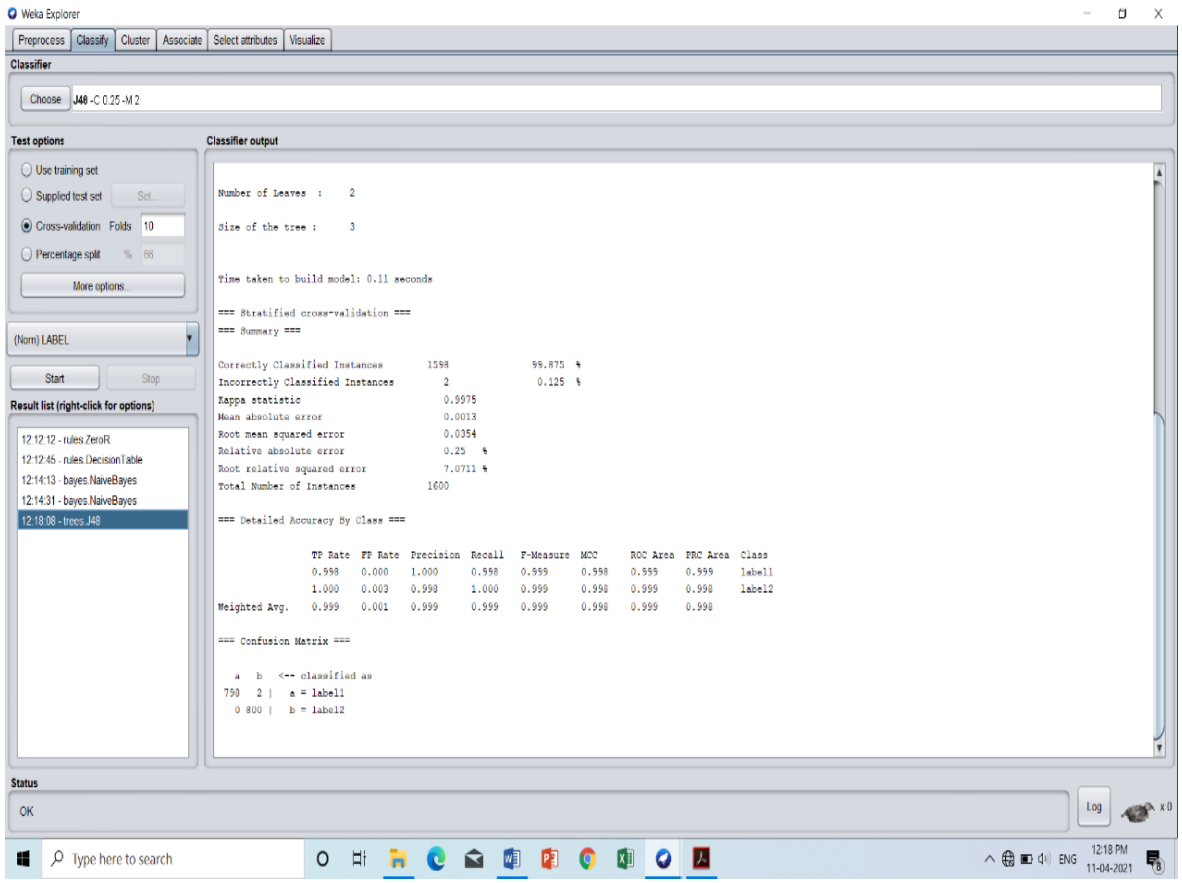


Fig 5: Output of J48

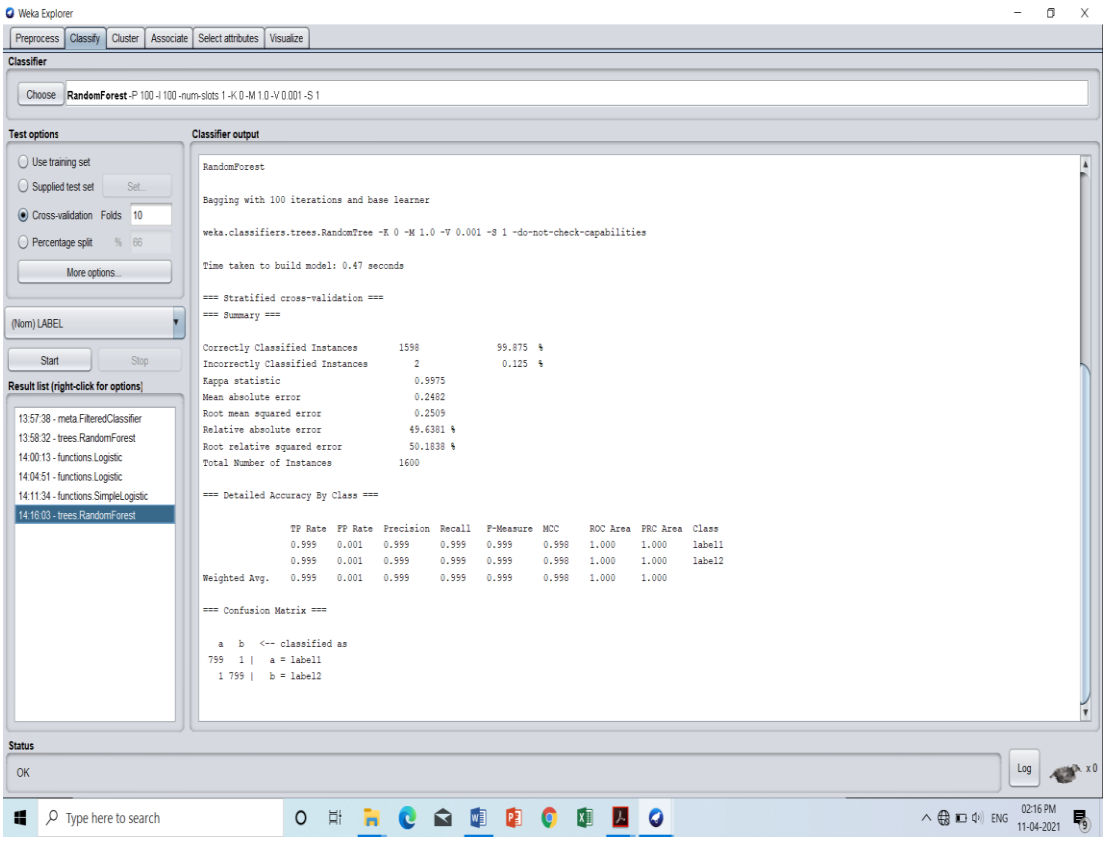


Fig 6: Output of random forest

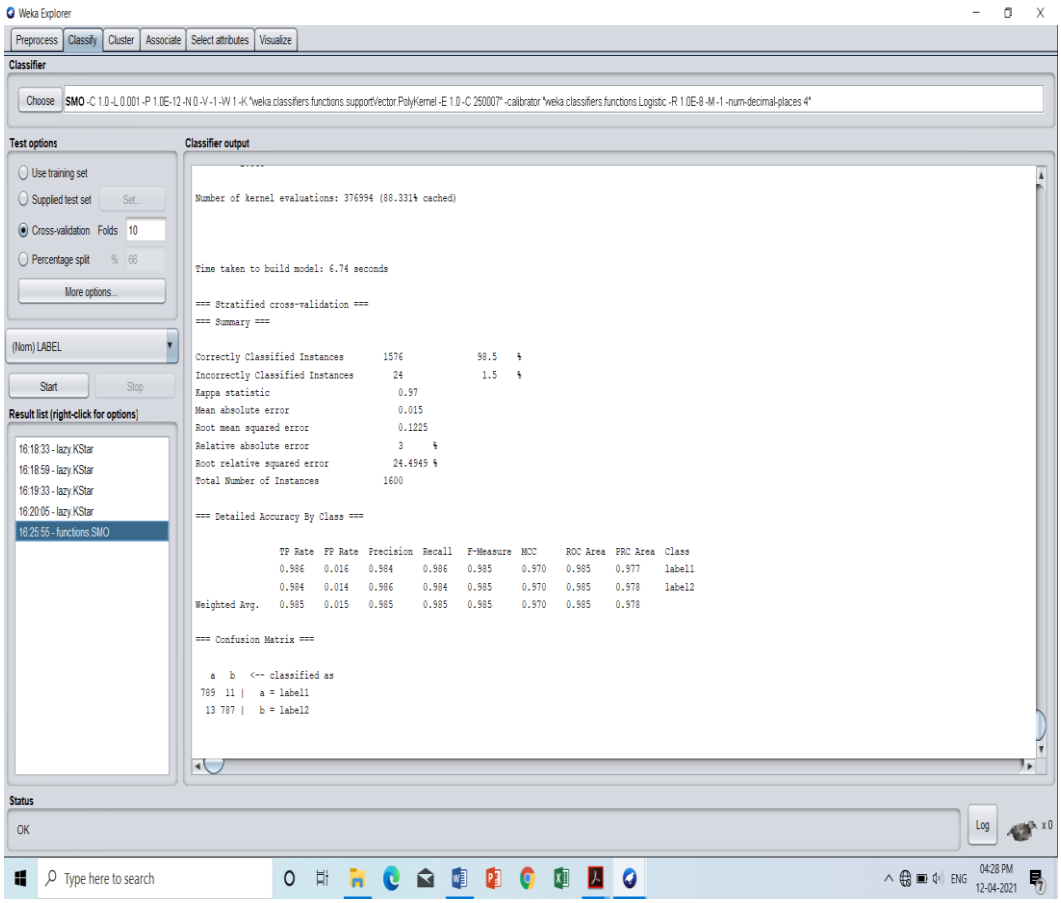


Fig 7: Output of SVM

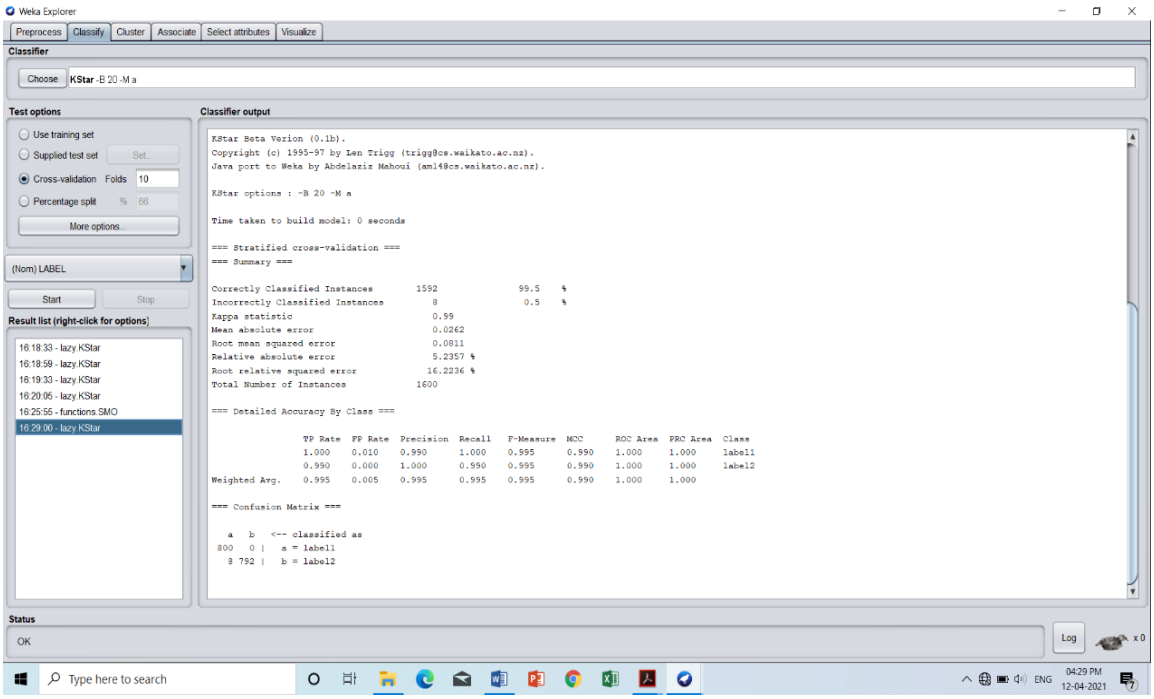


Fig 8: Output of k- star

IV. EXPERIMENTS AND EVALUATION

The proposed methodology has evaluated with 1600 reviews from amazon dataset. The output of all the classification techniques in previous section. Now, in this section we describes the experiments in our research. We present experiment from supervised classification techniques to classify fake and real reviews from our dataset which is compared with verified purchase and without verified purchase methods.

4.1Without verified purchase

4.1.1 Confusion matrix for all classification algorithm

We compare the number of real and fake reviews using classification algorithm with actual results as shown in confusion matrix Table without verified purchase. We get confusion matrix after implementing NB, DT-J48, RF, SVM, K* algorithms. Table 3 shows confusion matrix for Amazon review dataset.

Table 3: Confusion matrix without verified purchase

Classification Algorithms	Prediction	Real	Fake
Naïve Bayes	Real	789	11
	Fake	14	786
DT-J48	Real	798	2
	Fake	0	800
RF	Real	799	1
	Fake	1	799
K*	Real	800	0
	Fake	8	792
SVM	Real	**789**	**11**
	Fake	**13**	**87**

4.1.2. Evaluate and compare accuracy of all classifiers

We evaluate the performance of five classification algorithms. First we get a confusion matrix with real and fake reviews. In Table 4 we make a comparison of accuracy as described below. We obtain higher accuracy without verified purchase of SVM is 99.5%.

Table 4: Comparison of Accuracy of all classifiers

Classification Algorithms	Accuracy
NB	98.43%
DT-J48	99.47%
RF	99%
K*	99.45%
SVM	**99.5%**

Table 5: Time taken to build a model

Classification Algorithms	Time taken to build model(seconds)
NB	0.03
DT-J48	0.11
RF	1.01
K*	0.25
SVM	**1.74**

Table 5 shows the time taken to build a model using five algorithm like NB, DT-J48, RF, K*, and SVM. From this Table K* takes the shortest amount of time 0 seconds and SVM has longest amount of time is 6.74 seconds to build a model.

4.2 With verified purchase

4.2.1 Confusion matrix for all classification algorithm

In this section we compare the number of real and fake reviews using different algorithms with verified purchase. Table 6 shows confusion matrix with verified purchase.

Table 6: Confusion matrix with verified purchase

Classification Algorithms	Prediction	Real	Fake
Naïve Bayes	Real	607	245
	Fake	204	544
DT-J48	Real	604	248
	Fake	196	552
RF	Real	847	5
	Fake	716	32
K*	Real	618	234
	Fake	213	535
SVM	Real	**636**	**216**
	Fake	**205**	**543**

4.2.2 Evaluate and compare accuracy of all classifiers

We evaluate the performance of five classification algorithms. In Table7 we make a comparison of accuracy as described below. We obtain higher accuracy with verified purchase of SVM is 73.687%.

Table 7: Comparison of Accuracy of all classifiers

Classification Algorithms	Accuracy
NB	71.935%
DT-J48	72.25%
RF	54.93%
K*	72.0625%
SVM	**73.687%**

Table 8 shows the time taken to build a model using five algorithm like NB, DT-J48, RF, K*, and SVM. From this Table K* takes the shortest amount of time 0.12 seconds and SVM has longest amount of time is 2.03 seconds to build a model.

Table 8: Time taken to build a model

Classification Algorithms	Time taken to build model(seconds)
NB	0.01
DT-J48	0.01
RF	1.55
K*	0.12
SVM	**2.03**

V. RESULT ANALYSIS

In the last step we calculate the model to understand the performance. We make a comparison of performance without verified purchase and with verified purchase. We have calculated precision, recall and f-measure to compare the performance of all classifier.

Table 9: Performance comparison of all classifier without verified purchase

Classification Algorithms	Precision	Recall	F-measure
Naïve Bayes	0.984	0.984	0.94
DT-J48	0.99	0.99	0.99
RF	0.99	0.99	0.99
K*	0.985	0.985	0.985
SVM	0.995	0.995	0.995

Table 10: Performance comparison of all classifier with verified purchase

Classification Algorithms	Precision	Recall	F-measure
Naïve Bayes	0.721	0.719	0.720
DT-J48	0.725	0.723	0.723
RF	0.693	0.549	0.412
K*	0.73	0.73	0.73
SVM	0.721	0.721	0.721

Performance comparison of all classifier without verified purchase and Table 10 shows performance comparison of all classifier with verified purchase.

From the comparison of Table 9 and Table 10, SVM is suitable algorithm by accuracy for all experiment with verified purchase and without verified purchase. Table 10 shows performance comparison of all classifier with verified purchase. SVM is best method as comparison with other classifiers.

VI. CONCLUSION

The study is taken into account to understand and analyze the various techniques used to analyze the reviews or fake review detection. In this study, two parameters, without verified purchase and with verified purchase are used. Study presents classification algorithms to apply a dataset of Amazon reviews using supervised learning techniques. We use five classification algorithms like NB, DT-J48, RF, SVM, and K*. After comparing the performance of all classifiers, analyzed that SVM is more accurate than another classifier because it has higher accuracy in both cases without verified purchase and with verified purchase.

In our future work, we can work on other datasets such as Twitter dataset, yelp.com, or Tripadvisor.com.com and with different methods. We would like to extend these work tools such as Python and R or R studio, and then we will evaluate the performance of our work.

VII. REFERENCES

1) AshwiniMC, & PadmaMC. (2020). Efficiently analyzing and detecting fake reviews through opinion mining. In International Journal of Computer Science and Mobile Computing (Vol. 9, Issue 7). www.ijcsmc.com
2) Barbado, R., Araque, O., & Iglesias, C. A. (2019). A framework for fake review detection in online consumer electronics retailers. Information Processing and Management, 56(4), 1234–1244. https://doi.org/10.1016/j.ipm.2019.03.002
3) Dowari, G., jyoti Bora, D., Khyat, J., Aier, T., & Jyoti Bora, D. (2020). Fake Product Review Monitoring and Removal using Opinion Mining. UGC Care Group I Listed Journal, 10(5). www.junikhyat.com
4) Elmogy, A. M., Tariq, U., Ibrahim, A., & Mohammed, A. (2021). Fake Reviews Detection using Supervised Machine Learning. International Journal of Advanced Computer Science and Applications, 12(1), 601–606. https://doi.org/10.14569/IJACSA.2021.0120169
5) Elmurngi, A. G. E. (2017). Seventh International Conference on Innovative Computing Technology (INTECH 2017) : Luton, UK, August 16-18, 2017. The Sevength International Conference on Innovative Computing Technology, Intech, 107–114.
6) Elmurngi, E. I., & Gherbi, A. (2018). Unfair reviews detection on Amazon reviews using sentiment analysis with supervised learning techniques. Journal of Computer Science, 14(5), 714–726. https://doi.org/10.3844/jcssp.2018.714.726
7) Fong, A. S. and A. C. M. (2019). Proceedings of the Fourth International Conference on Contemporary Computing and Informatics (iC^{3}I 2019) : 12-14 December 2019, Amity Global Institute, Singapore.
8) Hatwar, P., Dhepe, S., Fale, S., Gedam, S., & Kumar Minz, N. (2019). E-Commerce Product Rating Based on Customer Review Mining. In International Journal of Engineering Science and Computing. http://ijesc.org/
9) Holla, Al. (2018). A Comparative Study on Fake Review Detection Techniques. International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 5(4).
10) Huy le, B. kim. (2020). Detection of fake reviews on social media using machine learning algorithms. Issues In Information Systems, 07(07), 7. https://doi.org/10.48009/1_iis_2020_185-194
11) Islam, R. H. & M. R. (2019). ECCE 2019 : 2nd International Conference on Electrical, Computer and Communication Engineering (ECCE) : conference digest : 07-09 February 2019, Cox's Bazar, Bangladesh.
12) Jain, G., Jain, G., & Agarwal, B. (2017). Spam Detection on Social Media Text. International Journal of Computer Sciences and Engineering, 5(5). https://www.researchgate.net/publication/322791510
13) Jain, P., Chheda, K., & Lade, M. J. | P. (2019). Fake Product Review Monitoring System. International Journal of Trend in Scientific Research and Development, Volume-3(Issue-3), 105–107. https://doi.org/10.31142/ijtsrd21644
14) Jindal, N., & Liu, B. (2008). Opinion Spam and Analysis. http://money.cnn.com/2006
15) karim, S. (2020). Predicting Fake online Reviews using Machine Learning. International Journal of Scientific Research and Engineering Development, 3, 269–273. www.ijsred.com
16) Kaur, P. (2019). 66. Prabh 2019 (14). Intterrnattiionall JJourrnall off Ellecttrroniicss Engiineerriing ((IISSN:: 0973--7383, 11(1).
17) Marriwala, N., Tripathi, C. C., Kumar, D., & Jain, S. (2020). Lecture Notes in Networks and Systems 140 Mobile Radio Communications and 5G Networks Proceedings of MRCN 2020. http://www.springer.com/series/15179
18) Narayan, R., Rout, J. K., & Jena, S. K. (2018). Mining. Springer, 273–279. https://doi.org/10.1007/978-981-10-3376-6
19) Ren, Y., & Ji, D. (2019). Learning to Detect Deceptive Opinion Spam: A Survey. In IEEE Access (Vol. 7, pp. 42934–42945). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2019.2908495
20) Rodrigues, J. C., Rodrigues, J. T., Gonsalves, V. L. K., Naik, A. U., Shetgaonkar, P., & Aswale, S. (2020, February 1). Machine Deep Learning Techniques for Detection of Fake Reviews: A Survey. International Conference on Emerging Trends in Information Technology and Engineering, Ic-ETITE 2020. https://doi.org/10.1109/ic-ETITE47903.2020.063

21) Rout, J. K., Dash, A. K., & Ray, N. K. (2018). A Framework for Fake Review Detection: Issues and Challenges. Proceedings - 2018 International Conference on Information Technology, ICIT 2018, 7–10. https://doi.org/10.1109/ICIT.2018.00014
22) Shashank Kumarr Chauhan, A. G. (2017). 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN) : 2-3 February, 2017, Amity School of Engineering and Technology, Noida, India.
23) Tadelis, S. (2016). The economics of reputation and feedback systems in e-commerce marketplaces. IEEE Internet Computing, 20(1), 12–19. https://doi.org/10.1109/MIC.2015.140
24) Tanjim UI Haque, Nudrat Nawal Saber, F. M. S. (2018). 2018 IEEE International Conference on Semiconductor Electronics (ICSE2018) : proceedings : 15-17 August 2018, Pullman Kuala Lumpur City Centre Hotel & Residence, Kuala Lumpur, Malaysia. 2018 IEEE International Conference on Innovative Research and Development (ICIRD), May, 85–88.
25) Vidanagama, D. U., Silva, T. P., & Karunananda, A. S. (2020). Deceptive consumer review detection: a survey. Artificial Intelligence Review, 53(2), 1323–1352. https://doi.org/10.1007/s10462-019-09697-5
26) Wahyuni, E. D., & Djunaidy, A. (2016). Fake review detection from a product review using modified method of iterative computation framework. MATEC Web of Conferences, 58. https://doi.org/10.1051/matecconf/20165803003
27) Wu, Y., Ngai, E. W. T., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. Decision Support Systems, 132(February), 113280. https://doi.org/10.1016/j.dss.2020.113280

5G SECURITY: BRIEF ANALYSIS OF THREATS

Dr. Amandeep Singh Bhandari, Dr. Charanjit Singh
Department of ECE, Punjabi University, Patiala, Punjab, India

Abstract- Wireless communication systems and devices have emanated during the past few decades and is continuously evolving. In today's scenario of advanced technologies, the fifth generation wireless system (in short 5G) will play a pivotal role in providing high-quality services as compared to 4G. In order to provide better platform for various applications such as healthcare, banking, education, media, entertainment, public transport, etc. by providing greater data bandwidth and ultra-low latency. To accomplish these goals, a secure architecture of 5G is very much desirable which should be very less vulnerable to security threats and attacks, thereby expected to be designed by taking a user-centric approach. This paper presents immense study about the security aspects related to 5G technology. The paper outset with introducing some basic concepts about 5G networks and its applications. This paper further provides brief survey about the security weaknesses in 5G networks and the suggestions to improve the systems.

Keywords- 5G, security, attacks, M2M, D2D, IoT, AKA.

Introduction- A mobile network is a wireless network spread over the earth through a web of cell sites. Wireless communication technologies are essential parts of our lives as now it is impossible to spend a day without using any of wireless devices. Being tremendous change in wireless mobile communication, 5G will also mitigate the gap between physical and digital world by affecting our lives. Primarily, it aims to provide a complete wireless communication without any limitations. In today's global economic development, mobile communication industry is playing the role of important pedestal, which is growing at very fast speed by changing its characteristics to change people's lifestyle. The drastic growth of use of the Internet was the motivation behind Mobile Broadband. After the great success of fourth generation (4G) wireless systems, it is expected that 5G technology will also provide better efficiency in terms of high bandwidth for better quality of voice and mobile broadband.

The 5G technology consolidates almost all the enhanced benefits of mobile phones, such as high speed dialling, cloud data storage and high definition (HD) downloading at high speed. New radio bands above 20 GHz are being assigned for 5G. This technology anticipates the implementation of high capacity broadband applications and services requiring the gigantic amount of spectrum.

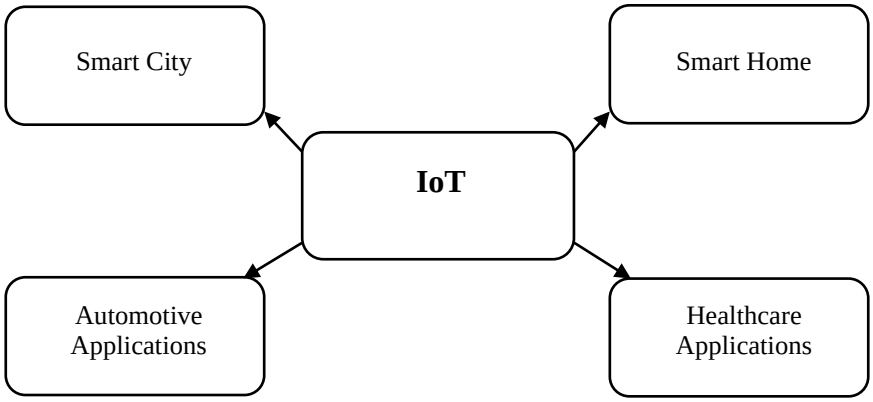


Figure: Application areas of IoT

Therefore, the availability of current spectrum bands is a key obligation for the provision of 5G services. Several bands are conventional for 5G distributed over a large range of frequencies so the air-interface has to be extensible enough to accommodate all such bands.

For the growth of Internet-of-Things (IoT), machine-to-machine (M2M) communication evolve into an important research paradigm that employs 5G wireless systems. The implementation of IoT becomes mandatory nowadays for the vision of smart cities or smart military bases, where the sensors and the ability to make decisions among machines results in no demand of human intervention [1].

In real world, 5G technology is becoming backbone of Automotive industry day-by-day. The 5G technology is assumed to provide smooth services to users in conjunction with supporting the massive number of connected devices. Therefore, in 5G wireless systems the reliability and high availability are the major requirements to be promised by the technology for the success rate of data transmission over a certain period of time.

The main applications of 5G can be classified into enhanced Mobile Broadband (eMBB), massive Machine Type Communication (MTC), Ultra-Reliable Low Latency Connections (URLLC) that needs the support of intense device density, energy efficiency and low-latency potential. MTC is a technology that enables communication among user-end devices and the underlying data transport infrastructure by application of different network technologies such as point-to-point, multi-hop, adhoc networks, etc.

The technologies used in 5G wireless systems can be mainly classified as per their usage in different layers, as massive Multi-Input Multi-Output (MIMO), millimetre wave, full-duplex communication, etc. can be endorsed by physical layer, while Software-Defined Networks (SDN), Network Functions Virtualization (NFV), etc. can be supported by logical layer. MIMO is a fundamental technique in the modern cellular system in which multiple antennas are to be used at both the transmitter and receiver sides to enhance the capacity of uplink and downlink operations.

For the proper development and deployment of 5G wireless technology in IoT environment, the security concerns must be considered, as the massive number of user devices are to be entertained by the 5G systems, thus it will be mandatory to protect the leakage of user personal information. The key security services, such as Confidentiality, Mutual Authentication, Key Forward/Backward secrecy, Integrity remains unchanged over the years. But for IoT infrastructure, lightweight cryptographic measures should be considered so that complexities among M2M communication can be reduced.

In distinction to usage of IoT for 5G wireless networks, the Artificial Intelligence (AI) also plays a vital role in order to provide security solutions to 5G networks from several vulnerable attacks generated by the eavesdroppers. The main motive behind the application of AI in 5G networks is to authenticate and authorize the devices in the network.

Related Work-

In 2020, Anshu Bhardwaj has discussed various security challenges for 5G technologies and also presented the current status about deployment of 5G technology worldwide. The requirement of huge infrastructure alongwith asymmetric resources, i.e., large bandwidth, power requirements are the biggest challenges for 5G technology due to M2M communication in large quantity. In addition to these challenges, security of 5G technology is also of major concern in order to avoid various attacks, such as MiTM, DDoS, etc. Furthermore, the utilization of 5G technology in military communications has been described. [4]

In 2020, Zaher Haddad et al. have proposed an Authentication and Key Agreement (AKA) protocol using Blockchain, the technology behind the ease of secure information distribution over the whole network, for 5G technology. The proposed scheme has also avoided various security threats by considering different security measures. In this scheme, the home network (HN) is more secure due to absence in authentication protocol, which is mainly responsible for initiating of Blockchain technology. [5]

In 2020, Jingjing Zhnag et al. have analysed the security aspects of 5G EAP-TLS authentication protocol applied in pi calculus by employing the ProVerif model checker, which is used mainly for IoT environments. From the analysis, the authors have discussed about several flaws in the security properties of the primary protocol and also suggested some strategies to countermeasure the violence of the security of the protocol. [6]

In 2020, Lili Jiang et al. have developed the continuous-time Markov chain (CTMC) model to analyse the dependability of 5G-AKA authentication when facing the service failure. The model has been designed to evaluate various parameter such as, number of authentication requests not served per million, authentication service's first restoration time from failure of Secure Anchor Function (SEAF) or Authentication Server Function (AUSF) and total cost of ownership (TCO). [7]

In 2020, Ma Jinsong and Mohammad Yamin have discussed about 5G network and its security issues. To provide high efficiency by increasing the speed of data transfer among devices, various aspects of 5G technology must be considered alongwith protecting the data transfer from third parties' attacks. To accomplish these goals, lightweight security algorithms and protocols should be employed in order to reduce severe load on storage devices and also mitigates the computational complexity further. [8]

In 2020, Dinh C. Nguyen et al. have surveyed about integration of Blockchain with 5G wireless systems and beyond. The Blockchain technology is a decentralized and immutable which is used to manage and enhance the capabilities of 5G networks for efficient data sharing for M2M communication. In this paper, the authors have also observed and elaborated many research challenges for 5G networks for its application in IoT environment to provide advanced techniques for smart grids, smart vehicles, smart cities, etc. [9]

In 2020 Misbah Shafi et al. have discussed about security issues and the countermeasures related to 5G New Radio (NR) that provides several services like M2M communication, spectrum sharing, etc. A system model has also been designed to observe the effects of artificial rain (AR) and artificial dust (AR) over the security of wireless communication networks. With the passage of time, vulnerable breaches in 5G networks are increasing resulting in risks of more attacks on downlink specifically to get the allocated resources from the authenticated users. [10]

In 2019, Ikram Gharsallah et al. have proposed a secure efficient and lightweight (SEL)-AKA protocol to overcome the flaws in fundamental 5G-AKA protocol as it is vulnerable to many threats such as Denial of Service (DoS) attack, linkability attack due to authentication failure, replay attack. The proposed protocol has successfully achieved mutual authentication and key agreement among user equipments and also prevented the aforementioned security threats alongwith verification using Security Protocol Animator (SPAN) software. [11]

In 2019, Jingjing Zhang et al. have analysed the security of 5G-EAP-TLS authentication protocol designed for IoT environment using Scyther model checker. In this process, the requirement of authenticating the subscribers by the home network and serving network has been verified. Many flaws in 5G-EAP-TLS protocol have been discovered and reported, so that the protocol is not suitable to be implemented in real systems. [12]

In 2019, AN Braeken et al. have proposed a modified 5G-AKA protocol by using random numbers instead of sequence numbers to avoid replay attacks significantly and reducing the communication phases by employing asymmetric encryption to overcome the flaws in the basic protocol. The proposed protocol has outperformed the existing protocol in terms of better forward security, better computation and communication efficiency and post-compromise security. [13]

In 2019, Ahmed A. Abd El-Latif et al. have proposed two efficient hash functions based on quantum walks (QWs), namely – QWHF-1 and QWHF-2 used to generate the cryptographic keys to secure data in 5G networks. These two hash functions have been developed to introduce an Authenticated Key Distribution (AKD) for efficient data sharing using proper encryption and an Authenticated Quantum Direct Communication (AQDC) for device-to-device (D2D) communication thereby providing protection against active well-known attacks in 5G networks. [14]

In 2018, Ijaz Ahmad et al. have discussed about security challenges in 5G networks in terms of concepts used at logical layer such as SDN, NFV, and Mobile Edge Computing (MEC). SDN is more vulnerable to DoS attacks, NFV is vulnerable to many threats such as side-channel attacks, flooding attacks, Virtual Machine (VM) migration related attacks, cloud specific attacks, and MEC is vulnerable to DoS attack, Man in The Middle (MiTM) attack, privacy leakages, and VM manipulation. The security solutions to these security challenges have also been discussed in this paper. [15]

In 2018, Dongfeng Fang et al. have discussed about security flaws in 5G wireless networks and also proposed a new architecture to provide flexible authentication and better firmness for protection against security threats. Efficient authentication is very much necessary in 5G networks to fulfil the requirement of low latency using SDN. A handover procedure has also been studied in this paper to verify the performance of the proposed security architecture. [16]

In 2017, Bin Han et al. have discussed about trust zone drafted for the intensification of authorization, authentication and accounting (AAA) for 5G networks in edge cloud to manage security administration and database accesses. Trust zone can mainly be classified as Connected, Weakly connected, Lost connection, Reconnecting and Disconnecting states, among which first three are steady states while remaining two are transient states. [17]

Conclusion- Nowadays, millions of devices communicating with each other by taking the advantage of internet technology, the concepts or disciples of M2M and MTC are highly researched. The direct adoption of the concepts of M2M communication in various applications with the use of several advanced technologies such as Blockchain and AI for the IoT environment also give rise to different challenges. The most prominent one being the security of the networks. In this paper, various threats for 5G wireless systems incorporating recent technologies are discussed. Literature study shows that many threats have been detected so far by various researchers and also the many suggestive concepts have been applied to mitigate the unauthorized access due to several attacks. By implementing these concepts, the forward secrecy and mutual authentication has been improved.

REFERENCES

[1] Madhusanka Liyanage, Ijaz Ahmad, Ahmed Bux Abro, Andrei Gurtov, and Mika Ylianttila, “A Comprehensice Guide to 5G Security”, John Wiley & Sons Ltd.

[2] Dushantha Nalin K, Jayakody Kathiravan Srinivasan, Vishal Sharma, “5G Enabled Secure Wireless Networks”, Springer.

[3] Erik Dahlman, Stefan Parkvall, Johan Skold, “5G NR: The Next Generation Wireless Access Technology”, Academic Press, Elsevier.

[4] Anshu Bhardwaj, “5G for Military Communications”, Third International Conference on Computing and Network Communications (CoCoNet’19), Elsevier, Procedia Computer Science 171(2020), pp 2665-2674.

[5] Zaher Haddad, Mostafa M. Fouda, Mohammed Mahmoud, and Mohammed Abdallah, “Blockchain-based Authentication for 5G Networks”, IEEE, 2020, pp 189-194.

[6] Jingjing Zhang, Lin Yang, Weipeng Cao, and Qiang Wang, “Formal Analysis of 5G EAP-TLS Authentication Protocol Using ProVerif”, IEEE Access, Volume 8, 2020, pp 23674-23688.

[7] Lili Jiang, Xiaolin Chang, Jing Bai, Jelena Misic, Vojislav Misic, and Zhi Chen, ‘Dependability Analysis of 5G-AKA Authentication Service from Server and User Perspectives”, IEEE Access, Volume 8, 2020, pp 89562-89574.

[8] Ma Jinsong and Mohammad Yamin, “5G Network and Security”, 7th International Conference on Computing for Sustainable Global Development (INDIACom), IEEE, 2020, pp 249-254.

[9] Dinh C. Nguyen, Pubudu N. Pathirana, Ming Ding, and Aruna Seneviratne, “Blockchain for 5G and beyond networks: A state of the art survey”, Journal of Network and Computer Applications 166 (2020), Elsevier, pp 1-38.

[10] Misbah Shafi, Rakesh Kumar Jha, and Manish Sabraj, “A survey on security issues of 5G NR: Perspective of artificial dust and artificial rain”, Journal of Network and Computer Applications 160 (2020), Elsevier, pp 1-25.

[11] Ikram Gharsallah, Salima Smaoui, and Faouzi Zarai, “A Secure Efficient and Lightweight authentication protocol for 5G cellular networks: SEL-AKA”, 15th International Wireless Communications & Mobile Computing Conference (IWCMC), IEEE, 2019, pp 1311-1316.

[12] Jingjing Zhang, Qiang Wang, Lin Yang, and Tao Feng, “Formal Verification of 5G EAP-TLS Authentication Protocol”, Fourth International Conference on Data Science in Cyberspace, IEEE, 2019, pp 503-509.

[13] AN Braeken, Madhusanka Liyanage, Pardeep Kumar, and John Murphy, "Novel 5G Authentication Protocol to Improve the Resistance Against Active Attacks and Malicious Serving Networks", IEEE Access, Volume 7, 2019, pp 64040-64052.

[14] Ahmed A. Abd El-Latif, Bassem Abd-El-Atty, Salvador E. Venegas-Andraca, and Wojciech Mazurczyk, "Efficient quantum-based security protocols for information sharing and data protection in 5G networks", Future Generation Computer Systems 100 (2019), Elsevier, pp 893-906.

[15] Ijaz Ahmad, Tanesh Kumar, Madhusanka Liyanage, Jude Okwuibe, Mika Ylianttila, and Andrei Gurtov, "Overview of 5G Security Challenges and Solutions", IEEE Communications Standards Magazine, Volume 2, Issue 1, 2018, pp 36-43.

[16] Dongfeng Fang, Yi Qian, and Rose Qingyang Hu, 'Security for 5G Mobile Wireless Networks", IEEE Access, Volume 6, 2018, pp 4850-4874.

[17] Bin Han, Stan Wong, Christian Mannweiler, Mischa Dohler, Hans D. Schotten, "Security Trust Zone in 5G Networks", 2017, IEEE.

[18] Amandeep Singh and Dr. Charanjit Singh, "GbAS: Group-based Authentication System for Machine to Machine (M-M) Communication", International Journal of Advanced Science and Technology, ISSN: 2005-4238 IJAST, Vol. 29, No. 03, (2020), pp. 4671- 4680.

[19] Amandeep Singh and Dr. Charanjit Singh, "Bio-inspired Authentication of MTC in Long Term Evolution Networks", Turkish Journal of Computer and Mathematics Education (TURCOMAT), Vol.12 No. 7 (2021), pp. 1618-1630.

UBER and LYFT CAB FARE PREDICTION in BOSTON CITY USING REGRESSION TECHNIQUES

Avanthika Karthikeyan[1], Rhithika Sree K S[1], Deivarani S[2]
M.Sc. Artificial Intelligence and Machine Learning[1], Department of Data Science[2]
Coimbatore Institute of technology, India
karthikeyanavanthika@gmail.com
rhithupk542002@gmail.com
deivarani@cit.edu.in

ABSTRACT— Cabs play an important role in transportation. It acts as an alternative transportation vehicle in many cities. In developed countries, cabs tend to be used as a substitute for private vehicles by passengers. Passengers use the service for convenience reasons, or they use it if they do not own a car. The two major companies blooming in the Cab transportation sector are UBER and LYFT. Their cab prices are not fixed for the same source and destination, rather it keeps on changing dynamically based on various external factors. But the prices can be predicted using Machine Learning models with the available datasets. The Regression Models are best suited for such prediction. The chosen dataset contains the prices charged by UBER and LYFT for certain trips in the city of Boston. The aim of the paper is to predict the fare charged by UBER and LYFT in the city of Boston. The dataset that is used in this paper has been taken from Kaggle. The techniques implemented in this paper are Linear Regression and Random Forest Regressor. This paper also discusses the performance comparisons of each model and also shows which model predicts more accurately

KEYWORDS— linear regression, random forest, cab fare prediction, uber cab fare, regression models

INTRODUCTION

Cab services are an important aspect for facilitating tourism. One can easily make an online cab booking just by double clicking on the Internet. Uber and Lyft are the top two competitors in the United States of America where they provide many different types of cab services. This paper limits the prediction of Cab fare only to the city of Boston. Uber and Lyft determine the fees and terms on which drivers transport riders. Both companies use a dynamic pricing model. Customers are quoted with the fare in advance. Fare fluctuates depending on the local supply and demand at time of service. There are also other external factors that determine the ride price such as weather conditions, traffic and so on. Such factors also result in causing a surge in the price. Thus, it is difficult for one to label a fixed price for every trip given such extreme conditions. Machine Learning models are very useful in these types of situations. A Machine Learning model can predict these cab fare by assessing various conditions required to predict the fare. The main aim of this paper is to develop a Machine Learning model which does this prediction for the given dataset. Regression Algorithms are chosen, since they have proven to be more useful in predicting such values than the other algorithms. A Linear Regression and Random Forest Regressor algorithm is developed to predict the fare. The dataset includes the trip details of cab services in Boston over a span of 8 months. It also consists of key factors that decide the cab fare. This will help the customers to view their approximate fare that would be charged for the given source and destination in the city of Boston.

RELATED WORK

The fare charged by cabs can be calculated using mileage and the duration of the ride. This type of calculation requires real time data to estimate the fare. There are even ways to calculate the fare of a cab without using real time data, instead using data acquired from cab ride. Such a method was used by [1] in predicting the cab fare charged for each trip so that these can be very useful in the future prediction. The models used to build such a prediction system are linear regression and random forest. The same prediction was modelled by [2] using neural networks to make the prediction more accurate. Similarly [3] uses deep neural networks and stacking classifiers to predict the taxicab based on various dynamic conditions. A similar problem is being tried to solve here by taking the cab data into consideration and also predicting the cab fare based on weather data without involving real time data

METHODOLOGY

A. Dataset Description

The dataset was downloaded from Kaggle. The dataset that has been chosen consists of two csv files. The first csv files namely cabrides contain the details about the rides.
It contains attributes named

- Distance
  This column contains the distance travelled by the cab
- Cab_type
  This column consists of information whether the customer used uber or lyft cab services.
- Time_stamp
  This column consists of the time of the trip.
- Destination
  This column contains the drop place names.

- Source
  VideoThis column contains the pickup place names.
- Price
  This column contains the price charged for the corresponding trip.
- Surge_multiplier
  This column contains the surge multiplier for the corresponding trip.
- Id
  This column consists of the id for that corresponding trip.
- Name

  Uber and lyft provide various types of cab service. This column consists of information regarding what type of cab the customer has requested.

The second csv namely weather consists of information regarding the weather conditions for the trips. The csv consists of attributes namely temp, location, clouds, pressure, rain, time_stamp, humidity, wind.

B. Exploratory Data Analysis

1) Correlation plot for lyft:

Fig. 1 lyft correlation plot

From the Fig 1 it is known that surge multiplier and distance are weakly correlated. Also surge multiplier and price are more correlated than above pair

2) Correlation plot for uber:

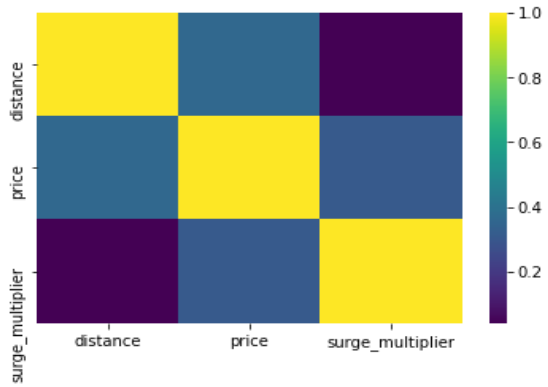


Fig. 2 Correlation plot for uber

From Fig 2 it is evident that price and distance are weakly correlated.

3) Bar plot to visualize the surge hike during the weekdays:

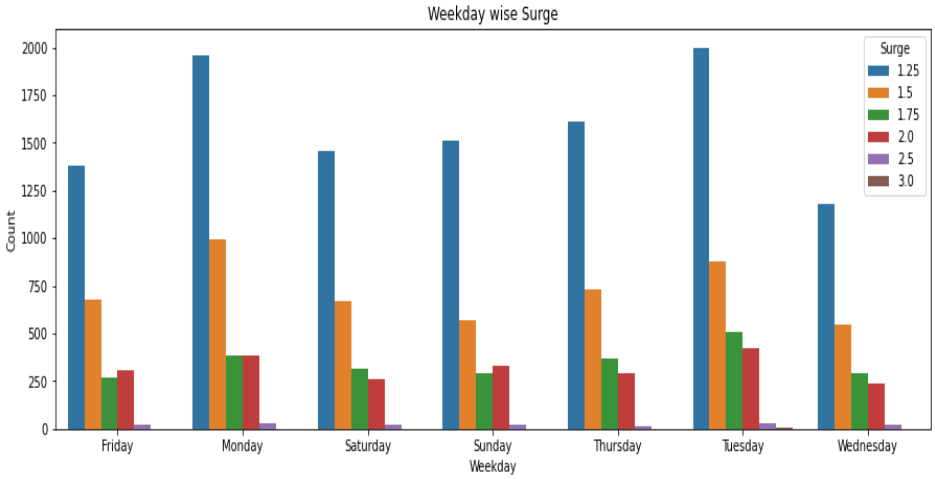


Fig. 3 Surge plot

From fig 3 it is evident that Tuesday has the highest surge hike.

4) Bar plot to find which company gets more rides when there is rain:

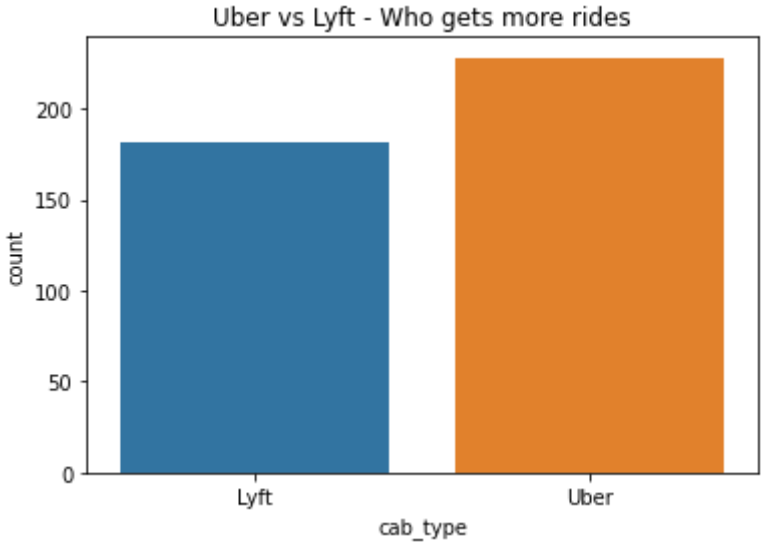


Fig. 4 Rain plot

It is evident from fig 2.4 that uber gets more rides compared to lyft when there is rain.

C. Data Pre-Processing

1) *Missing values:* Both the csv files contain some number of missing values. So, the missing value data are filled with 0 to make the prediction much easier.

2) *Dropping Unwanted Columns:* In cab_rides csv, the column product_id isn't required for price prediction, so it is dropped and in the weather csv file, the columns time_stamp and column_id isn't also required for the prediction, so those two columns are also dropped. Now the dataset doesn't contain any unwanted data.

3) *Shrinking and renaming the dataset:* The weather data set contains the weather conditions of each and every particular trip. The weather conditions rarely change when the trip is repeated. So, it is considered that the weather conditions remain the same during the prediction. Here, the mean value of every column is considered. Next step is to allocate weather data to both source and destination. So, the columns are renamed by source and destination weather.

4) *Merging the dataset*: The dataset has to be prepared for testing, training and splitting. Thus, the datasets are merged together, and the price column is dropped from the merged dataset since it is the target variable. The dataset is splitted into a x and y array where x contains all the other columns except for price and the array y contains the target variable, the price.

5) *Encoding:* It is known that the fit function used in prediction does not allow string variables to be fitted. Thus, the dataset must be encoded to convert all string variables into binary variables. The encoding technique used here is one hot encoding and after encoding all the columns, the string variable columns are dropped. Now the dataset does not contain any string variable.

The dataset is now ready for prediction.

D. Model Building Phase

1) *test_train_split:* In order to predict the fare and test it, the dataset must be split into training and testing subset. The training subset ratio is 75% which means the rest 25% is for testing the model.
2) *Algorithm Used - LINEAR REGRESSION:* Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).
3) *Algorithm Used - RANDOM FOREST REGRESSOR:* Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees.

The models are built and are evaluated to know which model out of these two has predicted accurately.

RESULTS AND DISCUSSION

Based on results from the evaluation metrics, the linear regression model has predicted much more accurately than the random forest regressor model.

Root mean square value of Linear regression model: 9.2383711
Root mean square value of Random Forest regressor model: 9.6643627
These values show that Linear regression model predicted accurately due to the low root mean squared value compared to the same value of random forest regressor model.

Actual vs Predicted Values

1) *Linear Regression:*

	actual	predicted
19554	13.0	14.158041
83511	7.5	12.539931
34788	16.0	10.365595
6148	7.0	10.770123
41776	0.0	15.852000
...	...	...
100530	0.0	14.790116
101886	32.5	20.984444
90060	20.5	15.599171
51193	16.0	13.677665
66658	19.5	13.778797

Fig 5. Value Comparison

In the fig 5 the actual and predicted values by linear regression are shown.

2) Random Forest Regressor:

	actual	predicted
729	16.5	18.430
34909	35.0	26.700
17089	20.5	15.465
10180	5.0	17.780
31659	8.5	17.930
...	...	...
13504	13.5	20.475
24907	3.5	14.900
16172	19.5	17.490
4325	11.5	28.070
17552	32.5	19.295

Fig 6. Value Comparsion

In the fig 6 the actual and predicted values by random forest regressor are shown.

CONCLUSION

As said in the goal, two models were built to predict the cab fare charged by uber and lyft in the city of Boston given the source and destination and the weather conditions. It was also analysed how the surge multiplier affects the price. The built models were also evaluated using evaluation metrics and based on the numeric values it was found out that the 'Linear Regression' model's prediction is 'more accurate' than the 'Random Forest Regressor' prediction. This conclusion was arrived at by the fact that the linear regression model has less root mean squared error than the random forest regressor model.

REFERENCES

[1] http://cs229.stanford.edu/proj2016/report/AntoniadesFadaviFobaAmonJuniorNewY orkCityCabPricing-report.pdf
[2] Van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. "Accurate freeway travel time prediction with state-space neural networks under missing data." Transportation Research Part C: Emerging Technologies 13.5 (2005): 347- 369.
[3] https://www.researchgate.net/publication/324706525_Taxi_Fare_Rate_Classificatio n_Using_Deep_Networks
[4] https://machinelearningmastery.com/regression-metrics-for-machine-learning/
[5] https://towardsdatascience.com/understanding-random-forest-58381e0602d2
[6] https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/code
https://www.taxifarefinder.com/main.php?city=ny &lang=en

HTCN-A3D: A DEEP LEARNING ENSEMBLE FOR UNSUPERVISED ANOMALY DETECTION OF HIGH DIMENSIONAL TIME SERIES DATA

Pritika Mehra[#1], Mini Singh Ahuja[*2]
[#] *Department of Computer Science and Engineering, Guru Nanak Dev University Amritsar India*
[*] *Department of Computer Science and Engineering, Guru Nanak Dev University Regional Campus Gurdaspur India*
[1]pritikacsc.rsh@gndu.ac.in
[2]miniahuja06@gmail.com

ABSTRACT: In the current situation of COVID 19, online business and online education systems are taking the place of offline work; the numbers of users using the internet are enormously increasing with increased online activities. The world is facing unfamiliar scenarios everywhere which is giving rise to anomalous events. These anomalous events can impact organizations to handle their work in a new normal way. The data is becoming larger and larger in size and dimension which poses a challenge in identifying anomalous events as deceitful activities are also increasing. Therefore Anomaly Detection is a need of an hour in every field. Anomaly detection is required to avoid the breakdown of organizations through frauds and malicious activities. This paper deals with the problem of anomaly detection for high dimensional time series data. Although there has been an extensive work on anomaly detection, most of the techniques are based on traditional supervised or unsupervised machine learning. This paper attempts to use an unsupervised deep learning ensemble in which two or more deep learning techniques can be combined to achieve better performance of anomaly detection of high dimensional time series data. In this paper, we have proposed an HTCN-A3D: A deep learning ensemble for unsupervised anomaly detection of high dimensional time series data using Temporal Convolutional Network augmented with attention mechanism.

KEYWORDS: *High Dimensional Data, Time Series Data, Ensemble, Deep Learning, Temporal Convolutional Network, Attention*

I. INTRODUCTION

Anomaly Detection is a process of identifying unexpected items or events in a dataset which differ from normal. These unexpected items are called anomalies. For detecting anomalies in a dataset, anomaly detection approaches are required. Anomaly detection approaches can be based on supervised, semi supervised or unsupervised machine learning methods. Supervised Anomaly detection is not suitable as it can detect only known anomalies. Thus Unsupervised Anomaly Detection will be preferred over supervised learning methods in this research work. Variety of Unsupervised Learning algorithms are used in previous research works. This research work will focus on deep learning Ensemble based Unsupervised Anomaly detection to improve the accuracy of detecting anomalies.

A. Anomaly Detection

There are several surveys of Anomaly Detection in the literature; however they concern different aspects of Anomaly Detection. E.g. Chandola et al. [1] and Agarwal & Agarwal [2] only review traditional algorithms for various applications which are not suitable for high dimensional time series data. Goldskein et al. [4] evaluates different unsupervised Anomaly detection algorithms on different datasets. Ahmed et.al [5] provides a survey of state-of-the-art real time big data processing Anomaly Detection.

B. High Dimensional Time Series

Nowadays large numbers of observations are being used which may have dimensions in the thousands or millions while only tens or hundreds of observations are accessible for study. It is tremendously complicated and demanding to handle High-dimensionality along with large datasets. Moreover, data can have distinctive features and high-dimensional data structures, which means that conventional analysis techniques do not work well. To analyze extra useful information from high-dimensional data, novel approaches are required.

Xu et al.[15,7] provides an overview of high dimensional data including neighbour ranking based method, subspace based method and ensemble learning based method which cannot be efficiently applied on time series data. Thudumu et al. [8] documents the state-of-art anomaly detection techniques for high dimensional big data but does not consider time series data.

A time series is a sequence of information that attaches a time period to each value. The value can be any measurable quantity that depends upon time in some way, like prices, humidity, or a number of people. [20]. Talagala et.al.[19] proposes a STRAY algorithm based on extreme value theory and feature engineering for detecting anomalies of both high dimensional and time series data but is evaluated only on pedestrian counting system dataset. The algorithm has not been evaluated for different datasets with diverse properties.

The existing algorithms and techniques in unsupervised anomaly detection do not work properly on High dimensional data especially when applied on time series. This is why there is a need for an ensemble approach that combines algorithms to provide an efficient solution of anomaly detection of time series data.

C. Deep learning Ensemble

In recent years, deep learning based Anomaly detection algorithms are becoming popular and completely surpasses Traditional methods. Goldthorpe et.al [24] denoising autoencoders are used to detect anomalies in correlated data. The model used is not efficient in removing errors in highly correlated data and may also not be effective for high dimensional time series data. Raghavendra et. al.[6] aims to provide a broad outline of state-of-art in Deep Anomaly Detection techniques.CNN, LSTM and RNN deep neural Network models are stated. Shoemaker et al.[17] compared supervised and semi-supervised ensemble methods with non ensemble traditional anomaly detection methods without any discussion of unsupervised ensemble methods. Since supervised ensemble training method requires labelled data for training and can only identify known abnormal types so its application scope is limited. Much attention is required towards unsupervised ensemble methods. Much efficient and optimized ensemble technique is required to detect anomalies for time series data. [6, 26] discuss DAD techniques such as CNN, GAN,DNN, DAE, VAE, SDAE, LSTM, RNN and GRU for time series data.

Gugulothu et. al. [22] proposes Sparse Recurrent Neural Network based Anomaly Detection (SPREAD) approach based on point wise non temporal dimensionality reduction using feedforward layer and recurrent layers for time series compression using recurrent autoencoder. Yaldiz et. al. [25] proposes a hybrid deep learning framework to solve unsupervised Anomaly Detection in spatio temporal data. This framework may not produce optimum performance for various real world applications as it has been tested for only spatio-temporal datasets. Zheng Gao et.al. [27] Propose a model combining the Adversarial Autoencoder and Recurrent NN but it may not work well with real world streaming data as it is tested on data of online recommendation and advertising only. Chao Meng e. al. [28] Propose a framework called Multidimensional Time Series Outlier Detection based on a Time Convolutional Network Autoencoder (MOTCN-AE). Execution time for MOTCN-AE is more for small dataset.

Based on the extensive and unifying review of previous research work, it is found out that deep learning ensemble techniques have gained a lot of attention recently as compared to traditional methods for detecting anomalies.

II. HTCN-A3D Framework

Here, we propose a High Dimensional Temporal Convolutional Network Attention based Autoencoder Anomaly Detector (HTCN-A3D). Our framework is inspired by MOTCN-AE [28] in which optimal weights of autoencoder are not taken. Therefore we incorporate attention mechanism [29] in TCN (Temporal Convolutional Network) which will help to find optimal weights of encoder output. Attention mechanism adds weighted features of neighbouring nodes. Also algorithm can be optimized by time series feature enrichment as in MOTCN-AE [28] and Time Series Approximation.

The HTCN-A3D framework proposed in this paper is shown in figure 1. The proposed method consists of following steps:

- Original time series is converted into enriched time series using statistical feature extraction.
- Enriched time series is approximated using time series approximation to generate enriched approximated time series.
- Enriched approximated Time Series is coded by TCN-AE.
- Then the attention mechanism produces optimal weights of encoder output by adding weighted features of neighbouring nodes.
- TCN- AE decodes the output to generate reconstructed time series.
- The enriched approximated time series data and the reconstructed TS are compared, and the differences are taken as anomalies.

Hypothesis: Adding Attention mechanism and time series approximation to MOTCN-AE [28] will produce optimal and efficient results in the detection of anomalies.

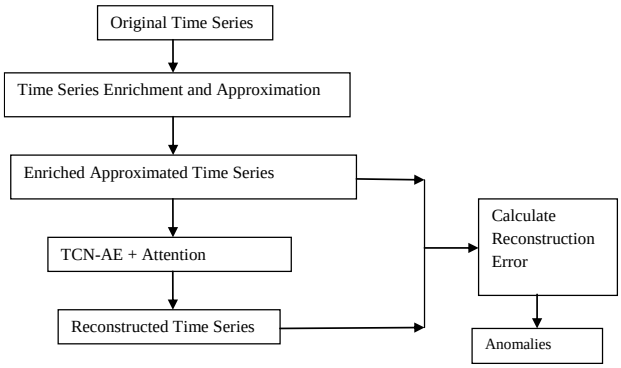


Figure 1: The framework of HTCN-A3D

To prove the hypothesis, the proposed framework will be implemented using different datasets and the results of the proposed framework will be compared with other traditional or deep learning ensemble methods based on the evaluation criteria.

III. EXPERIMENTAL SETTINGS

1) Datasets

We will validate our proposed method on five real datasets and a synthetic dataset. The datasets in consideration are: NAB, SKAB, KPI, Nifty Stock BitcoinHeistRansomwareAddress dataset and AI4I 2020 Predictive Maintenance Dataset.

NAB: NAB is a novel benchmark for evaluating algorithms for anomaly detection in streaming, real-time applications. It is composed of over 50 labelled real-world and artificial time series data files plus a novel scoring mechanism designed for real-time applications. The NAB dataset contains 58 univariate time series, and each time series contains 1000 to 22,000 vectors. The time series dataset is collected from various applications, such as network traffic monitoring, cloud server CPU usage, industrial device operation monitoring, and social media monitoring. Each time series contains a Boolean outlier tag that is used to help us identify outliers.

SKAB: Skoltech Anomaly Benchmark (SKAB) is designed for evaluating the anomaly detection algorithms. The dataset represents a multivariate time series collected from the sensors installed on the testbed.

KPI: The dataset consists of KPIs (key performace index) time series data from many real scenarios of Internet companies with ground truth label.

Nifty Stock Dataset: It comprises stock price data and can be available from Kaggle.

BitcoinHeistRansomwareAddress dataset: BitcoinHeist datasets contains address features on the heterogeneous Bitcoin network to identify ransomware payments. This dataset is available from UCI machine learning repository. It contains multivariate time series with 10 attributes.

AI4I 2020 Predictive Maintenance Dataset: The AI4I 2020 Predictive Maintenance Dataset is a synthetic dataset that reflects real predictive maintenance data encountered in industry. It is a multivariate time series with 14 attributes and 10,000 instances.

2) Approaches considered for comparison

We will compare HTCN-A3D with standard Encoder Decoder AD, Recurrent Autoencoder, Conventional CNN, TCN-AE and MOTCN-AE.

3) Performance Evaluation/ Metrics

The five methods will be evaluated using following metrics:

1) *Precision*: Precision is the number of correct predictions over the output size but it only measures the rate of false positives.

$$Precision = \frac{true\ Posiive}{true\ Positive + False\ Positive}$$

2) *Recall*: Recall is the opposite of precision; it measures false negatives against true positives.

$$Recall = \frac{True\ Positive}{true\ Positive + False\ Negative}$$

3) *Area under the curve (AUC):* AUC is the area under the receiver operating characteristic (ROC) curve, indicating the overall accuracy of a classification method. ROC plot is used to visualise the performance of a binary classifier. It gives us the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different classification thresholds.
4) *True Positive Rate*: True Positive Rate is the proportion of observations that are correctly predicted to be positive.

$$TPR = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5) *False Positive Rate*: False Positive Rate is the proportion of observations that are incorrectly predicted to be positive.

$$FPR = \frac{False\ Positive}{True\ Negative + False\ Positive}$$

For different threshold values we will get different TPR and FPR. So, in order to visualise which threshold is best suited for the classifier we plot the ROC curve.

6) F1-Score: the F1-score is a balanced metric that appropriately quantifies the correctness of models across many domains. It is a measure of a model's accuracy on a dataset.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

IV. CONCLUSION AND FUTURE WORK

In this paper, a High dimensional time series anomaly detection framework based on a TCN with TS approximation and Attention mechanism is proposed (HTCN-A3D). In particular, HTCN-A3D is an ensemble that combines a time convolution network, along with Time series approximation and attention mechanism. Our proposed method can be trained effectively on high-dimensional time series data and is an attempt to achieve accurate results on several benchmark datasets. We will compare the proposed method with the state-of-the art techniques using extensive experiments. The

proposed method aims to improve the detection accuracy as compared to the existing methods. In future, focus will be to implement the proposed method and show that it outperforms other benchmarks on above given datasets.

REFERENCES

[1] Chandola V., Banerjee A., Kumar V., Anomaly detection:A survey, ACM Computing Surveys (CSUR); 41(3);2009; p.15

[2] Agrawal S.,Agrawal J., Survey on Anomaly detection using data mining techniques, Procedia Computer Science 60; 2015; p.708-713

[3] Blomquist H., Miller J., Anomaly detection with machine learning; Examinator; 2015; 1650-8319;UPTEC STS15 014

[4] Goldstein M., Uchida S., A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, PLOS ONE 11(4);2016; doi: 10.1371/ journal.pone.0152173

[5] Habeeb R., Gani A., Nasaruddin F., Hasheem I., Real time big data processing for anomaly detection: A survey; doi:10.1016/j.ijinfomgt; 2018.08.006

[6] Raghavendra Chalapathy, Sanjay Chawla, Deep Learning for Anomaly Detection-A survry; arXiv:1901.03407v2, 2019

[7] Xiodan Xu, Huawen Liu, Minghai Yao, Recent progress of anomaly detection Hindawi; Complexity; 2019;2686378

[8] Thudumu, S., Branch, P., Jin, J. *et al.* A comprehensive survey of anomaly detection techniques for high dimensional big data. *J Big Data* **7,** 42 (2020). https://doi.org/10.1186/s40537-020-00320-x

[9] Karadayı Y, Aydin MN, Öğrenci AS. A Hybrid Deep Learning Framework for Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Data. *Applied Sciences*. 2020; 10(15):5191. https://doi.org/10.3390/app10155191

[10] Amarbayasgalan T, Pham VH, Theera-Umpon N, Ryu KH. Unsupervised Anomaly Detection Approach for Time-Series in Multi-Domains Using Deep Reconstruction Error. *Symmetry*. 2020; 12(8):1251. https://doi.org/10.3390/sym12081251

[11] https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

[12] https://towardsdatascience.com/effective-approaches-for-time-series-anomaly-detection -9485b40077f1

[13] https://towardsdatascience.com/unsupervised-learning-for-anomaly-detection-44c55a96b8c1

[14] https://towardsdatascience.com/unsupervised-anomaly-detection-on-time-series-9bcee10ab473

[15] Xiaodan Xu , Huawen Liu , Li Li , Minghai Yao, A Comparison of Outlier Detection Techniques for High-Dimensional Data, International Journal of Computational Intelligence Systems, Vol. 11 (2018) 652–662

[16] Ane Blazquez-Garcia and Angel Conde, et. al.,A review on outlier/anomaly detection in time series data, arXiv:2002.04236v1, ACM ,2020

[17] Lary Shoemaker, Lawrence O Hall, MCS'11: Proceedings of the 10th international conference on Multiple classifier systemsJune 2011 Pages 6–15

[18] Shweta B. Meshram, Sharmila M. Shinde, A Survey on Ensemble Methods for High Dimensional Data Classification in Biomedicine Field, International Journal of Computer Applications (0975 – 8887) Volume 111 – No 11, February 2015

[19] Priyanga Dilini Talagala, Rob J. Hyndman, Kate Smith-Miles, 2019, arXiv:1908.04000

[20] https://365datascience.com/tutorials/time-series-analysis-tutorials/time-series-data/

[21] Tung Kieu , Bin Yang, Chenjuan Guo and Christian S. Jensen, Outlier Detection for Time Series with Recurrent Autoencoder Ensembles,Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19).

[22] Narendhar Gugulothu, Pankaj Malhotra, Lovekesh Vig, Gautam Shroff, Sparse Neural Networks for Anomaly Detection in High-Dimensional Time Series, International Workshop on AI for Internet of Things, IJCAI 2018, Stockholm, Sweden.

[23] Nadeem Iftikhara, Thorkil Baattrup-Andersenb, Finn Ebertsen Nordbjerga, Karsten Jeppesena, Outlier Detection in Sensor Data using Ensemble Learning, Procedia Computer Science 176 (2020) 1160–1169

[24] Peter Goldthorpe, Antoine Desmet, Denoising autoencoder anomaly detection for correlated data, EUROPEAN CONFERENCE OF THE PROGNOSTICS AND HEALTH MANAGEMENT SOCIETY 2018

[25] Yeldız Karadayı , Mehmet N. Aydin and A. Selçuk Ö ğrenci, A Hybrid Deep Learning Framework for Unsupervised Anomaly Detection in Multivariate Spatio-Temporal Data, Appl. Sci. 2020, 10, 5191; doi:10.3390/app10155191, 2020

[26] Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Gregoire Montavon, Wojciech Samek, ´ Marius Kloft, Thomas G. Dietterich, Klaus-Robert Muller, A Unifying Review of Deep and Shallow Anomaly Detection , arXiv:2009.11732v2, 2020

[27] Zheng Gao, Lin Guo, Chi Ma, Xiao Ma, Kai Sun, Hang Xiang, Xiaoqiang Zhu, Hongsong Li, Xiaozhong Liu, AMAD: Adversarial Multiscale Anomaly Detection on High-Dimensional and Time-Evolving Categorical Data, arXiv:1907.06582v1, 2019

[28] Chao Meng , Xue Song Jiang , Xiu Mei Wei , And Tao Wei, A Time Convolutional Network Based Outlier Detection for Multidimensional Time Series in Cyber-Physical-Social Systems, 10.1109/ACCESS.2020.2988797

[29] Arnav Kundu, Abhijeet Sahu, Erchin Serpedin and Katherine Davis, A3D: Attention-based Auto-encoder Anomaly Detector for False Data Injection Attacks, Electric Power Systems Research. 189. 106795. 10.1016/j.epsr.2020.106795

[30] Zhao, Zhiruo, "Ensemble Methods for Anomaly Detection" (2017). Dissertations - ALL. 817. https://surface.syr.edu/etd/817

[31] Maya, S., Ueno, K. & Nishikawa, T. dLSTM: a new approach for anomaly detection using deep learning with delayed prediction. *International Journal of Data Science Anal*ysis **8,** 137–164 (2019). https://doi.org/10.1007/s41060-019-00186-0

[32] Sen, Rajat & Yu, Hsiang-Fu & Dhillon, Inderjit, Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. pages 4838-4847, 2019.

[33] Pereira, João. (2018). Unsupervised Anomaly Detection in Time Series Data using Deep Learning. 10.13140/RG.2.2.15967.07849.

A REVIEW OF DEEP LEARNING TECHNIQUES FOR SEGMENTATION OF MULTIPLE ORGANS

Harinder Kaur[#1], Navjot Kaur[*2], Nirvair Neeru[#3]
[#1,2,3] *Department of Computer Engineering, Punjabi University Patiala*
[1]narain1293@gmail.com
[2]navjot_anttal@yahoo.co.in
[3]nirvair_neeru@yahoo.com

ABSTRACT— Segmentation is the fundamental and crucial task of computer vision field. Deep learning (DL) algorithms, especially convolutional networks, have achieved enormous success in almost every domain, making it the technique of choice for medical image analysis. This paper presents a study of current deep learning methods for medical image segmentation, with a focus on multiple organs in different anatomies. To provide the basic insights of deep learning architectures, the fundamental networks like CNN, FCN and U-Net are detailed. The comprehensive survey of DL based multi organ approaches along with their issues and contributions is summarized. We conclude the article with open challenges and future directions to conduct the studies.

KEYWORDS—Segmentation, Deep learning, Convolutional Neural Network and Fully Convolutional Neural Network.

INTRODUCTION

Deep learning is the powerful concept, and it has recently achieved great success in computer vision applications named object detection [1], classification, segmentation etc. It can be generalized as an artificial neural network (ANN) with several layers including convolutional, de-convolutional, upsampling, downsampling, and so on. Since there is no fixed count of layers that is called deep, thus there is no explicit distinction between DL and ANN.

Multi organ segmentation is the process of classifying each and every voxel into some predefined organ category based on their homogeneous properties or some other criterion. It is the fundamental step in medical image applications like surgery planning, automated CAD systems, delineation of OAR is helpful in treatment planning, radiation oncology, etc. Thus it is required to effectively detect the organ structure. Multi organ segmentation based on DL is now the usual choice due to their gained success in every domain.

Radiotherapy is the key treatment to kill the head and neck (HaN) cancerous cells, but proper planning is essential to ensure that only the required amount of radiation is given to the sensitive or OAR like nerves and eyes. The experts of this field need five to more hours to do this manually [2]. Also, the radiologist takes long time to manually segment the abdominal organs which is helpful in tumour detection of various organs such as pancreas, kidney etc. Thus, manual segmentation of structures is time consuming task [3]. Hence, it is the necessary pre-processing condition to automatically segment organs at risk (OAR) for treatment delivery and planning [4]. Even the shape of the organ gives the idea of size of tumour which can elaborate how critical the patient is.

The traditional approaches for this organ segmentation are multi atlas, shape models and recently developed deep learning models. The major challenge that arises in MOS is the huge shape variability for e.g. parotid gland is challenging organ to segment due to shape variability and low soft tissue contrast. Although shape models helps to tackle this problem but large number training shapes are required for these shape models [4], [5] and landmark information is needed [6]. But deep learning models can address these issues effectively, deep learning models has capability to extract deep features without manual initialization. This paper aims to review deep learning approaches for multi organ segmentation in various anatomies. The remainder of the paper is organized as follows:

Section II provides the basic insights of deep learning architectures such as CNN, FCN etc., section III reviews the various deep learning approaches for multi organ segmentation, section IV concludes the review with future directions.

BACKGROUND

One of the most widely researched DL-based applications in the medical field is segmentation. As a result, there are several methodologies and network architectures to choose from. This section of the paper aims to give the basic understanding of the DL based architectures used for segmentation of organs. These DL based architectures are different from each other in various aspects including kind of training whether it is supervised, semi-supervised or transfer learning from pre-trained network, batch size, number and kind of hidden layers, number of iterations and so on. The basic architectures for deep learning are given below:

E. Convolutional Neural Network (CNN)

Convolutional neural network consists of many layers where the output of one layer acts as an input for the next layer. The convolutional layers extract simple features such as edges, corners and simple shapes. The ReLU layer, sigmoid, tangential added the non-linearities into the network which can extract the complex features. In order to avoid the overfitting in the architecture, regularization techniques can be used such as dropout. These extracted features will be generalized at final layer and segmentation task is performed. The pooling layer may help to reduce the number of network parameters. The major benefit of CNN is that spatial information is taken into account by it [7].

F. Fully Convolutional Networks (FCN)

FCN is the major structure to perform semantic segmentation. It comprises convolutional layers, fully connected layers, pooling layers, up sampling and down sampling. The up sampling layer helps to interpret the local features and downsampling is utilized to predict the context of the input image. The fine-grained spatial information may be lost during deep downsampling, thus skip connections helps to recover the spatial information. The results from deep layer and shallow layer can be fused together in order to enhance the performance [8].

G. U-Net

U-Net architecture consists of two paths: a contracting path and expansive path. The contracting path is the standard convolutional neural network with 3x3 convolutions followed by non-linear ReLU unit and pooling operation with downsampling. The feature channels are doubled at each downsampling layer. The expansive path comprises upsampling with 2x2 convolutions and here halve the feature channels followed by ReLU unit. The standard U-net architecture consists of 23 convolutional layers. Due to u-shaped paths, it is named as U-net [9].

METHODS

Recently, there is huge advancement in deep learning methods and improvement in segmentation accuracy over aforementioned methods such as multi atlas, shape models etc. Also, a comparison study performed in [10] suggests that CNN has strong recognition power over traditional approaches, hence a obvious choice. There is wide variety of DL variants and extensions available in literature for multi-organ segmentation which is reviewed in this section.

The deep learning algorithm (CNN) is successfully applied for HaN CT images to segment OAR [7]. During Convolutional Neural Network (CNN) training, positive intensity patches are extracted around the region of interest (ROI) i.e. OAR and negative intensity patches around the non-ROI. These patches are passed through the various layers of CNN which are capable to extract the local features edges, corners and boundaries. This trained network will segment the ROI and the results are further smoothed with Markov Random Field (MRF). Although significant results are obtained for mandible but it can be further improved for nerves and chiasm. This study demonstrates that CNN has potential to effectively delineate the OAR. The CNN can be further optimized around the boundaries of organs such as chiasm by adding prior information like MRI information [7]. A study is performed in [6] which incorporates prior information or shape constraint with CNN. In this approach, manual segmentation performed by experienced radiation oncologist is used as prior information which will help the CNN to find the target [6]. A method based on manifold learning and random patch forests is used to extract the features from patches and perform segmentation. But this could be improved by adding spatial locality and multiscale patches [4].

Due to diverse shape, low contrast and volume of small tissues such as chiasm, duodenum [24]., optical nerve, aorta etc. are difficult to segment. To address this challenge, interleaved 3D-CNN is proposed. The input image is divided into patches and CNN is applied to the extracted patch. Further, the neighbouring tissues are highly related to each other in anatomical terms. Thus, tentative results of one tissue can contribute to the segmentation result of neighbouring tissues. A complex network of interleaved and cascaded CNNs is constructed to join the results of all the tissues. It has been observed in this study that the relative positions of chiasm and optical nerve are inconsistent which suggests that consistent positions are not obtained during training due to small dataset and hence larger datasets are required to derive such consistent positions [11], [12].

3D-FCN is applied on thorax and abdominal anatomy for multi organ segmentation. Although promising results are obtained but they have tested with very small dataset and testing is limited to large volume organs only [13].

The integration of the traditional techniques with deep learning models can improve the segmentation accuracy. A technique is presented which combines the probabilistic atlas with deep CNN. The generated probabilistic atlas is utilized to locate the region of interest and this information is passed to CNN to segment the target in MRI images [14]. This model can be further fine-tuned to improve the results. The integrating practice is also opted in [15] to combine deformable image registration and CNN. This combination acts as a tool to obtain longitudinal data of the previously segmented images of the same patient. A technique is proposed to extend the standard adversarial networks with convolutional neural network. This integration helps to improve the generalization of deep learning model and to automatically locate the organs [16]. FCN-DecNet is applied in [17] with multiple convolutional, deconvolutional and fusion layers. The obtained rough results are refined with probabilistic atlases.

In order to address the low contrast problem and inter-patient variation, integration model is devised which combines shape representation model and fully convolutional network [18]. Furthermore, aforementioned patch based methods are ineffective due to redundancy in patches and learning is limited to local features only. The hybrid model devised in [18] avoids making patches and hence removes redundancy. Shape representation model is used to learn the shape of HaN. FCN training is constrained with this pre-trained shape representation model. This combined model is utilized to segment the multiple organs. Although this method is robust, but it must be tested on large dataset [18]. The different views of planes such as coronal, axial and sagittal views has been taken into account in [19]. Furthermore, multi-planar fusion is applied to generate more effective labels and perform better segmentation for abdominal anatomy.

3D U-Net with 8 levels is proposed in [20] to segment OAR and binary cross entropy is used for regularization. The major benefit of this technique is that a large dataset with 663 images is utilized [20]. The standard U-Net approach is applied first in [21] to assess the effect of dilated convolutional layers in U-Net. This will help to understand the global context as well as to extract the local features of the organs. A 3D U-Net based, two step segmentation techniques is proposed [22]. Firstly, image is divided into sub-tasks to locate the boundaries using bounding box and for each sub-task, dedicated 3D U-Net is utilized to perform segmentation [22]. Similar kind of concept is used on another study but this technique has trained two hierarchal neural networks for multi-organ segmentation. First neural network is trained to locate the organs of interest and second trained model is used to perform final segmentation [23]. The outline of DL approaches is presented in Table IV.

TABLE V
Overview of Deep Learning Based Approaches for Multi-Organ Segmentation

Study	Architecture	Organs	Anatomy	Data	Modality
2015 [4]	Manifold learning and Random forest patches	Salivary gland, Parotid glands	HaN	17 images	CT
2017 [7]	CNN	Spinal cord, Mandible, Parotid glands, Larynx, Pharynx, Eye globe, Optic nerve and Chiasm	HaN	50 images	CT
2017 [17]	Full Convolutional-Deconvolutional Network	Liver, kidney and spleen	Abdominal	70 slices from 12 patients	CT
2018 [11]	Interleaved CNN	Chiasm and optical nerve	HaN	48 images	CT
2018 [13]	3D FCN	Lungs, liver, spleen and kidneys	Thorax and abdominal	26 images	Dual -CT
2018 [14]	Probabilistic atlas and deep convolutional neural network	Trachea, spinal cord, parotis, SCM, carotis, jugularis	-	15 images	MRI
2018 [18]	Shape representation model and FCN	Brainstem, chaism, mandible, optical nerves, parotid and submandibular glands	HaN	Public domain database for computational anatomy (PDDCA)	CT
2018 [20]	3D UNET	Brain, brainstem, mandible, parotid, model, oncologist	HaN	663 images	CT
2018 [21]	Standard U-Net and Dilated U-Net	Esophagus, lungs and spinal cord	Chest	29 images	CT
2019 [23]	Hierarchal Neural Network approach	BrainStem, Chiasm Mandible, OpticNerve, Parotid	HaN	40 images	CT
2019 [19]	Deep Multiplanar co-training	Spleen, Kidney, Gall Bladder, Liver, Stomach Aorta IVC Veins Pancreas	Abdominal	326 images	CT
2020 [16]	Cascaded convolutional neural and adversial deep network	Liver, spleen, kidney	Abdominal	CHAOS	CT and MRI

CONCLUSIONS

The main deep learning methods for automated segmentation of multiple organs are reviewed in this article. The fundamental anatomies such as head and neck, abdominal and thorax are considered as benchmark to evaluate DL based methods. It can be inferred that low contrast and volume tissues such as patroid glands, chiasm, aorta, optical nerves, oesophagus, and so on pose significant challenges in multiorgan segmentation. Despite the fact that experiments have been undertaken to resolve these problems, but still there is a room for further improvement. The prevalence of DL-based hybrid methods means that future experiments can show studies using hybrid methods.

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," vol. 8828, no. c, pp. 1–14, 2016, doi: 10.1109/TPAMI.2016.2577031.

[2] C. Chu et al., "Applying machine learning to automated segmentation of head and neck tumour volumes and organs at risk on radiotherapy planning CT and MRI scans [version 1; referees: 1 approved with reservations]," F1000Research, vol. 5, no. 0, pp. 1–8, 2016, doi: 10.12688/F1000RESEARCH.9525.1.

[3] J. Y. Lim and M. Leech, "Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck," Acta Oncol. (Madr)., vol. 55, no. 7, pp. 799–806, 2016, doi: 10.3109/0284186X.2016.1173723.

[4] K. Fritscher, S. Magna, and S. Magna, "Machine-learning based image segmentation using Manifold Learning and Random Patch Forests," Imaging Comput. Assist. Radiat. Ther. Work. MICCAI 2015, no. October 2015, pp. 1–8, 2015.

[5] R. Gauriau, R. Cuingnet, D. Lesage, and I. Bloch, "Multi-organ localization with cascaded global-to-local regression and shape prior," Med. Image Anal., vol. 23, no. 1, pp. 70–83, 2015, doi: 10.1016/j.media.2015.04.007.

[6] J. Léger, E. Brion, U. Javaid, J. Lee, C. De Vleeschouwer, and B. Macq, "Contour Propagation in CT Scans with Convolutional Neural Networks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 11182 LNCS, no. Midl 2018, pp. 380–391, 2018, doi: 10.1007/978-3-030-01449-0_32.

[7] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks:," Med. Phys., vol. 44, no. 2, pp. 547–557, 2017, doi: 10.1002/mp.12045.

[8] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 640–651, 2017, doi: 10.1109/TPAMI.2016.2572683.

[9] W. Weng and X. Zhu, "INet: Convolutional Networks for Biomedical Image Segmentation," IEEE Access, vol. 9, pp. 16591–16603, 2021, doi: 10.1109/ACCESS.2021.3053408.

[10] G. Palareti et al., "Comparison between different D-Dimer cutoff values to assess the individual risk of recurrent venous thromboembolism: Analysis of results obtained in the DULCIS study," Int. J. Lab. Hematol., vol. 38, no. 1, pp. 42–49, 2016, doi: 10.1111/ijlh.12426.

[11] M. Barat et al., "Mass-forming lesions of the duodenum: A pictorial review," Diagn. Interv. Imaging, vol. 98, no. 10, pp. 663–675, 2017, doi: 10.1016/j.diii.2017.01.004.

[12] X. Ren et al., "Interleaved 3D-CNNs for joint segmentation of small-volume structures in head and neck CT images," Medical Physics., vol. 45, no. 5, pp. 2063–2075, 2018, doi: 10.1002/mp.12837.

[13] J. Torrents-Barrena et al., "Segmentation and classification in MRI and US fetal imaging: Recent trends and future prospects," Medical Image Analysis., vol. 51, pp. 61–88, 2019, doi: 10.1016/j.media.2018.10.003.

[14] S. Chen, S. Dorn, and A. Maier, "Automatic multi-organ segmentation in dual energy CT using 3D fully convolutional network," Midl, no. Midl, pp. 1–9, 2018.

[15] L. G. Varga and V. Szpisjak, "Automatic background-foreground segmentation of organs on MRI images of the head-neck area," IEEE 30th Jubil. Neumann Colloquium, NC 2017, vol. 2018-Janua, pp. 71–76, 2018, doi: 10.1109/NC.2017.8263253.

[16] L. Vandewinckele, D. Robben, W. Crijns, and F. Maes, Segmentation of head and neck organs-at-risk in longitudinal ct scans combining deformable registrations and convolutional neural networks, vol. 11045 LNCS. Springer International Publishing, 2018.

[17] P. H. Conze et al., "Abdominal multi-organ segmentation with cascaded convolutional and adversarial deep networks," arXiv, pp. 1–10, 2020.

[18] Y. Yang, H. Jiang, and Q. Sun, "A Multiorgan Segmentation Model for CT Volumes via Full Convolution-Deconvolution Network," vol. 2017, 2017.

[19] N. Tong, S. Gou, S. Yang, D. Ruan, and K. Sheng, "Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks," Med. Phys., vol. 45, no. 10, pp. 4558–4567, 2018, doi: 10.1002/mp.13147.

[20] Y. Zhou et al., "Semi-Supervised 3D Abdominal Multi-Organ Segmentation via Deep," 2019 IEEE Winter Conf. Appl. Comput. Vis., pp. 121–140, 2019, doi: 10.1109/WACV.2019.00020.

[21] S. Nikolov et al., "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," arXiv, pp. 1–31, 2018.

[22] U. Javaid, D. Dasnoy, and J. A. Lee, Multi-organ Segmentation of Chest CT Images in Radiation Oncology: Comparison of Standard and Dilated UNet, vol. 11182 LNCS. Springer International Publishing, 2018.

[23] Y. Wang, L. Zhao, Z. Song, and M. Wang, "Organ at risk segmentation in head and neck CT images by using a two-stage segmentation framework based on 3D U-Net," arXiv, 2018.

[24] E. Tappeiner et al., "Multi-organ segmentation of the head and neck area: an efficient hierarchical neural networks approach," Int. J. Comput. Assist. Radiol. Surg., vol. 14, no. 5, pp. 745–754, 2019, doi: 10.1007/s11548-019-01922-4.

MULTILEVEL COLLISION CONTROL OVER SCALABLE WIRELESS SENSOR NETWORKS

Shilpy Ghai[1], Dr. Vijay Kumar[2]
[1]*PhD Research Scholar, Department of Computer Science and Engineering, MMEC, Maharishi Markandeshwar Deemed to be University, Mullana,Ambala*
[2]*Department of Computer Science and Engineering, MMEC, Maharishi Markandeshwar Deemed to be University, Mullana,Ambala*
[1]ghai.shilpy2010@gmail.com
[2]katiyarvk@mmumullana.org

Abstract- In a wireless sensor network (WSN), parameters like sensor density may vary at any time (as per requirement) and WSN must be able to perform under the constraints of this scalable parameter. As the sensor density increases, simultaneous access to the shared channel may cause collision thus lead to packet drop/retransmission/extra control overhead etc. All these factors degrade the performance of the routing protocol. To manage the collision, in this paper, a collision control scheme is embedded with different protocols (LEACH/ PEGASIS/ TEEN) and its performance is analyzed under the constraints of scalable sensor density that varies from 50 sensors to 600 sensors using various performance parameters (i.e. Throughput/End-to-End Delay/Alive Sensors/Energy Consumption etc.).

Keywords- WSN, Scalability, Reliability, Resource optimization

I. INTRODUCTION

WSN comprises self-governing devices, called sensors that exchange the data using shared resources i.e. radio channel/bandwidth etc. Limited resources are used excessively used as the sensor density increases and it affects the performance of routing protocol as well as also degrades the lifespan of sensors.

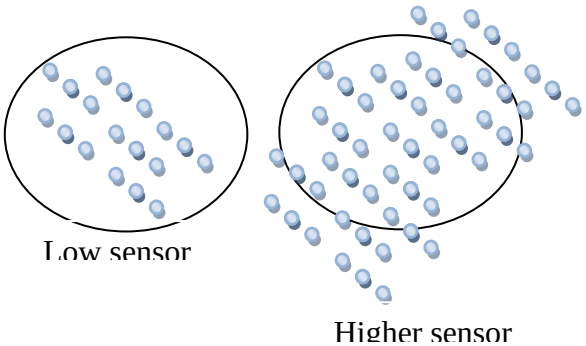


Fig. 1: Low/High Sensor density over WSN

II. IMPACT OF SENSOR DENSITY OVER DIFFERENT LAYERS

Impact of lower or higher sensor density over the different layers is explained [1-4] in detail below:

A) *Routing Layer:* As the number of sensors increase in a given network, routing overhead varies and complex routing computations may reduce sensor's lifespan. Routing protocol should be able to cope with the variations in sensor density, in order to reduce the extra control overhead. Fig.1 shows that if there are few seniors, than it is easy to maintain routing tables with frequent updates and it will produce the minimal control overhead whereas in case of higher sensor density, more efforts are required to manage the routing tables as well as it will be quite complex to adapt the changes in network topology. So routing protocols must be able to cope with this scalable factor. If the number of sensors increases, this factor also has an impact over MAC layer as discussed below.

B) *MAC Layer:* Shared channel cannot be utilized efficiently under the constraints of sensor density variations. Concurrent access to the shared medium can cause packet collision/drop followed by packet retransmission and all these factors unnecessary consume the energy resources. Data aggregation accuracy also suffers due to collisions. Fig.1 show that in case of minimal sensor density, wireless signal propagation is at optimal level thus reduces the probability of collision whereas in case of higher sensor density, more sensors will try to access the shared channel and probability of collision will increase accordingly and it may result in excessive packet drop/retransmission and finally, due to unfair utilization of resources, lifespan of network is declined.

Above discussed layers are directly affected by sensor density and there is need to resolve these issues. Section 6.2 introduces a scheme that can manage the collision under the constraints of scalable sensor density with respect to different routing protocols.

III. SOLUTIONS FOR LARGE SCALE NETWORK OPERATIONS OVER WSN

O. Singh et al. [5] did a simulation-based analysis using various WSN routing protocol (LEACH/SPIN/TEEN/FAIR/ EAMMH/SEP) under the constraints of scalable factors i.e. network size/node density etc. Simulation outcomes indicate that the performance of FAIR is better as compared to others and EAMMH does not support scalable features.

R. C. A. Alves et al. [6] explored that the performance of WSN can be further extended by integrating it with the Software Defined Network (SDN). The study found that SDN optimizes the resource consumption as well as also improves the overall network performance but the initial setup phase of SDN is a little bit complex. Results also indicate that SDN can also manage the total number of required sinks/processing cycles/control packets etc.

E. Fathelrhman et al. [7] developed a scalable routing solution for WSN by exploiting the operations over different layers. MAC layer uses the sleep cycle to void the collisions over the network. Multiple slots are used to forward the data and finally, the congestion control mechanism is invoked to adjust the current traffic load. Simulation results indicate that it outperforms in terms of higher network lifespan/optimal load balancing/energy consumption under the constraints of sensor node density as compared to traditional LEACH variants (LEACH/Multi-LEACH/CELL-LEACH).

P. Engelhard et al. [8] introduced a switch-based solution to minimize the collision at the MAC layer by optimizing the overhearing of frames. The analysis shows that it does not affect the performance of other layers and can be implemented for other MAC protocol variants.

H. Singh et al. [9] developed a clustering scheme for WSN that can estimate the number of the required cluster as per current requirements and clusters are formed by calculating the inter and intra distance between nodes. Simulation results show its performance in terms of optical resource consumption under the constraints of scalable network parameters and it can be further enhanced using multi-hop and multi-level clustering.

K. Shrivastav et al. [10] investigated the issues related to the Internet of Thing (IoT) based WSN and developed a scalable protocol that estimates the transmission range, residual energy and the distance between intermediate nodes to form the clusters. Simulation results indicate that it is more reliable and supports scalability features as compared to existing IoTs and it can be extended to support the heterogeneous WSNs.

A. Rajput et al. [11] introduced a fuzzy logic-based clustering scheme that uses input parameters (i.e. the coordinates of intermediate nodes and the number of the cluster over a given coverage area) for the fuzzy algorithm. The analysis shows that transmission under these constraints enhances the overall network performance as well as the survival interval of the sensors also. Its scope can be further extended using mobile sinks.

H. Singh et al. [12] investigated the issues related to routing over large scale WSN and presented a clustering scheme that forms the clusters based on node degree, residual energy and intermediate distance between them and finally, swarms optimization is used to balance the load over the network thus results in extended network lifespan. As per the simulation results, it retains the network performance under the constraints of scalable network size/ node densities etc. It can be further extended for hierarchical routing.

H. Singh et al. [13] introduced a hierarchical clustering based scheme that subdivides the entire WSN into multiple layers. Cluster heads are selected in each layer and load is equally distributed over each layer. Simulation results show that the optimization of Cluster head selection enhances the network lifespan as compared to existing schemes.

S. N. Mishra et al. [14] presented a cluster-based routing solution calculates the weight of each node based on various parameters (residual energy/node degree/transmission quality). The cluster head is selected based onthe highest weight. Simulation analysis shows its performance in terms of higher Throughput/Packet Delivery Ratio as compared to existing schemes (Survival Path Routing /Predictable energy-aware routing)

K. Kalaivanan et al. [14] developed a routing solution that forms the clusters using fuzzy logic using different input parameters i.e. residual energy, node's speed, number of intermediates nodes and connection time and a tree is built for data forwarding to reduce the delay factor. Simulation results show that it supports reliable/scalable routing under the constraints of the high mobile environment and it outperforms in terms of resource consumption/minimum delay/higher packet delivery ratio/as compared to the existing scheme (Mobility Based Clustering/optimized zone-based routing/Weighted Clustering Algorithm).

M. R. Senouci et al. [15] proposed a weighted sampling scheme for large scale WSN to reduce the computational overhead. It recognizes the region of interest over a specific coverage area for sampling thus reduces the computational overhead. It can be further extended to resolve connectivity issues.

G. P. Gupta et al. [16] used mobile agents for load balancing and data aggregation process over large scale WSN. Mobile agents build the optimal path by neglecting non-responsive intermediate nodes and after that load balancing is initiated for selected nodes over that path. Simulation results show the performance of agents in terms of optimal trip time/extended lifespan of the network under the scalable environment. Its performance can be further enhanced by introducing a sleep scheduling method and it can also be implemented for underwater WSN.

M. Koupaee et al. [17] introduced a fusion-based data forwarding method for WSN. Before transmission, it estimates the waiting interval and convergence conditions of the nodes and neighbors with idle conditions are selected for data

forwarding. Simulation results show its performance in terms of enhanced network performance and it can be further implemented for mobile WSNs.

Y. Touati et al. [18] investigated the issues of routing over large scale WSN. The study found various factors that can degrade the network per romance i.e. compatibility of routing protocols with the applications, excessive resource consumption due to complex computations and extra control overhead due to network size/topology etc. The analysis shows that all the above discussed factors can be optimized by selecting compatible applications and routing protocols.

M. A. Merzoug et al. [19] introduced a topology independent data aggregation over large scale WSN. For each aggregation cycle, it establishes new paths to avoid the link breaks and as per requirements, paths can be extended and data is forwarded hop by hop. Simulation results show that it can adapt to the topological updates and supports reliable transmission under the constraints of a scalable environment.

M. A. Merzoug et al. [20] investigated the issues related to data aggregation over large scale WSN andintroduced a searching method that builds optimal and shortest paths for structure free data collection. Single hop is used for forwarding purpose thus reduces the overall control overhead. Simulation results indicate its performance in terms of improved network performance.

IV. MULTILEVEL COLLISION CONTROL SCHEME FOR WSN

Wireless Sensor Network: WSN
Sensor Node: Snd
Collision: Cl
Collision Level: Clv
Collision Count: Cc
Packet Retransmission: Prt
Collision Threshold: ClTH
Packet Drop: Pktd
Routing Protocol: Rtp
Data:Dt
Interval: Itv
Initialize WSN
Initialize Snd
Initialize Rtp
Define ClTH
Set Cl=Cc=Pktd=0

Phase I: Initiate ClusterHead (CH) selection:

First of all, WSN is initialized and optimal CH(s) are elected and the joining process is invoked for the neighbors.

Phase II: Check for collision level at member level:

In a certain time slot, each member senses the shared channel to forward the data to its CH but simultaneously attempts to access the channel may cause collision thus results in packet drop. So at the node level, collision status is analyzed and if it exceeds from the threshold than members are identified those are responsible for the collision and finally, their status is marked as idle to resolve the collision and during this, data forwarding to CH_i is also halted.

If (Snd_i->Cl)
  Update (Snd_i->Cc)
  Update(Snd_i->Pktd)
End if
If (Snd_i->Cc > ClTH)
  Set Snd_i->Clv=HIGH
Status(Snd_i,IDLE, Itv)
Forward (CHi, Snd_i->Dt, FALSE)
Else If (Snd_i->Cc < ClTH)
  Update (Snd_i->Cc)
  Set Snd_i->Clv=LOW
  Forward (Snd_i, CHi->Dt, TRUE)
End if

Phase III: Reschedule the retransmission:

If (Snd_i->Pktd &&Snd_i->Clv=LOW)
  Retransmission (Snd_i->Pktd, TRUE, Itv)
End If

Phase IV: Check for collision level for ClusterHead(s):

At CH level, collision and the packet drop is monitored. If collision level increases from the collision threshold then mark its level as HIGH otherwise set it to the LOW level. In case of a higher collision level, stop the data forwarding to the base station to avoid packet loss and the number of retransmissions.

If (CH_i->Cl)

Update (CH_i->Cc)
Update(CH_i->Pktd)
End if
If (CH_i->Cc > ClTH)
Set CH_i->Clv=HIGH
Forward (BS, CH_i->Dt, FALSE)
Else If (CH_i->Cc < ClTH)
Update (CH_i->Cc)
Set CH_i->Clv=LOW
Forward (BS, CH_i->Dt, TRUE)

End if

Phase V: Reschedule the retransmission:
If (CH_i->Pktd==TRUE&&CH_i->Clv==LOW)
Retransmission(CH_i->Pktd, TRUE, Itv)
End If

As required, Phase II and Phase III are repeated.

V. SIMULATION SCENARIO

TABLE 1
SIMULATION CONFIGURATION

Simulation Parameters	Parameter Values
Routing Protocol	LEACH/PEGASIS/TEEN
Terrain	1000x1000
MAC Protocol	Mac/Sensor
Node Density	50/100/200/400/600
Propagation Model	TwoRay Ground
Data Type	CBR
Sampling Interval	1.0 ms
Simulation Time	600 seconds
Network Simulator	NS-2.34
Initial Energy	10.0j
rxPower	1
txPower	1
IFQ	50
Antenna Type	Omni
Simulation Scenario(s)	a. The normal scenario for LEACH/PEGASIS/TEEN (without Collision control scheme) b. Collision Control Scheme for LEACH/PEGASIS/TEEN

As per Table 1, NS-2.34 was used for analysis purpose with different parameters. Routing protocol is LEACH and its performance was analyzed under various simulation scenarios i.e.without Collision control scheme/with Collision control scheme, Terrain size is 1000x1000, MAC Protocol is Mac/Sensor, Node Density is 50/ 100/ 200, Propagation Model is TwoRay Ground, Data Type is CBR , Sampling Interval 1.0ms, Simulation Time is 600 seconds, Initial Energy 10.0j, rxPower/txPower is 1, IFQ 50, Antenna Type is Omni.

VI. SIMULATION PERFORMANCE AND RESULT ANALYSIS

Following section shows the performance analysis of collision control scheme using different protocols with above discussed parameters.

1) Throughput

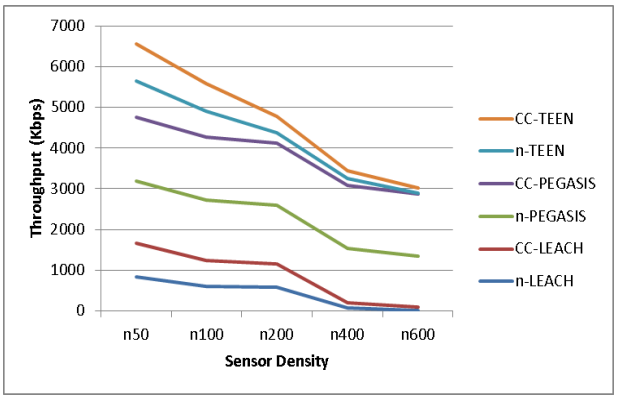


Fig. 2. Throughput

Fig. 2 displays the Throughput of various routing protocols i.e. LEACH/PEGASIS/TEEN under the constraints of sensor density that varies from 50-600.

In the case of sensor density 50, without using the collision control scheme, it is 829.39 Kbps for LEACH, 1524.65 Kbps for PEGASIS and 869.37 Kbps for TEEN.

Using a collision control scheme, it is 845.42 Kbps for LEACH, 1568.34 Kbps for PEGASIS and 917.32 Kbps for TEEN.

In case of sensor density 100, without using collision control scheme, it is 610.81Kbps for LEACH, 1474.6 Kbps for PEGASIS and 623.88Kbps for TEEN.

Using a collision control scheme, it is 638.42Kbps for LEACH, 1550.14Kbps for PEGASIS and 689.5Kbps for TEEN.

In the case of sensor density 200, without using the collision control scheme, it is 582.3Kbps for LEACH, 1423.62 Kbps for PEGASIS and 256.18Kbps for TEEN.

Using a collision control scheme, it is 583.84 Kbps for LEACH, 1541.04 Kbps for PEGASIS and 390.21Kbps for TEEN.

In the case of sensor density 400, without using the collision control scheme, it is 82.22Kbps for LEACH, Kbps for 1351.72PEGASIS and 181.24Kbps for TEEN.

Using collision control scheme, it is 111.94Kbps for LEACH, 1531.94Kbps for PEGASIS and 193.75 Kbps for TEEN.

In the case of sensor density 600, without using the collision control scheme, it is 20.28Kbps for LEACH, 1246.13 Kbps for PEGASIS and 30.84Kbps for TEEN.

Using a collision control scheme, it is 78.01Kbps for LEACH, 1528.3Kbps for PEGASIS and 114.52Kbps for TEEN.

2) End to End Delay

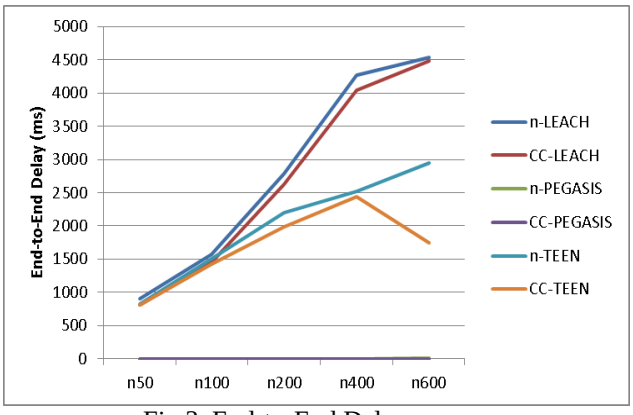


Fig 3. End-to-End Delay

Fig. 3 displays the End-to-End Delay of various routing protocols i.e. LEACH/PEGASIS/TEEN under the constraints of sensor density that varies from 50-600.

In the case of sensor density 50, without using the collision control scheme, it is 907.724ms for LEACH, 1.00E+00ms for PEGASIS and 823.213ms for TEEN.

Using a collision control scheme, it is 808.587ms for LEACH, 1.00E+00ms for PEGASIS and 812.072ms for TEEN.

In the case of sensor density 100, without using the collision control scheme, it is 1566.33ms for LEACH, 1.51E-05ms for PEGASIS and 1500.46ms for TEEN.

Using a collision control scheme, it is 1448.42ms for LEACH, 1.43E-05ms for PEGASIS and 1421.56ms for TEEN.

In the case of sensor density 200, without using the collision control scheme, it is 2787.87ms for LEACH, 3.46E-05ms for PEGASIS and 2201.39ms for TEEN.

Using a collision control scheme, it is 2629.16ms for LEACH, 3.20E-05ms for PEGASIS and 1989.3ms for TEEN.

In the case of sensor density 400, without using the collision control scheme, it is 4263.155ms for LEACH, 6.82E-05ms for PEGASIS and 2515.44ms for TEEN.

Using a collision control scheme, it is 4038.57ms for LEACH, 6.02E-05ms for PEGASIS and 2437.77ms for TEEN.

In the case of sensor density 600, without using the collision control scheme, it is 4534.373ms for LEACH, 10.00012ms for PEGASIS and 2950.33ms for TEEN.

Using a collision control scheme, it is 4485.47ms for LEACH, 9.52E-05ms for PEGASIS and 1740.91ms for TEEN.

3) Energy Consumption

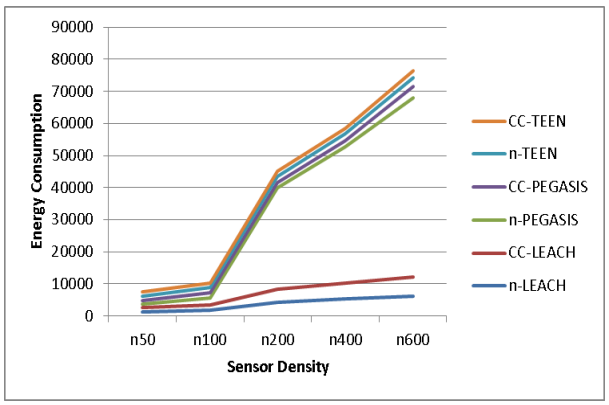


Fig 4. Energy Consumption

Fig. 4 displays the energy consumption of various routing protocols i.e. LEACH/PEGASIS/TEEN under the constraints of sensor density that varies from 50-600.

In the case of sensor density 50, without using the collision control scheme, it is 1264.832 for LEACH, 1259.003 for PEGASIS and 1278.849569 for TEEN.

Using a collision control scheme, it is 1230.359for LEACH, 1113.284456 for PEGASIS and 1278.771 for TEEN.

In the case of sensor density 100, without using the collision control scheme, it is 1811.283for LEACH, 1974.006 for PEGASIS and 1746.394881for TEEN.

Using a collision control scheme, it is 1726.436 for LEACH, 1672.248169 for PEGASIS and 1395.762 for TEEN.

In the case of sensor density 200, without using the collision control scheme, it is 4340.999 for LEACH, 31471.77 for PEGASIS and 1869.038063 for TEEN.

Using a collision control scheme, it is for 4043.21 LEACH, 1771.082258 for PEGASIS and 1716.763 for TEEN.

In the case of sensor density 400, without using the collision control scheme, it is 5272.465 for LEACH, 42501.92 for PEGASIS and 1936.262155 for TEEN.

Using a collision control scheme, it is 5066.416 for LEACH, 1920.168091 for PEGASIS and 1854.511 for TEEN.

In the case of sensor density 600, without using the collision control scheme, it is 6195.166 for LEACH, 55783.71 for PEGASIS and 2702.921521 for TEEN.

Using a collision control scheme, it is 5932.683for LEACH, 3708.811134 for PEGASIS and 2175.264 for TEEN.

4) Number of Alive Sensors

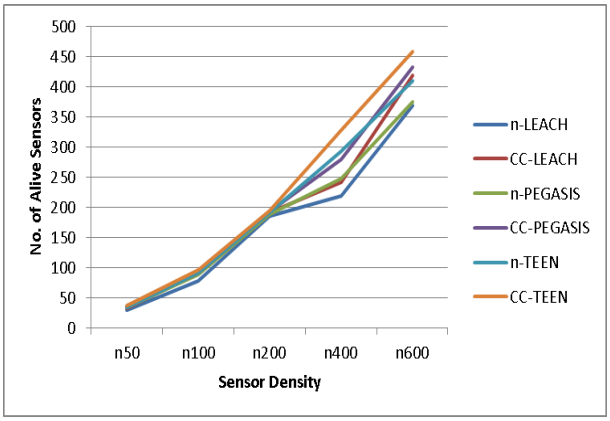


Fig 5. No. of Alive Sensors

Fig. 5 displays the number of alive sensors using various routing protocols i.e. LEACH/PEGASIS/TEEN under the constraints of sensor density that varies from 50-600

In the case of sensor density 50, without using the collision control scheme, these are 29 for LEACH, 34 for PEGASIS and 34 for TEEN.

Using a collision control scheme, these are 33 for LEACH, 35 for PEGASIS and 37 for TEEN.

In the case of sensor density 100, without using the collision control scheme, these are 78 for LEACH, 89 for PEGASIS and 91 for TEEN.

Using a collision control scheme, these are 92 for LEACH, 92 for PEGASIS and 96 for TEEN.

In the case of sensor density 200, without using the collision control scheme, these are 186 for LEACH, 189 for PEGASIS and 191 for TEEN.

Using a collision control scheme, these are 191 for LEACH, 193 for PEGASIS and 194 for TEEN.

In the case of sensor density 400, without using the collision control scheme, these are 219 for LEACH, 247 for PEGASIS and 293 for TEEN.

Using a collision control scheme, these are 241 for LEACH, 279 for PEGASIS and 327 for TEEN.

In the case of sensor density 600, without using the collision control scheme, these are 369 for LEACH,375 for PEGASIS and 409 for TEEN.

Using a collision control scheme, these are 419 for LEACH, 432 for PEGASIS and 458 for TEEN.

VII. CONCLUSION

In this paper, issues and solutions related to WSN scalability was explored and a collision control scheme was introduced. Its performance was analyzed using different routing protocols (LEACH/PEGASIS/TEEN) protocol under the constraints of sensor density that varies from 50 to 600. Simulation outcomes indicate the impact of sensor density and collision level over the performance of the protocol. It can be observed that various parameters are degraded due to collisions caused by scalable sensor density. Without managing the collision level, Throughput of each protocol is degraded as well as End-to-End Delay and energy consumption reaches its peak level thus reduces the overall lifespan of the sensors. In the case of the collision control scheme, it can be analyzed that there is a significant improvement in all performance parameters as well as it also enhances the overall lifespan of the sensors.

Finally, it can be concluded that the performance of the protocols and the sensor's life can be enhanced through an efficient collision management scheme. It can be further extended for other MAC and routing protocols.

REFERENCES

1. I. Snigdh, D. Gosain, "Analysis of scalability for routing protocols in wireless sensor networks", Vol.127, Optik, Elsevier-2016, pp.2535-2538.
2. K.P Noufal, "Wireless Sensor Networks – Scalability and Performance Issues: A Review", Vol.6 (1), IJCST-2015, pp.139-140
3. S. V. Dhage, A. N. Thakre, S. W. Mohod, "A Review on Scalability Issue in Wireless Sensor Networks", Vol.1 (10), IJIRAE-2014, pp.463-466
4. R.sudha, M. Infant Angel, "Wireless Sensor Networks – Scalability and Reliability Issues: A survey", Vol.7 (4), IJETTCS-2018, pp.1-5.
5. O. Singh, V, Rishiwal, "On the scalability of routing protocols in WSN",3rdInternational Conference on Advances in Computing,Communication & Automation (ICACCA), IEEE-2017, pp.1-6.

6. R. C. A. Alves, D. A. G. Oliveira, G. A. N. Segura, C. B. Margi, "The Cost of Software-Defining Things: A Scalability Study of Software-Defined Sensor Networks", Vol.7, IEEE Access-2019, pp.115093-115108.
7. E. Fathelrhman, A. Elsmany, M. A. Omar, T. C. Wan, A. A. Altahir, "EESRA: Energy Efficient Scalable Routing Algorithm for Wireless Sensor Networks", IEEE Access-2019 | Vol.7, pp. 96974-96983.
8. P. Engelhard, A. Zachlod, J. Schulz-Zander, S. Du,"Toward scalable and virtualized massive wireless sensor networks", International Conference on Networked Systems (NetSys), IEEE-2019, pp.1-6.
9. H. Singh,D. Singh, "An Energy-Efficient Scalable Clustering Protocol for Dynamic Wireless Sensor Networks", Wireless Personal Communications,Vol.109, Springer-2019, pp.2637–2662.
10. K. Shrivastav, K. D. Kulat, "Scalable Energy Efficient Hexagonal Heterogeneous Broad Transmission Distance Protocol in WSN‐IoT Networks", Journal of Electrical Engineering & Technology, Vol.15, Springer-2020, pp.95–120.
11. A. Rajput, V. B. Kumaravelu, "Scalable and sustainable wireless sensor networks for the agricultural application of Internet of things using fuzzy c-means algorithm", Sustainable Computing: Informatics and Systems, Vol.22, Elsevier-2019, pp.62-74.
12. H. Singh, D. Singh, "Multi‐level clustering protocol for load‐balanced and scalable clustering in large‐scale wireless sensor networks", The Journal of Supercomputing, Vol.75, Springer-2019, pp.3712–3739.
13. H. Singh, D. Singh, "Concentric Layered Architecture for Multi-Level Clustering in Large-Scale Wireless Sensor Networks", International Conference on Secure Cyber Computing and Communication (ICSCCC), IEEE-2018, pp. 467-471.
14. S. N. Mishra, M. Elappila, S. Chinara, "Development of Survival Path Routing Protocol for Scalable Wireless Sensor Networks", International Conference on Information Technology (ICIT), IEEE-2018, pp.210-215.
15. K. Kalaivanan, V. Bhanumathi, "Reliable location-aware and Cluster-Tap Root based data collection protocol for large scale wireless sensor networks", Journal of Network and Computer Applications, Vol.11815, Elsevier-2018, pp.83-101.
16. M. R. Senouci, H. E. Lehtihet, "Sampling-based selection-decimation deployment approach for large-scale wireless sensor networks", Ad Hoc Networks, Vol.75–76, Elsevier-2018, pp.135-146.
17. G. P. Gupta, M. Misra, K. Garg, "Towards scalable and load-balanced mobile agent-based data aggregation for wireless sensor networks Computers & Electrical Engineering, Vol.64, Elsevier- 2017, pp.262-276.
18. M. Koupaee, M. R. Kangavari, M. J. Amiri, "Scalable structure-free data fusion on wireless sensor networks", The Journal of Supercomputing vol.73, Springer-2017, pp.5105–5124.
19. Y. Touati, A. A. Chérif, B. Daachi, "Adaptive Routing for Large-Scale WSNs", Energy Management in Wireless Sensor Networks", Elsevier-2017, pp.53-63.
20. M. A. Merzoug, A. Boukerche, A. Mostefaoui, "Efficient informationgathering from large wireless sensor networks", Computer Communications, Vol.132, Elsevier-2018, pp.84-95.

A SURVEY ON VARIOUS GESTURE RECOGNITION TECHNIQUES FOR UAV CONTROL

Surbhi Kapoor[1], Akashdeep Sharma[2] and Amandeep Verma[3]

[1] Research Scholar UIET, Panjab University, Chandigarh, India, [2] Assistant Professor, UIET Panjab University, Chandigarh,India, [3]Assistant Professor, UIET, Panjab University, Chandigarh,India

[1]surbhi31892@gmail.com

[2]akashdeep@pu.ac.in

[3]amandeepverma@pu.ac.in

ABSTRACT- Drones are becoming ubiquitous in the society due its numerous applications facilitating human drone interaction (HDI). The aim of HDI is to make the communication with drone as natural as the interaction between humans. Gestures are considered as more natural, creative and intuitive way to communicate with drones. Gesture recognition approaches can be categorized into vision based approaches and sensor based approaches. Vision based approaches use cameras and video recorders to identify the behavior of the object and changes in the environment whereas sensor based approaches require a third party device to observe the changes in the environment. Sensor based approaches can be further categorized into wearable approaches and non wearable approaches. The core focus of this study is to provide a detailed study on vision based approaches and non wearable sensor based approaches to control the drones via gestures, not requiring the human to wear any device on the body in order to get recognized by the drone. A comparative study of the articles available in the literature based on the number of gestures, types of gestures and the corresponding body parts required to recognize the gestures is also presented in the paper. Attempts have also been made to highlight the publically available datasets for human drone interaction via gestures. This paper will help the researchers to get an insight of the studies available in this area as there is no survey presented in this field.

KEYWORDS- Drones, gestures, wearable, non wearable, vision based, sensor based, behavior.

I. INTRODUCTION

Human Robot Interaction (HRI) is an emerging research area and plays a vital role in the field of psychology, social science, computer science, robotics and engineering. The objective of HRI is to furnish the robots with various abilities that ease their interaction with humans. Human Drone Interaction (HDI) is a branch of HRI which focuses on flying robots that is drones. UAV's have been used for more than two decades, however most recent couple of years have been huge regarding drone adoption. Drones have become the backbone of many growing industries. In recent years, drones have become a popular choice for aerial photography, defense services and surveillance due to its mobility and economic feasibility. Unlike traditional cameras, which are fixed at one position and do not cover a larger area, drones can fly covering a wider range of view. Earlier, drones were used for commercial and military purposes only but now days, they are also used by hobbyists as toys. The drones are able to carry sufficient payload(sensors) due to growing technology and hardware which makes them useful for the researchers. HDI has its applications in four major areas as shown in figure 1.

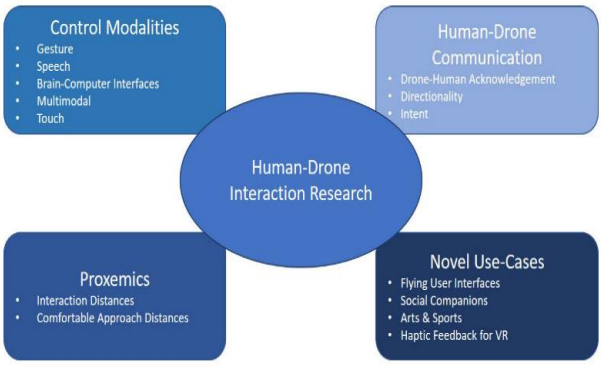


Fig. 1 Applications of Human Drone interaction [1]

Earlier, drones were controlled by ground stations[2][3] and a remote control[4] but due to the advancement in technology, drones can be controlled by natural user interfaces[5] which facilitates interaction with the drone through gestures[6], touch[7][8], gaze[9][10], brain signals[11] and speech[12][13]. The discussion in this paper is limited to controlling the drones via gestures. A gesture can be defined as a basic movement of the body part of a person, for example raising an arm. Gesture recognition is a challenging area in the field of human drone interaction. UAV's ability to identify the commanding actions of humans and take desirable actions is the upmost need in surveillance. The rest of the paper is organized as follows. Section II contains the discussion about gesture recognition approaches. Section III discusses the datasets and section IV summarizes the paper.

II. GESTURE RECOGNITION APPROACHES

Gesture Recognition is grabing more interest from the research community worldwide. Incorporating gestures in HDI is new area of research. Gesture recognition approaches can be broadly classified into the following two categories.

- Vision based approaches
- Sensor based approaches

Vision based approaches are gaining more impetus as the human need not carry any external device and can interact with the drone in a more interactive and natural way. Vision based algorithms use cameras and video recorders to identify the behavior of the object and changes in the environment. These algorithms are discussed in detail in section 2.1.

Sensor based approaches can be further categorized into wearable and non wearable approaches. Wearable approaches are those which require a sensor to be worn on the human body, rooted straight under the skin or indirectly embedded in devices which are more likely to be carried by the human like phone[14][15], forearm bands[16] smart watches[17] etc. The devices or sensors which are embedded are accelerometer[18], gyroscope[19], magnetometer[20] etc. This is a very tedious task due to the discomfort caused by wearing the devices and the cost of these devices add to the inconvenience. Non wearable approaches require a third party device to capture the images and convert them into depth images using depth sensors. Microsoft Kinect Sensor [21]and Leap Motion Controller[22][23] are two popular devices which fall under this category. Microsoft Kinect Sensor is motion sensing device designed for image capturing, voice recognition, gesture recognition and producing body skeletons. It consists of motorized pivot, RGB camera, a depth sensor and microphone. In contrast to Kinect sensor, leap motion controller is a USB powered device, specifically designed to track the movements of the hands rather than capturing the whole body. It consists of two monochromatic IR cameras and three IR LEDs to sense the motion made by hands.

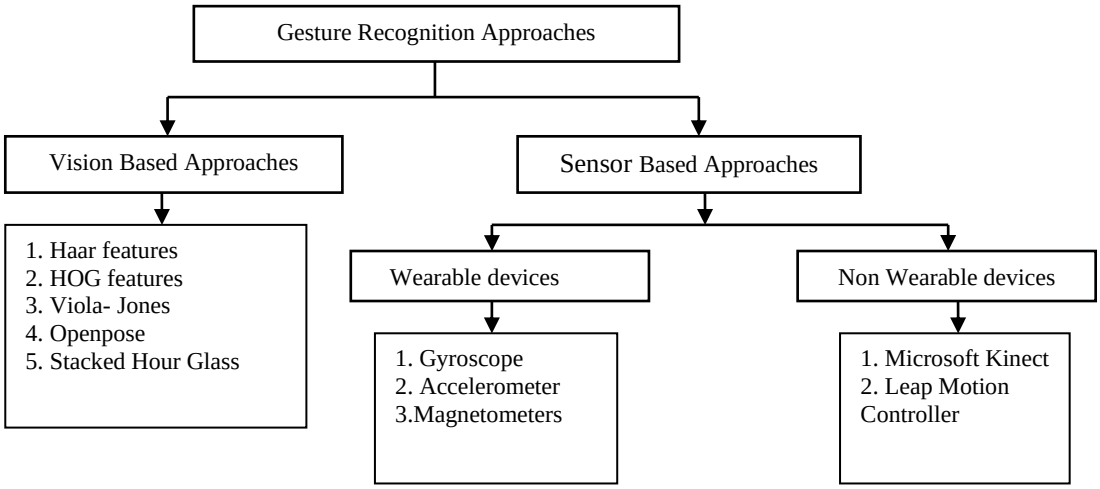


Fig. 2 Categorization of gesture recognition approaches

A. Vision Based Approaches

Vision based approaches uses images or video sequences captured by the cameras to recognize the gestures used to control the drone. These approaches do not require any human to wear any uncomfortable devices on their body. This is the major reason behind the gaining acceptability of these approaches as compared to sensor based approaches. This section provides a detailed discussion of the studies using vision based approaches.

A. G. Perera *et al.*[24] have used open pose to estimate joints on the human body followed by PCNN [25] to identify the gestures used to control the drone. The authors have introduced a new dataset, UAV - GESTURE for controlling UAV and recognizing the gestures and evaluated their dataset by obtaining an accuracy of 91.9%. Heba Aly *et al.*[26] proposed Maximum Entropy Markov Model which is a combination of hidden markov model and maximum entropy model to control the drone based on human skeleton information extracted from openpose[27] and obtained an accuracy of 80% J. Bolin *et al.*[28] used stacked hourglass[29] instead of openpose model to extract the human skeleton for pose identification. Finite state machine is used to select the actions to be performed by the drone. Two trails of experiments were conducted, each trail containing 30 gestures. Four parameters were considered for evaluation namely false positive rate, false negative rate, misinterpreted gestures and successful gestures. The average accuracy reported in case of successful gestures is 98.33%. Yale Song *et al.*[30] proposed a method for tracking body and hands. Upper body of the person is detected by multi-hypothesis Bayesian inference framework whereas the hand poses are estimated by HOG [31] features. They also introduced new multi signal gesture dataset, NATOPS and experimented their proposed methodology on their dataset. The authors used SVM classifier for hand pose classification and achieve

an accuracy of 99.94%. Gabriele Costante *et al*.[32] claimed that drones may not be able to generalize gestures as they may get confused by the way different humans perform the same gesture. In order to resolve this issue, they came up with the concept of personalized gestural interface based on transfer learning and Histogram of optical flow(HOOF)[33] features in which the system first detects the person and learns its identity followed by the identification of the gesture performed by the user by selecting the user specific classifier. The experiments were conducted on Keck gesture dataset and in real time. The comparison was drawn considering various combinations of the datasets. The average accuracy achieved by the authors is 87.9%. Nagi *et al*.[34] used the face estimates calculated by OpenCv's Viola- Jones face detector[35] and estimated the orientation of hand with respect to face location of the human. Natarajan *et al*.[36] used HAAR features to recognize the hand gestures to control the drones and achieved an average accuracy of 90%. Monajjemi *et al*.[37] used the hand gestures to command a group of UAV's which communicate with each other through a wireless network. Experimental results yielded an average accuracy of 76%. Monajjemi *et al*.[38] also presented a system in which a waving gesture is used to grab the attention of a drone in an outdoor environment. The entire idea of detecting a gesture is based on periodicity of the actions performed by the person which is determined by performing the frequency domain analysis on each track obtained from detecting and tracking a person in the UAV imagery. The system differentiates the other periodic movements i.e. running and walking people but confuses other stationary periodic movements such as digging action performed by a person. The overall accuracy reported by the authors is 81.8%. Jangwon Lee *et al*.[39] came up with the idea of predicting future hand gestures for human drone interaction by designing a convolutional network with VGG-16 [40] as a base network and testing it on their newly introduced dataset and obtaining an accuracy of 99.1%. A.Maher *et al*.[41] modified Yolov2 [42] network for detecting hand and face of the human. The detection results are used as an input to the gesture detection module based on angle ratio between hands and the face of the user. The authors evaluated their approach on their own dataset containing 6223 images having 23 participants collected in both indoor and outdoor settings. The map reported by the authors is 90.87%. Sepehr MohaimenianPour and Richard Vaughan[43] also used Yolov2 to detect hands and face of the human and the detection results are used for recognizing the gestures by considering the relative position of the face to the hands. J.Zhang *et al*.[44] also proposed a similar methodology as[13] for gesture recognition. Instead of using Yolov2, they have used mobile-Net SSD[45] for detection. The gestures are interpreted based on the angle of each hand to that of the face of the user. The authors also introduced a dataset containing 8 gestures, and 200 images for each gesture. The fps_mean obtained by the authors is 27.74. M. Monajjemi *et al*.[46] presented a system in which the person performs a waving gesture to grab the human attention. An appearance based tracker and cascade controller is used to identify the gestures of the humans. The authors reported a success rate of 71%. K. Haratiannejadi and R.R.Selmic [47] used RCNN[48] and skeleton detection module for detecting humans and their hands. A smart glove having four flex sensors and a motion processing unit is worn on the left hand to classify the left hand gesture, whereas right hand gesture are identified by CNN module without using any wearable sensor. The experiments were conducted in a multi subject environment, obtaining an average accuracy of 94.96%. Table 1 presents the analysis of articles based on types of gestures and the corresponding body parts required for drone control.

TABLE 1
ANALYSIS OF ARTICLES BASED ON TYPES OF GESTURES AND THE CORRESPONDING BODY PARTS

Reference	Year	Type of gestures	Body part	No.of gestures	Classifier
[24]	2019	Static, Dynamic	Full body	13	P-CNN
[26]	2017	Static, Dynamic	Full body	9	Naive-Bayes-Nearest-Neighbor
[28]	2017	Static	Full body	10	Manual classification
[30]	2011	Dynamic	Body, Hands	24	SVM
[32]	2014	Dynamic	Full body,face	5	SVM
[34]	2014	Static	Face, Hands	2	Haar classifier
[36]	2018	Static	Hands	5	AdaBoost
[37]	2013	Dynamic	Face, Hands	4	Manual classification
[39]	2018	Dynamic	Hands	5	CNN
[46]	2016	Dynamic	Face,Hand	2	Not Specified
[47]	2020	Static, Dynamic	Hands	20	CNN
[49]	2020	Static, Dynamic	Face, Hands	9	CNN

B. Sensor Based Approaches

Wearable sensor based approaches require the human to wear an external device to control the drone whereas non wearable sensor based approaches require a third party device to capture the human body. In this section, a detailed discussion about non wearable sensor based approaches is provided.

Popov *et al*.[50] controlled the flight operation of a drone by using the body postures (human skeleton) obtained from Microsoft Kinect sensor. Some rules are formed on the basis of the distance and angle formed between the limbs of the skeleton. The UAV takes the necessary action based upon the rules. The authors have presented the graphical results considering the parameters like roll angle, gesture control signal, velocity etc. P. Pfeil[51] *et al.* presented a study to

control the drone with 3D spatial interaction metaphors using the Kinect sensor. The authors have presented qualitative results in the form of comfort, naturalness, overall perception and likeability. Lichtenstern *et al.* [52] presented a system in which gestures recognized by UAV equipped with Microsoft Kinect is used to command the team of UAV's. The authors have not disclosed the results. Naseer *et al.*[53] described a method that allows a UAV to follow a person by responding to the hand gestures with the help of kinect sensor alongwith OpenNI tracker. The overall performance achieved by the system is 92.5%. Mantecon *et al.*[54] used Kinect sensor 2 along with SVM classifier to recognize the hand gestures. The authors used mean accuracy and standard deviation as the parameters to evaluate their approach. The mean accuracy and standard deviation reported by the authors are 94.62% and 5.24 respectively. B.Hu and J.Wang [55] developed a system in which Leap motion controller is used to recognize hands rather than a Kinect sensor. Three variants of deep neural networks namely 2-layer fully connected network, 5-layer fully connected network and 8 layer convolutional network are trained and tested for gesture recognition on their newly introduced dataset. The experimental results reveal that 5 layer and 2 layer networks produced an accuracy of 97% whereas on scaled data whereas 8 layer network works better on raw data. A. Mashood *et al.*[56] used Kinect depth sensor to detect and track the human movements followed by Flexible Action and Articulated Skeleton Toolkit' (FAAST)[57] program to from a human skeleton and generate commands to be sent to the UAV. The experiments were conducted in indoor setting having a person performing eleven different gestures and graphical results are presented considering speed, altitude, pitch, yaw and roll angles. S. Zhang *et al.*[58] presented a multimodal method using eyes, voice and gestures to control drones. They used Kinect sensor to extract the body joints on the human body and the angle between the body joints is used to recognize the gesture. The authors claim that success rate in all three modalities is above 90%. Sarkar *et al.*[59] explored a study to control the drone connected to the ground station via wifi. Hand gestures are recognized by the leap motion controller and transmit to the groundstation. Another work for recognizing the hand gestures using web technology is presented in [60] the authors presented the graphical results in the form of altitude and control latency. B.A.Tingare *et al.*[61] detected static and dynamic hand gestures based on distance between the finger tips detected by Leap motion Controller. They detected a total of seven gestures out of which two were dynamic gestures. The accuracy obtained for hand gesture recognition using leap motion controller is 95.6% and recognition accuracy using 2D cameras is 80.4%. T.Begum *et al.*[62] evaluated three different deep learning models namely VGG-16, ResNet-50[63] and simple CNN with a bit modification for recognizing the hand gestures. The authors collected a total of 544 images have 17 people as participants performing 4 gestures i.e. stop, forward, right and left. Simple CNN produces the highest accuracy of 92% as compared to other two models. Ramon A. Suarez Fernandez *et al.*[64] presented a natural user interface and a graphical user interface to command the drones using hand gestures, body position, speech and visual markers. The hand gestures are recognized by the use of leap motion controllers. Table 2 presents the list of datasets available for UAV control and gesture recognition. Table 2 presents the analysis of articles based on types of gestures and the corresponding body parts required for drone control.

TABLE 2
ANALYSIS OF ARTICLES BASED ON TYPES OF GESTURES AND THE CORRESPONDING BODY PARTS

Reference	Year	Type of gestures	Body part	No.of gestures	External device	Classifier
[50]	2016	Dynamic	Full body	9	Microsoft Kinect sensor	Manual classification
[51]	2013	Static	Full body	5	Microsoft Kinect sensor	Manual classification
[53]	2013	Static	Full body	3	Microsoft Kinect sensor	Manual classification
[54]	2014	Static,Dynamic	Hand	11	Microsoft Kinect sensor	SVM
[55]	2020	Dynamic	Hand	10	Leap Motion Controller	CNN
[56]	2015	Static	Full body	11	Microsoft Kinect sensor	Manual classification
[59]	2016	Static	Hand	2	Leap Motion Controller	Manual classification
[60]	2018	Static	Hand	9	Leap Motion Controller	Manual classification
[62]	2020	Static	Hand	4	Leap Motion Controller	Manual classification
[64]	2016	Static, Dynamic	Hand	Not specified	Leap Motion Controller	Manual classification

III. DATASETS

Gesture recognition via drones is an emerging research area. However, the studies on human drone interactions are still inadequate. One major reason that contributes to this inadequacy is unavailability of the datasets required for the exploration of this field. The studies available in this area are restricted to either indoor settings or performing static gestures, making it unrealistic for the real world problems. The experiments presented in the available literature are conducted in real time. The researchers have not released the datasets publically. Only two datasets are available publically which can help the researchers to explore this field. NATOPS[30] is a multi signal gesture database for aircraft handling. It contains 24 gestures captured by VICON and stereo cameras having 9600 frames with a resolution of 320x240 pixels captured at 20fps. Videos are collected in an indoor environment involving 20 persons repeating each of 24 gestures 20 times, totaling to 400 samples for each category. Another dataset was released by [24], UAV- Gesture, developed for Gesture recognition and UAV control. The dataset was collected by rotorcraft UAV, flying at a lower altitude in a wheat field (outdoor setting). The videos have a resolution of 1920x1080 captured at 25fps. A total of 13 gestures are recorded with a total duration of 24.76 min. The total frames present in the dataset are 37151. Annotations are available for bounding box and body joint and are performed using VATIC Tool. It is an evident observation that more number of datasets shall be developed by the research community in order to explore this field. Table 3 summarizes the datasets available for UAV control and gesture recognition.

TABLE 3
LIST OF DATASETS AVAILABLE FOR UAV CONTROL AND GESTURE RECOGNITION

Year	Dataset	Description	Format	Resolution	Environment	No.of Gestures	Reference
2019	UAV-Gesture	Gesture Recognition	37151 Frames	1920x1080	Outdoor	13	[24]
2011	NATOPS	Multi-signal gesture database	9600 frames	320x240	Indoor	24	[30]

IV. CONCLUSION

Drones have the capability to fly autonomously or through various control devices like joysticks and certain interfaces like voice, emotions, human brain and gestures. Human Drone Interaction is new area of research having lot of fascinating applications. To the best of our knowledge, this is the first survey which focuses on drone control using gestures. In the present study, a comprehensive survey has been done on the studies focusing on drone control using vision based algorithms and non wearable sensor based approaches, without the need of the human user to wear anything on the body. A detailed discussion about the datasets is also presented in this study. The analysis withdrawn from the papers is that the overall performance of the gesture recognition module highly depends on the extraction of body joints in the primary stage. In case of non wearable sensor based approaches, manual classification is preferred over standard classifiers. Analysis of the articles presented in the paper indicates that hand gestures are most preferred way of interacting with the drones. Additionally, it is observed that only two datasets are available publically for gesture recognition i.e. UAV-GESTURE, recorded in outdoor environment and NATOPS, recorded in indoor environment. More number of datasets shall be developed so as to make advancements in this field of research. It is quite apparent that genuine research focus is required on developing the datasets for controlling the drone via gestures. This study will be useful for the emerging researchers who are interested in exploring the field of human drone interaction.

REFERENCES

[1] D. Tezza And M. Andujar, "The State-Of-The-Art Of Human-Drone Interaction: A Survey," *IEEE Access*, Vol. 7, Pp. 167438–167454, 2019.

[2] Abdelhamid, P. Zong, And B. Abdelhamid, "Advanced Software Ground Station And UAV Development For Nlos Control Using Mobile Communications," *Discrete Dynamics in Nature and Society*,2015.

[3] V. Kangunde, R. S. Jamisola, And E. K. Theophilus, "A Review On Drones Controlled In Real-Time," *International Journal of Dynamics and Control*,pp. 1-5, 2021.

[4] J. M. Fernandez Gonzalez, P. Padilla, J. F. Valenzuela-Valdes, J. L. Padilla, And M. Sierra-Perez, "An Embedded Lightweight Folded Printed Quadrifilar Helix Antenna: UAV Telemetry And Remote Control Systems*," IEEE Antennas and Propagation Magazine*, Vol. 59, no. 3, Pp. 69–76, 2017.

[5] E. Peshkova, M. Hitz, And B. K. Kaufmann, "Natural Interaction Techniques For An Unmanned Aerial Vehicle System," Vol. 16, No. 1, pp. 34–42, 2017.

[6] J. R. Cauchard, L. E. Jane, K. Y. Zhai, And J. A. Landay, "Drone & Me: An Exploration Into Natural Human-Drone Interaction Jessica," *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pp. 361–365, 2015.

[7] H. Kang, H. Li, J. Zhang, X. Lu, And B. Benes, "Flycam: Multitouch Gesture Controlled Drone Gimbal Photography*," IEEE Robotics and Automation Letters*, Vol. 3, No. 4, pp. 3717–3724, 2018.

[8] P. Abtahi, D. Zhao, J. E., And J. Landay, "Drone Near Me: Exploring Touch-Based Human-Drone Interaction," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 1, No. 3, Pp. 1–8, 2017.

[9] J. P. Hansen, A. Alapetite, I. S. Mackenzie, And E. Møllenbach, "The Use Of Gaze To Control Drones," *Proceedings of the symposium on eye tracking research and applications*, pp. 27–34, 2014.

[10] B. N. Pavan Kumar, A. Balasubramanyam, A. K. Patil, B. Chethana, And Y. H. Chai, "Gazeguide: An Eye-Gaze-Guided Active Immersive UAV Camera," *Applied. Sciences.*, Vol. 10, No. 5, 2020.

[11] A. Nourmohammadi, M. Jafari, And T. O. Zander, "A Survey On Unmanned Aerial Vehicle Remote Control Using Brain-Computer Interface," *IEEE Transactions on Human-Machine Systems*, Vol. 48, No. 4, pp. 337–348, 2018.

[12] A. Menshchikov Et Al., "Data-Driven Body-Machine Interface For Drone Intuitive Control Through Voice And Gestures," *IECON 2019-45th Annual Conference of the IEEE Industrial Electronics Society*,pp. 5602–5609, 2019.

[13] R. Contreras, A. Ayala, And F. Cruz, "Unmanned Aerial Vehicle Control Through Domain-Based Automatic Speech Recognition," *Computers*, Vol. 9, No. 3, pp. 1–15, 2020.

[14] B. Zhao, X. Chen, X. Zhao, J. Jiang, And J. Wei, "Real-Time UAV Autonomous Localization Based On Smartphone Sensors," *Sensors*, Vol. 18, No. 12. 2018.

[15] T. Bonny And M. B. Abdelsalam, "Autonomous Navigation Of Unmanned Aerial Vehicles Based On Android Smartphone," *International Journal of Advanced Computer. Science and Appications*, Vol. 10, No. 11, Pp. 589–598, 2019.

[16] A. Stoica, F. Salvioli, And C. Flowers, "Remote Control Of Quadrotor Teams, Using Hand Gestures," *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pp. 296–297, 2014.

[17] S. Byun And S. Lee, "Implementation Of Hand Gesture Recognition Device Applicable To Smart Watch Based On Flexible Epidermal Tactile Sensor Array," *Micromachines*, Vol. 10, No. 10, 2019.

[18] J. Akagi, B. Moon, X. Chen, And C. K. Peterson, "Gesture Commands For Controlling High-Level UAV Behavior," *International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 1023–1030, 2019.

[19] H. Han And S. W. Yoon, "Gyroscope-Based Continuous Human Hand Gesture Recognition For Multi-Modalwearable Input Device For Human Machine Interaction," *Sensors*, Vol. 19, No. 11, 2019.

[20] L. F. Sanchez, H. Abaunza, And P. Castillo, "Safe Navigation Control For A Quadcopter Using User's Arm Commands," *International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 981–988, 2017.

[21] "Kinect-Windows App Development." [Online]. Available: Https://Developer.Microsoft.Com/En-Us/Windows/Kinect/. [Accessed: 19-Feb-2021].

[22] F. Weichert, D. Bachmann, B. Rudak, And D. Fisseler, "Analysis Of The Accuracy And Robustness Of The Leap Motion Controller," *Sensors*, Vol. 13, No. 5, Pp. 6380–6393, 2013.

[23] J. Guna, G. Jakus, M. Pogačnik, S. Tomažič, And J. Sodnik, "An Analysis Of The Precision And Reliability Of The Leap Motion Sensor And Its Suitability For Static And Dynamic Tracking," *Sensors*, Vol. 14, No. 2. Pp. 3702–3720, 2014.

[24] A. G. Perera, Y. W. Law, And J. Chahl, "UAV-GESTURE: A Dataset For UAV Control And Gesture Recognition," *Lecture Notes Computer Science. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, Vol. 11130 LNCS, pp. 117–128, 2019.

[25] G. Cheron, I. Laptev, And C. Schmid, "P-CNN: Pose-Based CNN Features For Action Recognition," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3218–3226, 2015.

[26] H. Aly, C. Arnold, A. Hari, P. Nguyen, And S. Shrestha, "Human Gesture Recognition For Drone Control," 2017.

[27] Z. Cao, H. Gines, T. Simon, S. E. Wei, And Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 43, No. 1, pp. 172–186, 2019.

[28] J. Bolin, C. Crawford, W. Macke, J. Hoffman, S. Beckmann, And S. Sen, "Gesture-Based Control Of Autonomous Uavs," *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, Vol. 3, pp. 1484–1486, 2017.

[29] A. Newell, K. Yang, And J. Deng, "Stacked Hourglass Networks For Human Pose Estimation," *European conference on computer vision*, 2016, Vol. 9912 LNCS, pp. 483–499.

[30] Y. Song, D. Demirdjian, And R. Davis, "Tracking Body And Hands For Gesture Recognition : NATOPS Aircraft Handling Signals Database," *In 2011 IEEE International Conference On Automatic Face And Gesture Recognition And Workshops*, 2011, pp. 500–506.

[31] N. Dalal And B. Triggs, "Histograms Of Oriented Gradients For Human Detection," *In Proceedings - 2005 IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, CVPR2005, 2005, Vol. 1, pp. 886–893.

[32] G. Costante, E. Bellocchio, P. Valigi, And E. Ricci, "Personalizing Vision-Based Gestural Interfaces For HRI With Uavs : A Transfer Learning Approach," *In 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2014, No. Iros, pp. 3319–3326.

[33] R. Chaudhry, A. Ravichandran, G. Hager, And R. Vidal, "Histograms Of Oriented Optical Flow And Binet-Cauchy Kernels On Nonlinear Dynamical Systems For The Recognition Of Human Actions," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1932–1939, 2009.

[34] J. Nagi, A. Giusti, G. A. Di Caro, And L. M. Gambardella, "Human Control Of Uavs Using Face Pose Estimates And Hand Gestures," 2014 *9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*,pp. 252–253, 2014.

[35] P. Viola And M. Jones, "Rapid Object Detection Using A Boosted Cascade Of Simple Features," *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition.*pp. 511–518, 2001.

[36] K. Natarajan, T. H. D. Nguyen, And M. Mete, "Hand Gesture Controlled Drones: An Open Source Library," 2018 *1st International Conference on Data Intelligence and Security (ICDIS)*, pp. 168–175, 2018.

[37] V. M. Monajjemi, J. Wawerla, R. Vaughan, And G. Mori, "HRI In The Sky : Creating And Commanding Teams Of Uavs With A Vision-Mediated Gestural Interface," *In 2013 IEEE International Conference On Intelligent Robots And Systems*, 2013, pp. 617–623.

[38] M. Monajjemi, J. Bruce, S. A. Sadat, J. Wawerla, And R. Vaughan, "UAV, Do You See Me? Establishing Mutual Attention Between An Uninstrumented Human And An Outdoor UAV In Flight," 2015 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3614–3620, 2015.

[39] J. Lee, H. Tan, D. Crandall, And S. Abanovi, "Forecasting Hand Gestures For Human-Drone Interaction," *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 167–168, 2018.

[40] K. Simonyan And A. Zisserman, "Very Deep Convolutional Networks For Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, pp. 1–14, 2014.

[41] A. Maher, C. Li, H. Hu, And B. Z. B, "Realtime Human-UAV Interaction," *Chinese Conference on Biometric Recognition*, Springer, Cham, pp. 511–519, 2017.

[42] J. Redmon And A. Farhadi, "YOLO9000: Better, Faster, Stronger," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6517–6525, 2017.

[43] S. Mohaimenianpour And R. Vaughan, "Hands And Faces , Fast : Mono-Camera User Detection Robust Enough To Directly Control A UAV In Flight," 2018 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5224–5231, 2018.

[44] J. Zhang, L. Peng, W. Feng, And Z. Ju, "Human-AGV Interaction : Real-Time Gesture Detection Using Deep Learning," *International Conference on Intelligent Robotics and Applications*, Springer, Cham, pp. 1–12.

[45] A. G. Howard Et Al., "Mobilenets: Efficient Convolutional Neural Networks For Mobile Vision Applications," *arXiv preprint arXiv*:1704.04861, 2017.

[46] M. Monajjemi, S. Mohaimenianpour, And R. Vaughan, "UAV , Come To Me : End-To-End , Multi-Scale Situated HRI With An Uninstrumented Human And A Distant UAV," *In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4410–4417, 2016.

[47] K. Haratiannejadi And R. R. Selmic, "Smart Glove And Hand Gesture-Based Control Interface For Multi-Rotor Aerial Vehicles In A Multi-Subject Environment," *IEEE Access*, pp. 227667–227677, 2020.

[48] R. Girshick, J. Donahue, T. Darrell, And J. Malik, "Rich Feature Hierarchies For Accurate Object Detection And Semantic Segmentation," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[49] M. A. Kassab, M. Ahmed, And A. L. I. Maher, "Real-Time Human-UAV Interaction : New Dataset And Two Novel Gesture-Based Interacting Systems," Vol. 8, 2020.

[50] V. L. Popov, K. B. Shiev, A. V. Topalov, N. G. Shakev, And S. A. Ahmed, "Control Of The Flight Of A Small Quadrotor Using Gestural Interface," *2016 IEEE 8th International Conference on Intelligent Systems (IS)*, pp. 622–628, 2016.

[51] K. P. Pfeil, S. L. Koh, And J. J. Laviola, "Exploring 3D Gesture Metaphors For Interaction With Unmanned Aerial Vehicles," 2013 *international conference on Intelligent user interfaces*, pp. 257–266, 2013.

[52] M. Lichtenstern, M. Frassl, B. Perun, And M. Angermann, "A Prototyping Environment For Interaction Between A Human And A Robotic Multi-Agent System," *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 185–186, 2012.

[53] T. Naseer, J. Sturm, And D. Cremers, "Followme : Person Following And Gesture Recognition With A Quadrocopter," *In 2013 IEEE International Conference on Intelligent Robots and Systems*, 2013, Pp. 624–630.

[54] T. Mantecon, C. R. Del Blanco, F. Jaureguizar, And N. Garcia, "New Generation Of Human Machine Interfaces For Controlling UAV Through Depth-Based Gesture Recognition," *Unmanned Systems Technology XVI*, vol. 9084, p. 90840C, 2014.

[55] B. Hu And J. Wang, "Deep Learning Based Hand Gesture Recognition And UAV Flight Controls," *International Journal of Automation and Computing*, Vol. 17, No. 1, Pp. 17–29, 2020.

[56] A. Mashood and H. Noura, "A Gesture Based Kinect for Quadrotor Control," *In 2015 International Conference on Information and Communication Technology Research (ICTRC)* , pp. 298–301, 2015.

[57] E. A. Suma, B. Lange, A. Rizzo, D. M. Krum, and M. Bolas, "FAAST: The Flexible Action And Articulated Skeleton Toolkit," *2011 IEEE Virtual Reality Conference*, pp. 247–248, 2011.

[58] S. Zhang, X. Liu, J. Yu, L. Zhang, and X. Zhou "Research On Multi-Modal Interactive Control For Quadrotor UAV," *2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC)*, 2019.

[59] A. Sarkar, K. A. Patel, R. K. G. Ram, And G. K. Capoor, "Gesture Control Of Drone Using A Motion Controller," *In 2016 International Conference on Industrial Informatics and Computer Systems (CIICS)*, 2016.

[60] Z. Zhao, H. Luo, G. H. Song, Z. Chen, Z. M. Lu, And X. Wu, "Web-Based Interactive Drone Control Using Hand Gesture," *Review of Scientific Instruments*, Vol. 89, No. 1, 2018.

[61] M. Tingare, "Controlling The Drone With Hand Gestures By Using LEAP Motion Controller," *International Journal of Pure and Applied Mathematics*, Vol. 118, No. 24, Pp. 1–10, 2018.

[62] T. Begum, I. Haque, And V. Keselj, "Deep Learning Models For Gesture-Controlled Drone Operation." *In 2020 16th IEEE International Conference on Network and Service Management (CNSM)*, pp. 1-7, 2020.

[63] K. He, X. Zhang, S. Ren, And J. Sun, "Deep Residual Learning For Image Recognition," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

[64] R. A. S. Fernandez, J. L. Sanchez-Lopez, C. Sampedro, H. Bavle, M. Molina, And P. Campoy, "Natural User Interfaces For Human-Drone Multi-Modal Interaction," *2016 IEEE International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 1013–1022, 2016.

A COMPARATIVE STUDY OF K-ANONYMITY AND DATA ENCRYPTION, BASED ON TIME AND SPACE

Nishant Agnihotri[1], Aman Kumar Sharma[2]
[1]PhD Research Scholar, Himachal Pradesh University, Shimla.
[2]Professor, Department of Computer Science, Himachal Pradesh University, Shimla.

Abstract: As the usage of the data over the cloud is growing in different applications related to different day-to-day working of the people is giving great challenge for the researchers in security management. There are different types of security requirements for the data which can protect the data from being accessed unethically. There requires continuous research in the area of security for refining the techniques further which can help in mitigating different types of security threats. Different threats are dynamic always keep change itself requires great effort. The current study will identify different types of gaps in the existing researches in comparison to k-anonymity so that some innovation can be brought in techniques in the future, which can protect the data from malicious access.

Keywords: *Encryption, Anonymization, Space, Time, Big data, Cloud, AES.*

I. INTRODUCTION

Big data is the most usable concept of today's time when we say various emerging fields are creating a large amount of data daily. These organizations use these data items at different levels of decision-making processes. It is unprecedented that the quality of the decisions has improved over the past decade. This has created a positive environment for big data usage as more and more companies want to use the big data different features.
Lately, various machine learning algorithms have emerged that have refined the prediction system with a higher level of accuracy. These prediction algorithms are using different sizes of training and testing sets. These training sets are used for the training of the machine. Once the machine starts learning the knowledge will be applied to the testing set for generating higher accuracy of the prediction **(E. Hossain, et al. 2019)[2]**

The large data generated as row data will be stored in the cloud-based storage. These cloud-based storages have a sufficient level of security so that no one can access the data illegally. The challenges for data security are getting changed over the period. These challenges create far greater threats to data privacy. These new and emerging challenges need to be handled with innovative techniques. There are different types of techniques that are currently being used for securing the data is having greater vulnerability points that provide the data leakage chances to the attackers. The new challenges that are being created need to be having a proper address mechanism. The researches in the direction of security are the hottest topic in big data analytics **(L. Xu., et al. 2014)[7]**

1.2 General introduction

As data is growing in many folds, its advantages are also growing with the emergence of various new techniques that are capable of extracting its hidden factors. Various techniques identify the direct and indirect relationship between the attributes of the data. These attributes are having greater challenges for identifying the interrelationship and also their impact on the business growth. These attributes are having higher level privacy requirements **(M. Iqbal, et al.2018)[8]**. The leakage of these data items belong to the private domain will create a major security concern for the organization. These challenges need to be addressed with suitable security means. It will enhance the trust factor for the users into the technology. Different techniques are required for making sure that the data is in the secure zone. These security threats keep on changing so there requires regular review of the data so that new threats can be identified and suitable security measures can be taken place for protecting the data from malicious access **(Bill Morrw,2012)[1]**.

1.3 Different types of threats

There are various types of threats that can destroy the privacy of the data. These security threats are having higher-level challenges that require to be mitigated with appropriate selection of the tools **(L.A. Maghrabi, 2014)[6].**

- Phishing: It is the most occurring attack where the user will be given the malicious URL which may be a replicated page of some original interface. This replicated page can lead to major security threats for the privacy of the user data **(Tom N. et al. 2007)[16]**.
- Eavesdropping: It is the security threat for accessing the data when the data will be transferred from source to destination. The attack stays in between the source and destination. The attacker generates a replica of data, resulting in a security breach **(Madhukar Anand, et al. 2005) [9].**
- DOS and DDOS: there are DOS and DDOS are the most prevalent type of attacks where any malicious user will be stopped from accessing the services of the legitimate users. In place of the legitimate user, the malicious user will be accessing the data with a bad intention **(Xiaowei Yang, et al. 2005)[17]**

1.4 Techniques for mitigating threats.

Various techniques are used to ensure the data privacy.

- Encryption: A security where data will be encoded with some public or private key. Those users who have appropriate decryption keys can read the data. The malicious users who do not have the legitimate key will not be allowed to access the data **(N. Agnihotri, et al. 2020)[10]**

- Anonymization: anonymization is the technique where data will be subdivided into multiple categories. The most critical data will be anonymized with appropriate identity hiding. This approach is currently being used for various types of financial-related data **(R. J. Bayardo, et al. 2005)[14]**

II. LITERATURE SURVEY

Osama M Ben Omran et al. (2014)[11]: Cloud is the cost-effective solution for different types of services for the clients. These services are in the category of IaaS, SaaS, PaaS. The clients belonging to different fields access these services available on the cloud. The major issue is for such architecture is for the security of the data. The cloud service provider is providing the secure solution by a technique of subdivision of the large data into smaller segments. The priority for each segment is allocated by its importance to the client. The scheme is followed for allocating the code for each segment.

Hadeal Abdulaziz et al. (2017) [3]: The new addition to big data analytics in the health sector where a large number of patient data will be stored in the central storage. Various big data processing machines will process the big data entity for identifying the suitable conclusive facts which are useful for the better diagnosis of the patients' disease. The major challenge for such an application is for keeping the data secure so that the patient data cannot be accessed illegally by any of the users without permission. Further, the paper has proposed a FOG-based solution to ensure, patient data. Pair-based encryption is used which is highly suitable for keeping the data in the encoded format. The malicious user if somehow access the data will not be having a decryption key so data would not get retrieved in a usable format.

X. Jin, B. Cui et al. (2018)[18]: there is a large growth in the data generating and processing applications which have opened new fields which are making people life easier and more comfortable. The quality of the decision-making is becoming more acceptable for the organization. The growth in the data has also increased the level of threats to the system. These threats will be a great challenge for the data managers who manages the data. A further proposed solution is providing security for the data based on an adaptive mechanism. The security expert will first check the network traffic, alert incidents, and external threat intelligence, this provides them with the idea to mitigate data breach risks by ensuring a certain level of security.

Preet Chandan Kaur et al. (2016)[12]: there are great security threats for the data being used over the networks. These threats will create a question of privacy of the private data belonging to the clients. The data requires security from both internal and external of the system. The internal segment is related to the co-workers and the external is related to the outside world. Thereafter, an anonymization technique for securing the data from malicious access has been proposed. The anonymization-based technique such as generalization, packetization, slicing, has been applied to avoid retrieval of data from the database.

Kajal Rani et al. (2017)[5]: The cloud-based PaaS and IaaS services are the most usable platform for different types of clients. These services are about having a general platform for keeping their own data processing and storage abilities. These clouds are public clouds having various clients access the services on requirements. These public platforms are having security as a great challenge for a service provider. The types of threats are changing in nature. The level of the threats to the system is highly at the risky level. An Encryption, Compression, and Splitting based multi-layer architecture for data security has been proposed. The author proposed technique is having three layers of works, in the first layer there is encryption of the data, the second layer includes the compression of the data, the third layer is regarding providing splitting the compressed data. The data in multiple segments will be stored at different locations so that collective data access by the malicious users will be stopped.

J. Stephy et al. (2018)[4]: the data is having its importance for the client. The level of importance will decide the security requirements for the data. The data which is requiring a higher level of security will be bonded into the different types of security techniques. The most usable technique from ancient times is Steganography. This technique will hide the data in some images. The image from the outside looks to be normal but the data is hidden inside the image. The real receiver knows that data is located into the image so will extract the data from the image. The data will be protected from malicious users.

Samuel S. Wu et al.(2017)[15]: the data in the social media platform and the health sector is suffering from various types of threats. These threats are making data to be more vulnerable to losing its security. Various techniques are used for providing security for the data. These security solutions maybe sometimes less in front of the level of the threats. The author of this paper has proposed collusion resistant multi-matrix masking technique for preserving the privacy of the data. There is a quite high number of threats that can be mitigated with the proposed technique. The technique is based on inferring statistical parameters of the data that remain the integral properties of the data. Various statistical techniques provide suitable solutions for data security and protection.

AES is the symmetric key encryption, works on the plain text of 128 bits. For its functioning of converting data in cyphertext, it adds around key after every rotation. having the ability to choose the variable size block and variable size encryption and decryption key, which is selected on a requirement basis. T **(P. Dhawan, 2013)[13] (N. Agnihotri, et al. 2020) [10]**

III. RESEARCH GAP

- Various researchers are working for the security mechanism are applying the security principles which are weaker for the current context.

- Various researchers identify encryption techniques as the suitable mean for providing security for the data. The time and space requirements for the data will be increased substantially.
- Different researchers talk about the anonymization-based security for the data, which will lead to creating a lower level of protection.
- There requires an improvement in data privacy and security using improved anonymization. This will improve the performance in terms of protection and performance.

IV. RESULTS FOR THE COMPARISON OF ANONYMIZATION AND ENCRYPTION

4.1 Dataset

The standardized dataset for checking the relative strength of the data security techniques is taken. The dataset is having higher strength such that comparative analysis based on different parameters is taken. The dataset for the k-anonymity is taken from URL ***https://www.kaggle.com/demodatauk/ full-banking-transaction-log-sample.*** **The dataset is having attributes as transactionID, ccountno,depositamt etc.**

4.2 Motivation

There are various security mitigation procedures currently being used for the preservation of the security of the data. There are plenty of techniques available, but adoption of the technique depends upon the time and the memory requirements for the specific technique while performing security for the data. Paper has primarily focused on checking the time and the memory space requirements for the specific techniques i.e. k-anonymity and AES.

4.3 Experimental setup

The whole process of testing different techniques for the relative strength is done with a standardized environment, which has been completely described in Table 1.

Parameter	Value
Tool used	MATLAB 2015
Dataset	Standardized
Processor	I5
Memory	4GB RAM

Table. 1 Environmental Setup

4.4 Parameters

Two parameters are used for measuring the performance of the two data security techniques.

- Memory: It is the total memory required for storing the process while execution. The higher is the memory requirement, the lower will be the compatibility for the client machine.
- Time: the time is the total processing time required for data security. Higher is the time lower will be the compatibility of the technique.

4.5 Comparison of memory required for security using anonymization and the AES encryption

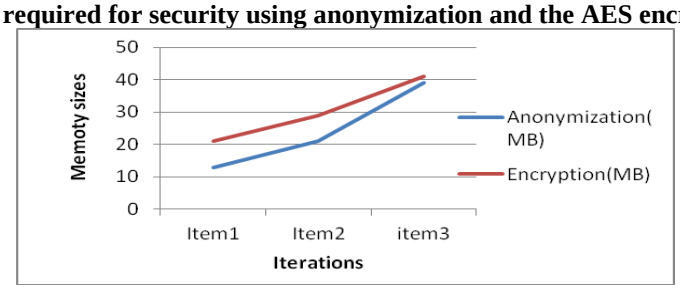


Fig.1 Memory comparison

Fig. 1 shows the memory requirements for the different dataset items for two different techniques. The memory requirements for Anonymization is comparatively lower compared to the encryption

4.6 Comparison of Time required for security using anonymization and the AES encryption

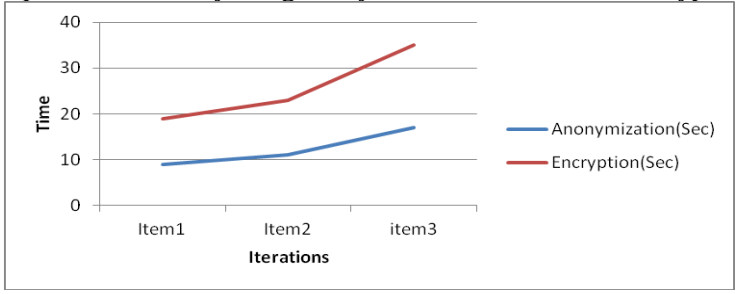


Fig. 2 Time comparison

Fig. 2 shows the comparison of the two techniques based on time. The anonymization is having better time parameter results compared to the time.

4.7 Complete Comparison

Parameter	Anonymization	Encryption
Memory	24.33	30.33
Time	12.33	25.66

Table. 2 Absolute improvement comparison

Table 2 shows the absolute improvement in the time and memory for the different datasets on two different parameters like memory and time.

V. CONCLUSION

Big data is growing in its applications in different fields. There is a various critical application which requires a higher level of security because the data is related to the highly secret zone. These applications can only grow in this service domain if data will be provided with enough level of security and trust-building. Various security means are used in the current environment. These security means are facing challenges from different types of attacks. These security means need to be having improvements for making them resilient for the current environmental protection challenges. There is wide scope in this line that can help in the protection of the data with greater trust. The encryption and anonymization-based techniques can be improved further for enhancing their performance in terms of time and space requirements.

VI. FUTURE WORK

There are high-class techniques that are currently being used for protecting the data from any malicious access. Further research can be done to reduce the response time of any such technique for security mitigation. It will help in opening more avenues for real-time use in subsequent fields.

REFERENCES

[1] Bill Morrow, BYOD security challenges: control and protect your most sensitive data, Network Security, Volume 2012, Issue 12, 2012, Pages 5-8, ISSN 1353-4858,

[2] E. Hossain, I. Khan, F. Un-Noor, S. S. Sikander and M. S. H. Sunny, "Application of Big Data and Machine Learning in Smart Grid, and Associated Security Concerns: A Review," in IEEE Access, vol. 7, pp. 13960-13988, 2019, doi: 10.1109/ACCESS.2019.2894819.

[3] H. A. Al Hamid, S. M. M. Rahman, M. S. Hossain, A. Almogren and A. Alamri, "A Security Model for Preserving the Privacy of Medical Big Data in a Healthcare Cloud Using a Fog Computing Facility With Pairing-Based Cryptography," in IEEE Access, vol. 5, pp. 22313-22328, 2017, doi: 10.1109/ACCESS.2017.2757844.

[4] J. Stephy and V. Subramaniyaswamy, "Analysis of digital image data hiding techniques," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, Palladam, India, 2018, pp. 140-144, doi: 10.1109/I-SMAC.2018.8653794.

[5] K. Rani and R. K. Sagar, "Enhanced data storage security in cloud environment using encryption, compression and splitting technique," 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), Noida, 2017, pp. 1-5, doi: 10.1109/TEL-NET.2017.8343557.

[6] L.A. Maghrabi, "The threats of data security over the Cloud as perceived by experts and university students," 2014 World Symposium on Computer Applications & Research (WSCAR), 2014, pp. 1-6, doi: 10.1109/WSCAR.2014.6916842.

[7] L. Xu, C. Jiang, J. Wang, J. Yuan and Y. Ren, "Information Security in Big Data: Privacy and Data Mining," in IEEE Access, vol. 2, pp. 1149-1176, 2014, doi: 10.1109/ACCESS.2014.2362522.

[8] M. Iqbal, S. H. A. Kazmi, A. Manzoor, A. R. Soomrani, S. H. Butt and K. A. Shaikh, "A study of big data for business growth in SMEs: Opportunities & challenges," 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), 2018, pp. 1-7, doi: 10.1109/ICOMET.2018.8346368.

[9] Madhukar Anand, Zachary Ives, and Insup Lee. 2005. Quantifying eavesdropping vulnerability in sensor networks. In Proceedings of the 2nd international workshop on Data management for sensor networks (DMSN '05). Association for Computing Machinery, New York, NY, USA, 3–9. DOI: https://doi.org/10.1145/1080885.1080887

[10] N. Agnihotri and A. K. Sharma, "Comparative Analysis of Different Symmetric Encryption Techniques Based on Computation Time," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2020, pp. 6-9, doi: 10.1109/PDGC50313.2020.9315848.

[11] O. M. Ben Omran and B. Panda, "A new technique to partition and manage data security in cloud databases," The 9th International Conference for Internet Technology and Secured Transactions (ICITST-2014), London, 2014, pp. 191-196, doi: 10.1109/ICITST.2014.7038803.

[12] P. C. Kaur, T. Ghorpade and V. Mane, "Analysis of data security by using anonymization techniques," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 287-293, doi: 10.1109/CONFLUENCE.2016.7508130.

[13] P. Dhawan, "Data Security Model for Cloud Computing," International Journal of Research in Science And Technology, Vol. 2(2), 2013

[14] R. J. Bayardo and Rakesh Agrawal, "Data privacy through optimal k-anonymization," 21st International Conference on Data Engineering (ICDE'05), 2005, pp. 217-228, doi: 10.1109/ICDE.2005.42.

[15] S. S. Wu, S. Chen, A. Bhattacharjee and Y. He, "Collusion Resistant Multi-Matrix Masking for Privacy-Preserving Data Collection," 2017 ieee 3rd international conference on big data security on cloud (bigdata security), ieee international conference on high performance and smart computing (hpsc), and ieee international conference on intelligent data and security (ids), Beijing, 2017, pp. 1-7, doi: 10.1109/BigDataSecurity.2017.10

[16] Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. Commun. ACM 50, 10 (October 2007), 94–100. DOI: https://doi.org/10.1145/1290958.1290968

[17] Xiaowei Yang, David Wetherall, and Thomas Anderson. 2005. A DoS-limiting network architecture. SIGCOMM Comput. Commun. Rev. 35, 4 (October 2005), 241–252. DOI: https://doi.org/10.1145/1090191.1080120

[18] X. Jin, B. Cui, J. Yang and Z. Cheng, "An Adaptive Analysis Framework for Correlating Cyber-Security-Related Data," 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA), Krakow, 2018, pp. 915-919, doi: 10.1109/AINA.2018.00134.

COMPARATIVE ANALYSIS OF THE FEATURE EXTRACTION TECHNIQUES USED IN DETECTING MELANOMA CANCER

Ramandeep Kaur[1]
Research Scholar Sri Guru Granth Sahib World University,Fatehgarh Sahib ramancheemachahal@gmail.com

Dr. Navdeep Kaur[2]
Professor
Sri Guru Granth Sahib World University,Fatehgarh Sahib
drnavdeep.iitr@gmail.com

ABSTRACT: Skin cancer is a potentially fatal disease is caused by abnormal melanocytic cell proliferation. Skin cancer is sometimes known as melanoma because of its malignancy. Melanoma develops on the skin as a result of sun exposure and hereditary factors. As a result, melanoma lesions appear in black or brown color. Skin cancer early identification has the potential to reduce death and morbidity. The strategies utilized in detecting melanoma skin cancer are presented in this experiment. The goal of feature extraction is to extract features from a lesion image in order to classify the melanoma.

KEYWORDS: *Melanoma, Skin Cancer, Feature Extraction, Deep Learning.*

1. INTRODUCTION

Skin cancer has recently been identified as one among the most dangerous types of cancer seen in humans.Basal-Cell Carcinoma, Squamous Cell Carcinoma, and Melanoma are the most common kinds of skin cancer. Melanoma is the utmost lethal kind of skin cancer (Majumder and Ullah, 2018). Only 4% of all skin malignancies are caused by UV rays yet account for 75% of deaths from cancer of the skin. Melanoma is caused by the existence of Melanocytes anywhere on the body.This malignancy is primarily caused by excessive sun exposure to the UV rays. It is critical to discover melanoma in its initial stages,if the cancer is not detected and treated at an early stage, it aggressively spreads to adjacent sections of the body when it becomes difficult to treat and eventually causes death (Majumder and Ullah, 2018).

There are four major sections in this study. Section 2 examines the efficacy of feature extraction strategies for the detection of skin cancer (SC). It includes a table that compares the accuracy, sensitivity, and specificity values obtained using various feature extraction algorithms. Section 3 presents an in-depth assessment of SC detection strategies based on the evaluation of selected research publications. Section 4 provides a summary of the entire study and a brief conclusion.

2. FEATURE EXTRACTION TECHNIQUES

The skin cancer detection framework is made up of following innovative algorithms:

* Pre-processing * Image Segmentation * Feature extraction

Feature extraction is utilized to draw out traits that accurately characterize a melanoma lesion, alike those visually observed by dermatologists. Because of its efficiency and simplicity of implementation, many computerized melanoma detection systems rely heavily on the traditional clinical algorithm of the ABCD-rule of dermoscopy for feature extraction. Its success stems from the presence of important melanoma lesion characteristics like asymmetry, border irregularity, color, and divergent structures, from which quantitative values could be obtained.

Internal features and exterior features are the two types of features. Internal features like globules, pigmented networks, non-uniform streaks, blue white veil, malignant area, and so on can be extracted from dermoscopic images. External aspects comprise information obtained from the patient, such as irritation on the skin, age, family history, etc. Some attributes can be extracted from a dermatoscopic image. For example, pixel contrast or intensity, correlation, energy, homogeneity, mean, skewness, kurtosis, entropy, distribution, standard deviation, and so on. Many approaches are employed in the diagnosis process, including the ABCD rule, the Menzies technique, wavelet transformation, the seven-point checklist technique, and pattern analysis.

2.1 Wavelet Transform

Wavelets are a Fourier analysis extension. Fourier analysis mathematics extends back in time to the nineteenth century. However, it wasn't until the mid-20th century that it became widespread. This technique is broadly employed in signal analysis, and it has had an impact on almost every scientific subject.

Wavelets are a mathematical methodology for the hierarchy of the frequency domain functions while the spatial domain is being maintained. This function can be used to differentiate items in noisy images from the background and other objects based in different frequency bands on their frequency response. There are numerous contemporary images that provide various information about the patient, but their use is somewhat restricted for a variety of reasons. Discrete data is frequently encountered in medical image processing applications. In order to solve target segmentation difficulties in pictures, wavelet transforms are applied.

2.2 Local Binary Pattern (LBP)

This approach is used to retrieve features, which is the most crucial phase in face recognition. It efficiently depicts the contour and texture of an image. LBP is an optimum texture operator that determines binary outcomes by thresholding each pixel neighborhood and marking picture pixels. Setting a center-pixel threshold allows the LBP to work between each neighboring pixel. If the value of the neighboring pixel exceeds or equaled center pixel, it is represented by a 1, otherwise by a 0.

2.3 Eigenvector-based Feature Extraction

Feature extraction is a method of reducing dimensionality that uses some functional mapping to take a subset of the original collection and extract a collection of new features while retaining as much information as feasible in the data. One of the most extensively used feature extraction techniques is conventional Principal Component Analysis (PCA), and it is focused on identifying the axes with the most variability in the data. The LDA or NWFE method is used to compute the eigenvectors initially. The eigenvectors serve as a matrix for spatial transformation. Second, via a matrix transformation, the primary data is projected into a new environment, resulting in a projection value that is suitable for classification (Wang et al. 2019).

2.3.1 Principle Component Analysis (PCA)

In data science, a typical feature extraction approach is Principle Component Analysis (PCA). PCA comprised mainly of four major components: feature covariance, Eigen decomposition, principal component transformation, and component selection based on explained variance. PCA selects the vectors of a covariance matrix with the highest values of their own, and then uses them to project the data onto a new, equal to or smaller dimensional subspace. In practice, PCA is a method for transforming an *n*-feature matrix into a new dataset with less than *n* features. In other words, by building a new, smaller collection of variables, it minimizes the number of features that capture most information available in original characteristics.

2.3.2 Linear Discriminant Analysis (LDA)

LDA is a dimensionality reduction technique that is frequently employed in the pre-processing stage of pattern-classification and machine learning applications. The goal is to reduce over-fitting and processing costs by projecting a dataset onto a lower-dimensional space with acceptable class separation.

2.4 Texture Features

The texture features are thought to be crucial in the identification of melanoma. The texture of a picture is defined by the spatial distribution of pixels in its vicinity. The spatial dependency of grey levels is represented by the GLCM, a two-dimensional matrix used for global texture analysis of an image. The GLCM matrix determines the texture of an image by indicating how frequently pairs of pixels with specified values appear in an image. Following that, the statistical measure is taken from the GLCM matrix. Textural features show that grey-tone variances in a certain area are spatially distributed. The pixels are associated in the photos and spatial values are obtained from the redundancy of the nearby pixels (Kavitha et al. 2017).

2.4.1 Gray Level Co-occurrence Matrix (GLCM)

GLCM technique is used for obtaining statistical texture information of second order. The method has been employed in a variety of applications. The relationships between three or more pixels are taken into account in third and higher order textures. Although these are technically viable, they are rarely employed due to calculation time and interpretation issues (Mohanaiah et al. 2013). GLCM functions characterize the texture of an image by determining how frequent pairs of pixels occur in an image, generate a GLCM, and extract statistical measures from this matrix in a certain spatial connection.

2.5 Histogram of Oriented Gradients (HOG)

HOG is a functional descriptor used in computer vision and image treatment to recognize objects. The methodology counts the number of times a gradient orientation appears in a limited region of an image. The HOG description focuses on an object's structure or shape. For object detection, the HOG characteristics are commonly utilized. HOG divides the picture into small squared cells, calculates a histogram in each cell of directional gradients, normalizes the result in a block wise motif and returns a description for each cell.

2.6 Gabor Features

In image processing, because of its remarkable properties, Gabor filters are commonly utilized. Optimal spatial frequency location joint and the ability to duplicate the receptive fields in visual cortex of simple cells (Wang et al. 2005).

2.7 ABCD Rule

The ABCD (Fidalgo Barata et al. 2020) technique distinguishes benign from nasty melanoma. *D* is the diameter used by some studies; if the diameter is higher than 6 millimeters and/or expands in the following month, it is malignant melanoma. In the ABCD technique (Thanh et al. 2019), each assigns a score and multiplies it by a factor. TDV is the abbreviation for 'Total Dermoscopic Value'. One simple method to memories common melanoma traits is to think alphabetically - the

ABCDEs of melanoma. ABCDE is an abbreviation for asymmetry, border, color, dimension, and evolving. These are the skin damage characteristics that clinicians look for when diagnosing and classifying melanomas (Mehta and Shah, 2016).

- *Asymmetry:* Melanoma frequently has poorly defined or uneven borders, but non-cancerous moles often have smooth, well-defined edges.
- *Border:* Melanoma frequently has poorly defined or uneven borders, but non-cancerous moles often have smooth, well-defined edges.
- *Color:* Melanoma lesions are frequently multicolored or shaded. Benign moles are typically one color.
- *Diameter:* Melanoma tumors are often greater than 6mm in diameter, or about the size of a common pencil.
- *Evolution:* Melanoma frequently changes its appearance, such as size, shape, or color. Melanoma tends to change over time unlike most benign lesions.

2.8 Menzies Method

To diagnosis any lesion, it must have neither negative nor one or more of the nine positive characteristics to be malignant or benign. Figure 1 shows positive and negative features. This is why many scientists use this method to determine malignant melanoma or benign melanoma for the most susceptibility.

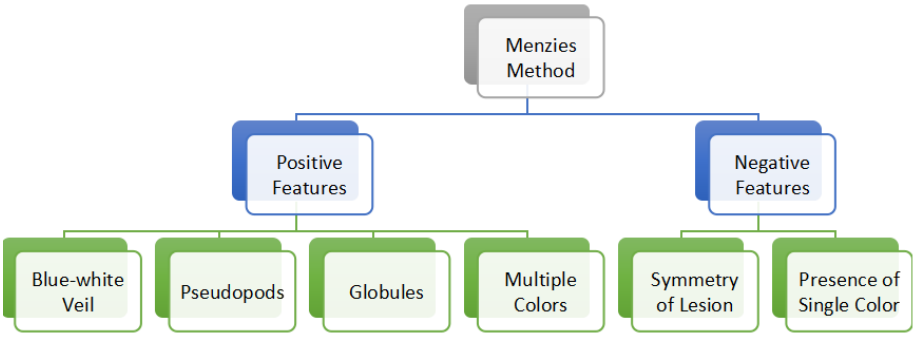


Figure 1. Types of Menzies Method

- *Symmetry of lesion:* Pattern symmetry need all the axis across a lesion center and no form symmetry.
- *Presence of single color:* Colors like blue, brown, black, gray, and tan are scored. White isn't counted as a color.
- *Blue-white veil:* Irregular regions with convergent blue pigmentation with an overlaying white "ground-glass" coating.
- *Multiple Brown dots:* Several dark brown spots in the skin lesion region.
- *Radial Streaming:* It is the radially oriented linear structures in the growth direction of pigment at the margin of a lesion.
- *Psuedopods:* It manifests as finger-like extensions of dark pigment along the periphery of the lesion.
- *Scar-like depigmentation:* White areas with conspicuous, uneven extension.
- *Globules:* Black dots discovered in or near the region of interest.
- *Multiple colors:* Colors found in the region of interest incorporate black, grey, blue, dark brown, tan, and red.
- *Multiple blue-gray dots:* Patterned focus of several blue or grey dots are frequently narrated as "pepper-like."
- *Broadened network:* A network made up of thicker, more unequal cords.

2.9 Pattern Analysis

These strategies seek out certain patterns, which can be global or local in nature. Global patterns include reticular, globe-shaped, homogeneous, starburst, parallel multi component, and unspecific patterns. Pigment network, uneven streaks, globules, or black spots, insufficient pigmentation, blue-white veil, regression structures, and vascular structures are examples of local patterns. This process is build on a qualitative evaluation of certain dermoscopic criteria. (Mehta and Shah, 2016).

2.10 Seven Point Checklist Scoring Method

This approach only specifies seven conventional dermoscopic criteria. A scale of 1 to 7 is provided. The method of scaling is based on the existence of major and minor criteria in the lesion. The presence of important criteria increases the score by two points, while the presence of minor criteria increases the score by one point. If the score is more than or equal to three, the tumor is categorized as malignant melanoma. When the performance of the ABCD rule technique and the seven-point checklist technique is compared, the seven-point checklist approach permits unskilled observers to acquire a better diagnostic accuracy value. This approach pertains to the chromatic properties, form, and texture of the lesion (Mehta and Shah, 2016).

Table 1. Seven-point Checklist Method

Criteria	Score
Major Criteria	
A typical-pigmented network	2
Blue-white veil	2
A typical vascular pattern	2
Minor Criteria	
Asymmetrical streaks	1
Irregular pigmentation	1
Uneven globules	1
Regression structure	1

3. LITERATURE SURVEY

Many researchers have worked on feature extraction strategies for the detection of skin cancer in the past.

Kumar *et al.* (2020) developed an improved technique for early detection of three types of skin cancer. A skin lesion photograph is being regarded as an input, which the system would use the proposed method to classify as malignant or non-cancerous. To distinguish homogeneous image regions, Fuzzy C-means is used for clustering images. Preprocessing is performed using various filters to improve image qualities, while other aspects are evaluated by combining RGB color-space, Local Binary Pattern and GLCM approaches. Furthermore, an artificial neural network is taught for classification using the differential evolution (DE) methodology. In addition, for classification, the differential evolution algorithm is used to train an artificial neural network (ANN). The originality of the work demonstrates that DE-ANN is higher than other typical detection accuracy classifiers, as stated in the section outcomes. The results demonstrate that 97.4% of skin cancer is detected efficiently in the proposed methodology. When compared to other traditional methodologies in the same domain, the results are very accurate.

Hakeem and Hassoun (2020) proposed an image-processing-based system for detecting skin cancer tissue, in which images are pre-processed (resize, median filter), and using Gabor features, the most important information is obtained from vector images. For cancer detection, the simulations results are efficient with an accuracy of 94.117%, based on the ANN approach.

Mohan Kumar et al. (2019), who used entropy characteristics based on the Stationary Wavelet and a Random Forest classification, produced an efficient skin-image classification approach. SWT decomposes the supplied skin pictures. The entropy of broken skin images is used to derive skin image features, which are then categorized using an RF classifier. The accuracy, sensitivity, and specificity of the system are all measured. The results show that at the third level of SWT decomposition using entropy characteristics and an RF classifier, the classification accuracy is 91.5 percent, and the sensitivity and specificity are 90 percent and 93 percent, respectively.

Amin et al. (2019) described a method for segmenting and classifying skin lesions. Deep features were collected and combined to create a single cancer classification features vector. This work produced improved results because of the following strategies: first, the fusing of divergent datasets; second, deep feature extraction from two pre-trained transfer learning models, Alex net and VGG 16; and third, serial fusion and optimization using PCA. When compared to existing methodologies, the proposed methodology obtained 0.9952 SE, 0.9841 SP, 0.9859 PPV, 0.0158 FNR, 0.0047 FPR, and 0.9900 ACC.

Lattoofi *et al.* (2019) devised and implemented the ABCD method using TDS as the classifier on skin lesion images from the PH2 dataset. In this study, authors rely on a pre-processing phase to remove hair. It is constructed around a morphological filter. In order to extract features to diagnose melanoma, the ABCD algorithm was utilized for the image. These characteristics include form structural properties and uneven color distribution. This approach produced accuracy, specificity, and sensitivity in the 93.2 percent, 92.59 percent, and 90.15 percent ranges, respectively. The experimental results indicate that the methodology used enhances the early diagnosis of skin lesions with high accuracy, resulting in less false positive prediction.

Alzahrani *et al.* (2019) proposed a pattern analysis method for detecting melanoma that incorporates a seven-point checklist and use a convolutional neural network to automatically extract lesion features. The proposed models were realized by combining automatic lesion feature extraction obtained by multi-input CNN using standardized and non-standardized pictures (dermoscopy) (clinical). As demonstrated, the proposed methodology performs well in terms of accuracy, sensitivity, and specificity.

Kavitha *et al.* (2017) reported global and local texture feature extraction utilizing several GLCM. A texture feature descriptor called Speeded up Robust Features is used to extract an image's local texture properties (SURF). The categorization results determine the performance of feature extraction. SVM and the KNN classifier are used in the classification methodology. Several metrics are used to assess performance, including sensitivity, specificity, accuracy, precision, and F1 score. The experimental results show that the local texture feature extracted using SURF outperforms global feature extraction (GLCM) and other descriptors such as Scale Invariant Feature Transform (SIFT). When used in conjunction with the SVM - RBF classifier, the SURF local feature descriptor enhances classification accuracy.

Bakheet (2017) created an efficient framework for a CAD (Computer-Aided Diagnosis) system for melanoma skin cancer that is based on an optimized collection of HOG based skin lesion descriptors. Research has demonstrated that the proposed framework is a strong supporter of advanced alternatives with great sensitivity, specificity and precision (98.21%, 96.43% and 97.32%, respectively) utilising the approach described for a large public dermoscopic picture dataset.

A powerful method to detect, identify and classify skin lesions by PCA and SVM was presented by (Alquran *et al.* 2017). The proposed technique, which used SVM based on PCA characteristics, was successful in classifying the retrieved lesion ROI. In this study, the following steps are taken: Database collection for dermoscopy picture, preprocessing, threshold segmentation, GLCM, Asymmetric, border, color, diameter (ABCD) statistical feature extraction, feature selection using Primary Component Analysis (PCA), Total dermoscopic calculation score, and SVM-type classification.

The method for identifying photos with LBP of melanoma and non-melanoma of skin diseases was reported by (Adjed *et al.* 2016). The LBP gathers information on local texture from pictures of skin cancer and then produces statistics that distinguish between melanoma and non-melanoma tissues. The feature matrix is classified into two classes using SVM (malignant and benign). The approach achieves 76.1 percent classification accuracy, 75.6 percent sensitivity, and 76.7 percent specificity.

Table 2. Accuracy, Sensitivity, Specificity Table for Feature Extraction Techniques for Detecting Skin Cancer

S.No.	Reference	Technique	Classifier	Accuracy	Sensitivity	Specificity
1.	Mohan Kumar et al. (2019)	Stationary Wavelet Transform	Random Forest	91.5%	90%	93%
2.	Adjed et al. (2016)	Local Binary Pattern	SVM	76.1%	75.6%	76.7%
3.	Kavitha et al. (2017)	GLCM	SVM, KNN	87.3%	86.2%	88.4%
4.	Alzahrani et al. (2019)	Seven-point Checklist Method	CNN	0.6430	0.5537	0.8926
5.	Lattoofi et al. (2019)	ABCD Rule	-	93.2%	90.15%	92.59%
6.	Bakheet et al. (2017)	HOG	SVM	97.32%	98.21%	96.43%
7.	Hakeem & Hassoun (2020)	Gabor Filter	ANN	94.117 %	-	-

4. SUMMARY

One of the most common malignancies in human beings is skin cancer. It is divided into two types: Non-melanoma and Melanoma skin cancer. Extensive research solutions have been developed by building feature extraction algorithms in order to diagnose skin cancer quickly and at the earliest stage, as well as to overcome some of the aforementioned concerns. This paper discusses feature extraction strategies for detecting skin cancer. The metrics in Table 2 show the accuracy, sensitivity, and specificity of feature extraction approaches for identifying melanoma skin cancer.

REFERENCES

1. Adjed, F., Faye, I., Ababsa, F., & Gardezi, J. (2016). Classification of skin cancer images using local binary pattern and SVM classifier. *4th International Conference on Fundamental and Applied Sciences (Icfas2016)*.
2. Alquran, H., Qasmieh, I. A., Alqudah, A. M., Alhammouri, S., Alawneh, E., Abughazaleh, A., & Hasayen, F. (2017). The melanoma skin cancer detection and classification using support vector machine. *2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*.
3. Alzahrani, S., Al-Nuaimy, W., & Al-Bander, B. (2019). Seven-Point Checklist with Convolutional Neural Networks for Melanoma Diagnosis. *2019 8th European Workshop on Visual Information Processing (EUVIP)*.
4. Amin, J., Sharif, A., Gul, N., Anjum, M. A., Nisar, M. W., Azam, F., & Bukhari, S. A. C. (2019). Integrated Design of Deep Features Fusion for Localization and Classification of Skin Cancer. *Pattern Recognition Letters*.
5. Bakheet, S. (2017). An SVM Framework for Malignant Melanoma Detection Based on Optimized HOG Features. *Computation*, 5(4), 4.
6. Fidalgo Barata, A. C., Celebi, E. M., & Marques, J. (2018). A Survey of Feature Extraction in Dermoscopy Image Analysis of Skin Cancer. *IEEE Journal of Biomedical and Health Informatics.*
7. Hakeem, S. I. & Hassoun, Z. A. (2020). Skin Cancer Detection based on Terahertz Images by using Gabor filter and Artificial Neural network. *IOP Conference Series: Materials Science and Engineering*, 928.
8. Hosny, K. M., Kassem, M. A., & Foaud, M. M. (2018). Skin Cancer Classification using Deep Learning and Transfer Learning. *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*.
9. Kavitha, JC., Suruliandi, A., & Nagarajan, D. (2017). Melanoma Detection in Dermoscopic Images using Global and Local Feature Extraction. *International Journal of Multimedia and Ubiquitous Engineering,* 12(5), 19-28.

10. Kumar, M., Alshehri, M., AlGhamdi, R., Sharma, P., & Deep, V. (2020). A DE-ANN Inspired Skin Cancer Detection Approach Using Fuzzy C-Means Clustering. *Mobile Networks and Applications*.
11. Lattoofi, N. F., Al-sharuee Israa F., Kamil, M. Y., Obaid, A. H., Mahidi, A. A., Omar, A. A., & Saleh, A. k. (2019). Melanoma Skin Cancer Detection Based on ABCD Rule. *2019 First International Conference of Computer and Applied Sciences (CAS)*.
12. Mohan Kumar, S., Ram Kumar, J., & Gopalkrishnan, K. (2019). Skin Cancer Diagnostic using Machine Learning Techniques - Stationary Wavelet Transform and Random Forest Classifier. *International Journal of Innovative Technology and Exploring Engineering (IJITEE).* 9(2), 4705-4708.
13. Majumder, S., & Ullah, M. A. (2018). Feature Extraction from Dermoscopy Images for an Effective Diagnosis of Melanoma Skin Cancer. *2018 10th International Conference on Electrical and Computer Engineering (ICECE)*.
14. Mehta, P., & Shah, B. (2016). Review on Techniques and Steps of Computer Aided Skin Cancer Diagnosis. *Procedia Computer Science*, 89, 309-316.
15. Mohanaiah, P., Sathyanarayana, P., & Gurukumar, L. (2013). Image Texture Feature Extraction Using GLCM Approach. *International Journal of Scientific and Research Publications*, 3(5), 1-5.
16. Smaoui, N., & Derbel, N. (2018). Melanoma Skin Cancer Detection based on Image Processing. *Current Medical Imaging Reviews*, 14.
17. Thanh, D. N. H., Prasath, V. B. S., Hieu, L. M., & Hien, N. N. (2019). Melanoma Skin Cancer Detection Method Based on Adaptive Principal Curvature, Colour Normalisation and Feature Extraction with the ABCD Rule. *Journal of Digital Imaging*.
18. Wang, W., Mou, X., & Liu, X. (2019). Modified eigenvector-based feature extraction for hyperspectral image classification using limited samples. *Signal, Image and Video Processing*.
19. Wang, X., Ding, X., & Liu, C. (2005). Gabor filters-based feature extraction for character recognition. *Pattern Recognition*, 38(3), 369–379.

A COMPARATIVE STUDY OF VIDEO WATERMARKING BASED ON DWT AND SVD

Ms. Anuradha Saini[#1], Dr. Sushil Bhardwaj[#2]
[1]*PhD Research Scholar, Department of Computer Applications, RIMT University, Mandi Gobindgarh, Punjab, India*
[2]*Assistant Professor, Department of Computer Applications, RIMT University, Mandi Gobindgarh, Punjab, India*
[1]anuradhasaini1984@gmail.com
[2]sushilbhardwaj2010@gmail.com

ABSTRACT:- The constant innovations in the speed of Internet access have paved a promising way for digital image and video streaming. On the hind side, maintaining the originality and authenticating the videos too, poses a challenge for the stakeholders. Counterfeit videos being circulated on the digital media is an area of concern. Intensive research is being carried out in the area of digital theft. One of the trending solution is Digital video watermarking. The watermark is some additional information that is added to the host video or image to secure and authenticate the data. This paper presents the importance of digital video watermarking and its different techniques. This paper also provides review of different techniques followed by the quantitative comparisonof output of these techniques on the basis of three different parameters i.e. PSNR, NC and MSE. In addition to this, comparison is also done on the basis of various attacksthat are performed on the videos by various techniques to test the robustness.

KEYWORDS:- Watermarking, Discrete wavelet transform(DWT), Singular Value Decomposition(SVD), PSNR, NC, MSE

I. INTRODUCTION

With the development of Internet, the problem of data authentication and data security emerged as a big issue. As the use of computers has increased a lot over the years, the circulation of information through internet is becoming faster, easier and simpler. Now a days, the illegal distribution of the patented and copyrighted data like image, audio, video etc. is very common and a topic of serious concern for advertising companies, film industries, gaming etc. The owner is always concerned for its original data and to prove the ownership and authenticity, he hides a unique or secret data with the original data, through which he can claim his ownership in case of piracy or unauthorized use of data.

An illegal copy of a digital video can be easily distributed to a worldwide audience[1], as there are so many online streaming sites are available through internet. In this modern era, a large amount of multimedia data is generated and broadcasted all around the world through various social networking websites and portals with the help of Information and Communication Technology (ICT)[2] which proves to be an indispensable and cost efficient technique for propagation of the multimedia documents. With the increase in online data dissemination, there exists a need for prevention of copyright violation, authenticity, confidentiality and ownership identity theft[3] which has been considered as the potential issues in the field of multimedia. In order to settle these issues, generally a watermark information is embedding into data file to guarantee the genuineness of the multimedia data as required in several important applications like e-health, fingerprinting, forensic, protection of social digital contents, E-Voting and driver licenses, military, remote education, media file archiving, broadcast monitoring and digital cinema[4]. Thus, there is need to develop effective watermarking methods that can offer good trade-off between the benchmark parameters for the above considered applications.

Digital watermarking is the process in which a unique image or piece of information is taken as a watermark and same is embedded permanently in the original image or in the frames of video by means of some technique. Image that has been used as watermark can be extracted from the watermarked image or video via some operations which can further be produced as the evidence of originality by the owner or for any other copyright issue.

II. WATERMARKING TECHNIQUES

Watermarking techniques as shown in Figure 1, can be grouped into two major classes depending upon the domain; spatial-domain watermarking techniques and frequency-domain watermarking techniques.

Watermarking Techniques	**Spatial Domain**	**> LSB** **> Correlation based Technique**
	Frequency Domain	**> DCT** **> DWT** **> DFT** **> SVD**

Figure 1: Different Watermarking Techniques

1. Spatial Domain Watermarks

Spatial-domain techniques embed a watermark in the frames of a given video/content by changing its pixels directly. These techniques are easier to implement thereby requiring fewer computational resources; however, they do not fare well against common digital signal processing operations such as image/video compression.Algorithms in this class have some common characteristics. One of them is, during watermark design or embedding no transforms are applied to the host signal. Another one is, spread spectrum modulation is used to drive the watermark from the message data. Last is, combining operation of watermark in the host data is done through simple operations like addition or replacements and this

operation is performed directly on the pixel values and detection is done by correlating the received signal with the expected pattern. There are some strengths of pixel domain methods:

- They are conceptually simple
- They have very low computational complexities.
- They have been mostly used for video watermarking applications where the primary concern is real-time performance.

However, they have some limitations also like the need for absolute spatial synchronization leads to high susceptibility to de-synchronization attacks, only use of spatial analysis technique the watermark optimization becomes difficult, in the absence of temporal axis the video processing and multiple frame collusion is exposed and becomes a danger to it.

2. Frequency/Transform Domain Watermarking

Frequency-domain techniques modify the coefficients of the transformed video frames according to a prescribed embedding structure. It disperses the watermark in the spatial domain of the video frame, making it very difficult to remove the embedded watermark. Thus, to achieve the digital watermarking algorithm's imperceptibility and robustness the frequency-domain watermarking techniques are more effective [5]. Two of the techniques which are extensively used by the authors to enhance the robustness, perceptibility and security of video are SVD and DWT, the working of these techniques are explained further:

3. Singular Value Decomposition (SVD)

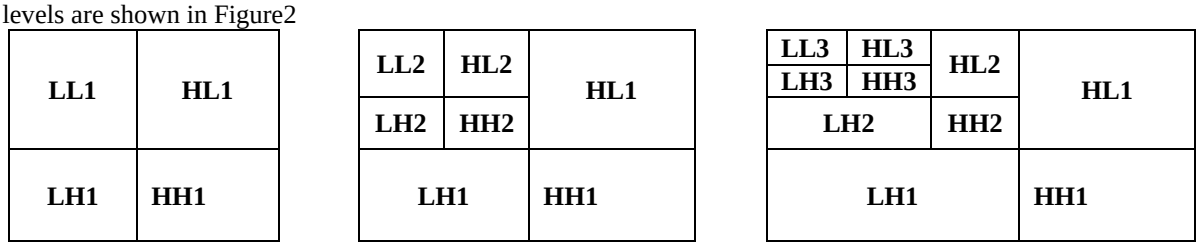
It is a mathematical technique which extracts the 3 matrices from one image matrix. Considering a image I which is a frame of a video sequence to be a mxm matrix then its SVD can be defined as:

$$I=USV^T$$

Where S is a diagonal matrix and U and V are unitary (or orthogonal) matrices. The columns of matrix V are the right singular vectors of the image I, whereas the left singular vectors of the image I are available in the columns of matrix U [5]. In SVD based watermarking technique the watermark is embedded by modifying either U and V or S matrix. Suppose I is an mxn matrix then U is an mxm unitary matrix, V is an nxn unitary matrix (V^T is transpose of V) and S is mxn diagonal matrix.

4. Discrete Wavelet Transform (DWT)

In this technique the frame or image is divided in four sub bands known as LL, LH, HL and HH, where LL is lower-resolution approximation image and HL is horizontal, LH is vertical, and HH diagonal are high frequency part of image with detail components. The process can be repeated to compute multiple "scale" wavelet decomposition[6]. In 2-level DWT the LLsub band of first level is divided in above said four sub band and we can call them LL2, HL2, LH2, and HH2 respectively. And same division is followed in 3-level DWT i.e. LL2 is divided into LL3, HL3, LH3 and HH3. All of these three levels are shown in Figure2



(a) (b) (c)

Figure2: Sub bands generated by DWT at different levels(a) DWT / 1-level DWT (b) 2-level DWT (c) 3-level DWT

Some of the authors [9][13] used SVD technique which is very effective and gives good result but most of the authors [7][8][10][11][12] have clubbed two or more techniques to improve the robustness and security. All these are discussed in the following part of the paper.

III. LITERATURE REVIEW

Nadesh R.K et al. (2019) [7], proposed a hybrid technique which is formed by combining the DWT and SVD techniques. PCA technique is used to minimize the correlation between 2 wavelet coefficients. Firstly the DWT is implemented on a video frame to divide it into four parts and then SVD is used to embed the watermark, and a secret key is used for authorization during extraction. This technique can hide information in the video and ensures that only the authorized user can extract the watermark with the help of the secret key.

Shafali Banyal et al. (2016) [8], proposed a robust and effective video watermarking algorithm for copyright protection based on Discrete Wavelets Transform (DWT) and Singular Value Decomposition (SVD). Owing to the exceptional spatio-frequency localization properties, DWT is employed to find out areas where a watermark can be embedded imperceptibly in the host video frame. Hence, the video watermarking algorithm proposed by the author is based on DWT and SVD. The proposed algorithm has two components - a fifteen step procedure to embed the watermark in the original video followed by a seven step procedure to extract the watermark. The algorithm is executed on three images from the

frame sequence of digital video. The paper concludes by comparing its peak signal ratio (64.28) to that of the existing techniques (42). The proposed algorithm claims to have achieved greater degree of compression by the use of daubechies (db10) based watermarking technique and reduced SVD.

Imen et al. (2018) [9], suggested a new and competent strategy of video watermarking based on the Singular Value Decomposition (SVD) and Multi Resolution-SVD domain. This method picked only the fast motion frames of the authentic video to insert the watermark which makes it exceptional and time-efficient from other algorithms. For incorporation of the watermark, the process of Quantization Index Modulation (QIM) is employed. In this scheme, the partial frames were to be worked upon, raising the imperceptibility of the watermark. Thus, the amalgamation of these techniques SVD, MR-SVD, QIM, results in constructing this approach safer than the other existing methods. The final outcome of this approach says that the method is imperceptible and can survive most of the categories of the attacks.

Jane et al. (2014) [10]The authors worked upon given applications of video watermarking i.e., data authentication and copyright protection. The major area of concern in this field was to maintain the security of the watermark from the disturbance occurred from the channel during transfer. The current survey of the literature gave some marked methods of insertion of a watermark into a video element. In this paper, a mixed method composes of the Discrete Wavelet Transform (DWT) is used where its coefficients are changed according to the information of watermark to be inserted into the object and Singular Value Decomposition (SVD) technique has been proposed. The DWT method has been applied to lower frequency bands followed by the inverse of the same. After the implementation, the computations of the proposed scheme are checked and it has been observed that the suggested approach gives significantly better, strong and dependable results as compared to other methods. Here, diverse attacks like filtering, scaling, Gaussian noise, compression, rotation, cropping etc. are applied on the video object and reasonable results are seen in terms of PSNR value after embedding the watermark in the lower frequency sub band (LL).

C. Sharma et al. (2018) [11], suggested a new technology which is the result of the combination of DWT, SVD and Rail fence encryption method. Rail fence encryption method is applied on the 10's complement watermark image to obtain encrypted watermark image. DWT and SVD is implemented to embed this encrypted image and inverse of same process is implemented to retrieve the watermark. This technique is robust to basic attacks but facing problem with live streaming videos.

Jantana et al. (2018) [12], advised an imperceptible video watermarking technique for data verification based on Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT). In the said article, the watermark was inserted in the Y-component of all the frames of the video and a binary image as a watermark was rooted in the mid-frequency coefficients of wavelet and cosine transforms. The result of the approach came positive as the extraction does not require the actual video and the quality of the extracted watermark came out to be good. This approach is more robust against High Efficiency Video Coding stream compression. It is hence expected that it could be efficiently implemented for the number of areas of relevance in the field of digital video watermarking like copyright protection and authentication.

Wubiao Chen et al. (2018) [13],proposed and designed a video watermarking algorithm based on SVD and secret sharing. The authors justified the use of SVD technique as it has strong stability, portrays the algebraic parameters of an image and its singular value is not affected for an image due to small interferences making the implementation of embedding and extracting the watermark process easier. The proposed algorithm consists of the Secret Sharing Scheme, Scrambling transformation, Secret sharing phase and Secret reconfiguration phase discussed briefly. It is followed by Embedding of Secret Watermark Image consisting of Generation Strategy of Watermark Image, Singular Value Decomposition of Video Image Frames and Extraction and Recovery of Watermark Image. The resulting algorithm based on SVD and Secret Sharing technique ensures guard against geometric attacks in addition to improved and greater robustness and security which is validated through the experimental results obtained.

IV. PERFORMANCE TESTING PARAMETERS

In this part of paper, we havediscussed the quantitative comparison of the output of the watermarked video on the basis of some objective measures and parameters which are mentioned below:

Mean Square Error (MSE): It calculates the mean absolute error between original and distorted frames [9]. It represents the cumulative squared error between host frame and the watermarked frame, lower the value of MSE means lower the error and is calculated as:

$$\textbf{MSE}= \sum_{i=0\,j=0}^{M-1\,N-1} \frac{1}{M*N} (OI(i,j), DI(i,j))^2$$

Where OI is Original Frame, DI is distorted Fame, M is number of Rows and N is number of Columns.

Peak Signal to Noise Ratio (PSNR): It is the most commonly used quality measure to check the performance of the technique. Imperceptibility is related to the visibility of the watermark, it should be embedded without affecting the quality of the video frame. So there is a need to measure the perceptibility in video because the amount of distortion and visibility is strongly depends on the video object. Peak Signal to Noise Ratio (PSNR) is the measuring tool which is used to get the perceptibility. The value of PSNR is expressed in decibel(dB). It is used to show the change of frame's quality before and after embedding of watermark and calculated as

$$\textbf{PSNR= 10 log 10 } (\frac{255^2}{MSE})$$

Normalized Correlation (NC): It is a method used to find out the common pattern or template matching. It is the matrix which is used to compare the similarities between two images or frames and the formula to calculate NC is given below

$$\mathbf{NC}= \frac{\sum_{(i,j)}(w'(i,j)w(i,j))}{\sum_{(i,j)}w2(i,j)}$$

Bit Error rate (BER): It is one of the commonly used method to measure the quality of the data transmission system. The said parameter is computed by making the comparison between sequence of bits transmitted and received at the destination to the count of number of errors occurred during transmission. The basic formula for measuring the BER is given which is affected by varied factors like distortion, noise etc.

$$BER = \frac{N\,Err}{Nbits}$$

To measure BER in a process, the most evident approach is to transfer the bits through the system and accordingly, compute the value of BER. For most of the processes for which we calculate BER, we just need to check if the measured value is less than the threshold value taken.

V. Comparison of Different Techniques

TABLE I

Quantitative Comparison of Results of different techniques on the basis of different parameters (PSNR: Peak Signal to Noise Ratio, NC: Normalized Correlation, MSE: Mean Square Error)

Measures	[7]	[8]	[9]	[10]	[11]	[12]	[13]
Techniques used	**DWT and SVD**	**DWT and SVD**	**SVD**	**DWT and SVD**	**DWT and SVD**	**DWT and DCT**	**SVD**
PSNR	43	62.5863	40.2	47.9272	52.234	37.9538	39.89
NC	0.0035	--	1	--	--	0.9757	0.9895
MSE	0.0035	0.0318	--	--	--	--	--

TABLE II

PSNR values of different techniques

Attacks	[7]	[10]	[11]
SALT AND PEPPER	--	12.364	34.567
GAUSSIAN	12	29.993	37.891
FRAME AVERAGING	--	--	41.453
CROPPING	11	13.046	34.324
ROTATION	43	11.427	31.123

TABLE III

NC and BER values of different techniques

	NC			BER	
Attacks	**[7]**	**[9]**	**[13]**	**[9]**	**[11]**
SALT AND PEPPER	--	0.9238	0.9895	0.0303	0.0289
GAUSSIAN	0.9986	0.9008	--	0.0404	0.026
FRAME AVERAGING	--	0.8	--	0.0909	0.024
CROPPING	0.0036	--	--		0.029
ROTATION	0.9989	--	--		0.032

From the TableI, we can find out that [8] is giving best result on the basis of PSNR value but the author has not implemented any attack on the watermarked video, so [8] has achieved better perceptibility but has not worked upon the robustness of the technique. As robustness also has same importance like perceptibility: [7], [9], [10], [11] have implemented various attacks and maintained the balance between the perceptibility and the robustness. In [13] author has implemented only one attack (Salt and Pepper), so we cannot compare its robustness through one attack only. As [11], [10] and [7] also got good PSNR value 52.234, 47.9272, 43 respectively as compared to [9] and [13] having values 40.2 and 39.89 respectively and achieved robustness also. [11], [10] and [7] are using hybrid techniques (DWT and SVD), whereas [9] and [13] have used standalone (SVD) technique. From Table I it is clear that the combination of techniques leads to better outcomes and improved results.

Table II shows the PSNR values while implementing various attacks by [7], [10] and [11]. As it is clearly shown in Table II technique [11] is giving better result as compared to [7] and [10] with all the mentioned attacks except the rotation attack (when compared with [7]).

From Table III, we can conclude that [7] is giving good result with NC values but [11] is giving lesser BER as compared to [9]. So, after comparing all the parameters of perceptibility and robustness after attacks, we can conclude that [11] is better among all compared techniques, but the unavailability of its NC and MSE values act as a hindrance in providing the accurate comparison.

VI. CHALLENGES/ISSUES

Most of the above reviewed techniques are robust to different types of Noise Attacks and Geometric Attacks. Only few technique has worked with the Video Compression Attacks and few with Frame Synchronization Attacks. So there is a scope of research work to improve the result with Video Compression Attacks and Frame Synchronization Attacks.These techniques are facing problem in real time videos. So there is also a great scope of research to work with live streaming videos. Another domain in which research can be performed is Ambiguity attacks and Collision Attacks on videos. Very less work has been done in Video Watermarking with these attacks.

VII. CONCLUSION

After study it is concluded that all the previous frequency domain techniques are helpful in data security and is a valuable repository for researchers but the combinations of different techniques are giving better results as compared to the working of individual standalone techniques. As reviewed it has been found that the SVD technique is giving better results than other frequency domain techniques. But its results improves when it is combined with other watermarking techniques in one or more phases of watermarking embedding process. This comparison elucidates that the merged and combined techniques are more robust to various attacks than the standalone techniques. Standalone techniques are very common and can be detected by the hackers easily, so a robust technique is still required which should be capable of hiding watermark in such a way that the finding of watermarked frames and extraction of watermark becomes challenging task. So to achieve this, the existing techniques can further be merged with some algorithms of finding fast motion frames or with some other data hiding or authentication techniques to make it more reliable and hybrid robust technique.

REFERENCES

[1] Mohd., A., Mohd. J.A., Lambert, A.J., and Pickering, M.R. (2013, November). A blind high definition video watermarking scheme robust to geometric and temporal synchronization attacks, IEEE

[2] Singh, A.K., Kumar, B., Dave, M., Ghrera, S.P., and Mohan, A., (2016). Digital Image Watermarking: Techniques and Emerging Applications, B. B. Gupta et al. (Eds.) Handbook of Research on Modern Cryptographic Solutions for Computer and Cyber Security, IGI Global, USA, 246-272.

[3] Lee, J., and Jung, S.H., (2001). A survey of watermarking techniques applied to multimedia, Proceedings 2001 IEEE International Symposium on Industrial Electronics (ISIE2001), 1, 272 -277.

[4] Singh, A.K., Kumar, B., Singh, S.K., Ghrera, S.P. and Mohan, A., (2016). Multiple watermarking technique for securing online social network contents using back propagation neural network, Future Generation Computer System, 86, 926–939.

[5] Mohd., A., and Pickering, M.R. (2016). An Overview of Digital Video Watermarking, IEEE Transactions on Circuits and Systems for Video Technology, 28(9).

[6] Paul, R.T., (2014, November). Video Watermarking Based on DWT-SVD Techniques, IJSR, 3(11), 1624-1629.

[7] Nadesh, R.K., Srinivasa Perumal, R., Arivuselvan, K., Aishwarya, K., (2019), A Hybrid Approach for Video Watermarking Using DWT and SVD, Innovations in Power and Advanced Computing Technologies (i-PACT)

[8] Banyal, S., Sharma, S. (2016, October). Digital Video Watermarking Using DWT and SVD Techniques, International Journal of Advanced Research in Computer and Communication Engineering, 5(10), 223-227.

[9] Nouioua, I., Amardjia, N., and Belilita, S. (2018). A Novel Blind and Robust Video Watermarking Technique in Fast Motion Frames Based on SVD and MR-SVD, Hindawi, Security and Communication Networks.

[10] Jane, O., Elbasi, E., llk, H.G.,(2014 August) Hybrid Non-Blind Watermarking Based on DWT and SVD, Journal of applied research and technology, 12, 750-761.

[11] Sharma, C., Amandeep, (2018). Video Watermarking scheme based on DWT,SVD, Rail Fence for quality loss of data, 4th International Conference on Computing Sciences.

[12] Panyavaraporn, J., and Horkaew, P. (2018). DWT/DCT-based Invisible Digital Watermarking Scheme for Video Stream, IEEE conference, 154-157.

[13] Chen, W., Li, X., Zhan, S., Niu, D., (2018). Multimedia Video Watermarking Algorithm Using SVD and Secret Sharing, 2nd IEEE Advanced Information Management Communicates Electronic and Automation Control Conference (IMCEC), 1682-1686.

ELECTRICITY CONSUMPTION FORECASTING SYSTEM USING ARIMA MODEL

Niharika[#1], Jaswinder Singh[#2], Harpreet Kaur[#3]
Department of Computer Science & Engineering, Punjabi University, Patiala (Pb.)
[1]sharmaniharika94182@gmail.com
[2] dr.jaswinder@pbi.ac.in
[3]Harpreet.ce@pbi.ac.in

ABSTRACT— This paper provides a comprehensive overview of research related to the machine learning model which is used in predictive analysis. Machine learning is in this new era, which has long been trying to rise to the promise of being consistent and accurate. This system is effectively trying to learn from the data and make a prediction out of it. So, machine learning and predictive analysis go hand in hand. The main goal and contribution of Vision are to support research in predictive analysis so that scientists can use it to build machine learning models and be effective, estimating future value with greater accuracy and efficiency. We have taken data of electricity consumption of India for the year 2019-2020 and have implemented a machine learning model and calculated the accuracy of the model. In this paper, we propose a method for predicting electricity consumption for the next 4 months, based on the ARIMA methodology. ARIMA methods and Seasonal ARIMA methods are used for time series analysis, and in the past were mainly used for load forecasting due to their accuracy and mathematical validity.

KEYWORDS— Electricity consumption, ARIMA, Forecasting, SARIMAX, Seasonality.

I. INTRODUCTION

Machine learning is a branch of artificial intelligence and is a method of data analysis that automates the analytical model building or which trains the machines to learn. The basic idea of the machine learning system is that it can learn from the data, identify the patterns and try to make decisions without any human intervention. Today, machine learning is becoming increasingly important due to the growing number and variety of data, calculations, data processing, that is, it is cheaper and more efficient, and available for data storage. This means that you can quickly and automatically create models that can analyze larger, more complex, and data for faster and more accurate results even at very large scales.

Predictive analytics - A section of advanced analysis that is used to predict unknown future events. It comprises a variety of statistical techniques and uses statistics to estimate or predict future outcomes. This will help us understand what's in the future by analyzing the past. Predictive analysis and machine learning go hand in hand because predictive models tend to be one car in an algorithm. Over time to time, these models can be trained to respond to new data or values, giving the results the business needs.

So, predictive models are of two types**:**

1. Classification models, that predict class membership.
2. Regression models, which predict a number.

These all models are made up of algorithms that perform the data mining and statistical analysis, determining the trends and the patterns in data.

Predictive analytics Process:

A. Defining Project

First of all, we will define the project their outcomes, objectives, deliverables, scoping of the effort as well as the input which will be used. And the identification of data sets that are going to be used. All data sources are available, up-to-date, and the expected size of the data analysis.

B. Collection of data

Since predictive analytics is the use of large amounts of data to get information about the latest events, and to advance in the game, the data collection stage is crucial for the success of the initiative. Data will be collected based on a specific project. Data mining for predictive analysis prepares data from multiple sources for analysis that means data from a lot of places. This also includes a picture of various customer interactions as a single view item.

C. Data Analysis

Once we get all the details that need to be in place, it's time to analyze them. The information can be verified, cleaned, transformed, and maintained. Find out useful information from it. Once this is completed, the results should be interpreted and achievable goals followed for far-reaching conclusions.

D. Statistics

Statistical analysis allows validating if the findings, assumptions, hypotheses are admirable to go ahead and test them using the standard statistical models. So that decisions are made based on numbers.

E. Modeling

Models often recommend using one of the existing tools. There are many libraries based on open source programming languages such as Python and R to explore all possible options and choose the best one for the job. Predictive modeling makes it possible to have an accurate predictive model of the future, which can choose the best option, which can be sorted using a multi-model.

F. Deployment

The deployment model allows you to select and implement analysis results that allow you to be effective in your decision-making process. It also helps you get results, reports, and outputs by automating simulation-based solutions.

G. Model Monitoring

Finally, a monitoring model should be created. Models are valid for a certain period, since the external conditions, in principle, do not change. In this way, the model can be managed and monitored to test the operation of the model to provide the expected results.

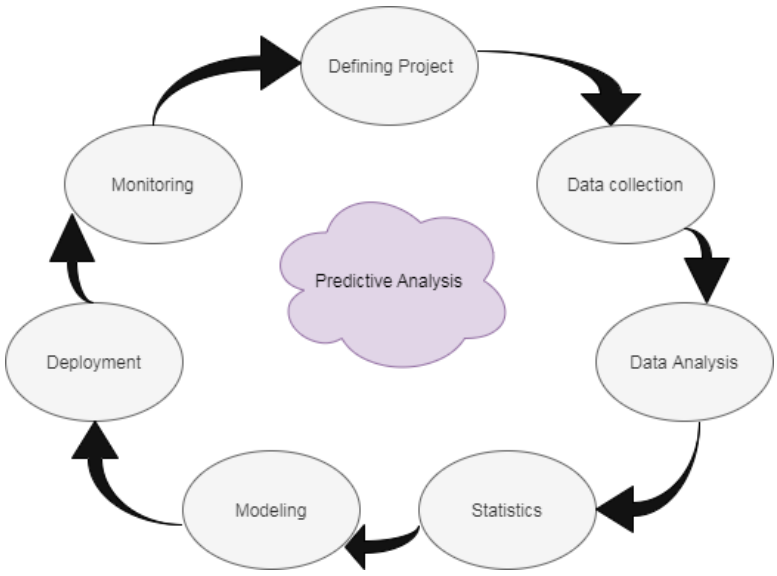


Fig.1. Predictive Analysis Flow Process.

In this paper, I have explained the time series analysis using ARIMA or SARIMA model.

Time series is a sequence where a metric is recorded over a regular time interval. Time series can be annual, quarterly, monthly, weekly, daily, hours, minutes, or even seconds. Time series forecasting is a method that can be used to predict events over subsequent periods. This method predicts future events by analyzing past trends and assumes that the trend in the future will look like a historical style. This method has been used in many areas of research.

The main two goals of time series analysis are:

1) To determine the nature of a phenomenon, which is represented by a sequence of observations and

2) Forecasting.

For the forecasting part, we need to build the model, which will help in predicting the future values.

The dataset used in this project is the power consumption data of the year 2019-2020 which has been scraped from the weekly energy reports of POSOCO (Power System Operation Corporation Limited) (www.kaggle.com). This data is a time-series data that contains data on a daily basis and my main objective is to compare the forecasted value against the actual data and also to forecast the electricity consumption for the next coming month. Because every time, the energy generation exceeds the energy consumption values due to this some states suffer from an insufficient amount of power generation and also resulting in a power loss. This model will help the corporation to generate a sufficient amount of electricity closer to the consumption so that there will be no shortage in any state and the energy would be saved as well.

An Auto-Regressive Integrated Moving Average (ARIMA) model has been already used to predict product prices [1], [2], such as oil [3] and gas [4]. ARIMA techniques were also been used for load forecasting in power systems [5], [6] which showed good results. Currently, with the restructuring process that occurs in many countries; Auto-Regressive (AR) models, such as the Norwegian system are used to predict weekly prices [7]. Besides methods of Artificial Neural Networks (ANN) that are widely used for load forecasting are now used for price prediction [8] [9][10] [11]. ARIMA–MetaFA–LSSVR model was superior to the others in terms of performance measures in predicting 1-day-ahead electricity consumption of air conditioners [12]. Somehow The ANN model is robust, efficient, and accurate, and it produces better results for any day of the week in forecasting Electricity Price for PJM by giving MAPE -9.72 [13].

This article focuses on the 4 months preceding forecasting electricity consumption using ARIMA models. That is, in this work, ARIMA models are used to predict the next daily consumption of electricity over 4 months. These models are based on time series analysis and provide accurate and reliable forecasts of electricity market prices of California [14] and Spain [15]. The following model is implemented in python language using a jupyter notebook.

II. METHODOLOGY

ARIMA and Seasonal ARIMA

ARIMA stands for Auto-Regressive Integrated Moving Averages.

ARIMA processes are a class of stochastic processes used to analyze time series [16]. And ARIMA is a type of model that explains the information in the past values of the time series i.e., its lags and delays in prediction errors, so that equation can be used to predict the future values. A time series that shows samples of white noise rather than random noise can be modeled using ARIMA models. The application of the ARIMA methodology for the study of time series analysis is due to Box and Jenkins [17].

The general statistical methodology of the ARIMA model is as follows :

A. Visualize the time series data
B. Make the time series data stationary
C. Plot the correlation and autocorrelation charts.
D. Construct the ARIMA model or seasonal ARIMA based on the data
E. Use the model to make predictions.

A. Visualizing the time series data

In the visualization of the data, we went through the electricity consumption data. The data has been cleaned by removing any NAN values and updating the column names. As our data is daily based data and contains the date, state-wise consumption, and total consumption. In this model, we are only dealing with the date and the total consumption. Making a date-time as an index, we are then further visualizing the data by plotting the time series data on the graph. And analyzing the pattern we find out that the data is seasonal data which means that at some point of time the consumption is high and at some point of time the consumption is somewhat low and the data is repeating that same pattern until the end.

The following graph is the generalized visualization of the total electricity consumption of India :

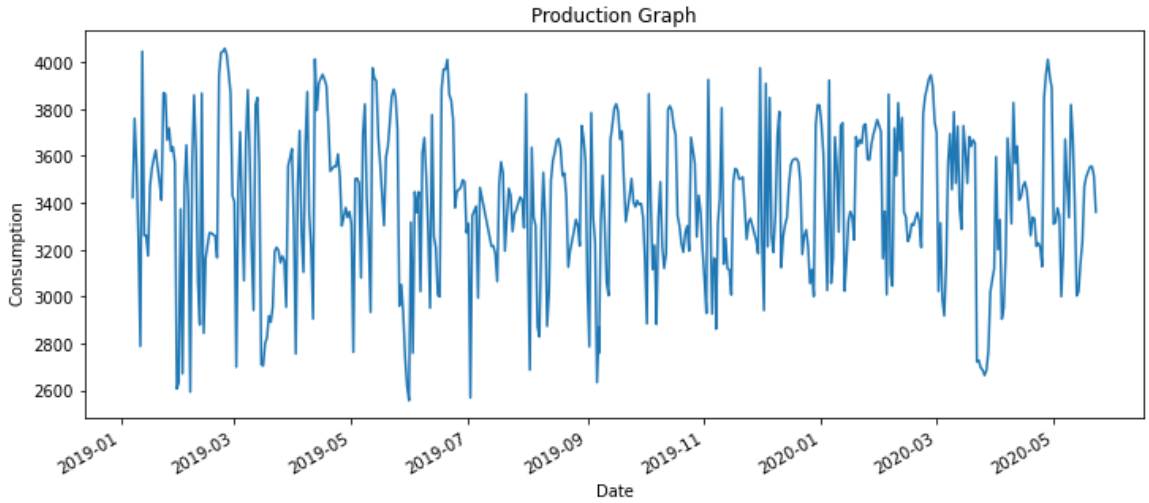


Fig.2. Graph of Total Consumption of Electricity of India for the year 2019-2020.

B. Stationarising the time series

After visualizing, a stationarity test is done on the time-series data. A stationary time series means whose statistical properties like mean, variance, autocorrelation, etc. are constant over time. A stationarised series is easy to predict, as we simply predict these statistical properties to be the same in the future as they have been in the past. Stationarising the time series is done to obtain meaningful sample statistics like mean, variances, and correlations with another variable. If the time series is not stationary, it needs to be converted into a stationary series. To check if a series is stationary or not, the following test will be done.

Dickey-Fuller Test

The Dickey-Fuller test is a unit root testing that test the null hypothesis that $\gamma=1$ in the following equation if $\phi=0$ in this model of the data:

$$y_t=\alpha+\beta t+\phi y_{t-1}+e_t \qquad \text{-------(1)}$$

which is written as :

$$\Delta y_t=y_t-y_{t-1}=\alpha+\beta t+\gamma y_{t-1}+e_t \qquad \text{-------(2)}$$

Where yt is the data, α is a constant, β is the coefficient on a time trend.
Using Eq(2), we can do a linear regression of Δyt against t and $yt-1$ and test if γ is different from 0. If $\gamma=0$, then we have a random walk process. If not and $-1<1+\gamma<1$, then we have a stationary process.

Augmented Dicky-Fuller test

The Augmented Dickey-Fuller test allows for higher-order autoregressive processes by including $\Delta yt-p$ in the model. But our test is still if $\gamma=0$.

$$\Delta yt=\alpha+\beta t+\gamma yt-1+\delta 1\Delta yt-1+\delta 2\Delta yt-2+\ldots+\delta p\text{-}1\Delta yt\text{-}p+1+et \quad \text{--------(3)}$$

Where yt is the data, α is a constant, β is the coefficient on a time trend, and p the lag order of the autoregressive process.

The null hypothesis is non-stationary for both tests.

H_0 (null hypothesis) = It is non- stationary.

H_1 (alternate hypothesis) = It is stationary.

For this test, we want to REJECT the null hypothesis, so we want a p-value of less than 0.05 (or smaller), to infer that a series is stationary.

Following is the table showing the values of the ADF (Augmented Dicky Fuller) test on the aforementioned time-series data:

TABLE.1.
VALUES OF ADF TEST

ADF value	-5.940089379866561
p-value	2.2717840848590585e-07
Number of lags	7
No. of observations used for ADF regression and critical values calculations	495
Critical Value	1% : -3.4436298692815304 5% : -2.867396599893435 10% : -2.5698893429241916

The electricity consumption data that we have taken rejected the null hypothesis, by giving a p-value less than 0.05 which means the data is stationary.

C. Plot the correlation and autocorrelation

Before building a predictive model, we need to determine the optimal parameters of the model; we will get a stationary series. And for those parameters, we need ACF and PACF plots.

A non-seasonal ARIMA model is classified as an 'ARIMA (p,d,q)' model, where

p is a number of autoregressive terms,

d is a number of non-seasonal differences needed for stationarity

q is a number of lagged forecast errors in the prediction equation.

So the value of p and q come through ACF and PACF plots.

Autocorrelation Function (ACF)

Statistical correlation estimates the strength of the relationship between two variables. Pearson's correlation is a number between -1 and 1 that describes a negative and positive correlation respectively. And zero value indicates a no correlation.

The correlation for time series observation can be calculated with previous time steps, called lags. And this is called serial correlation or autocorrelation. And a plot of the autocorrelation of a time series by lag is called the Autocorrelation Function (ACF). Identification of the MA model is done with ACF.

Partial Autocorrelation Function (PACF)
Partial autocorrelation to estimate the relationship between observations in the time series of observations in the previous stages of the relationship between observations is removed. Partial autocorrelation there is a correlation in the k lag, i.e. after removing the effects, a correlation with a short lag period is possible. Here it is, the Identification of an AR model can be done with the PACF.
The following graph is the ACF and PACF graph obtained from the data:

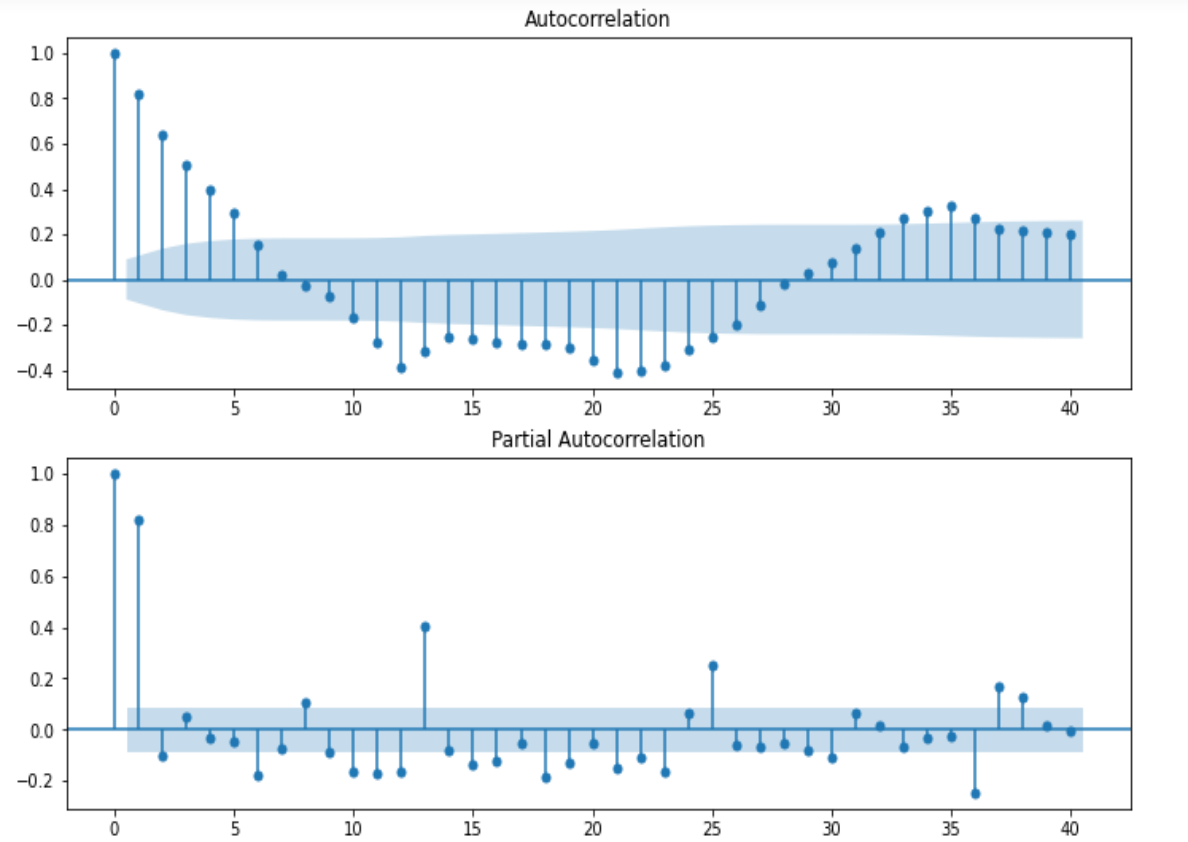


Fig.3. ACF (above) and PACF (below) Plot of the data.

D. Construct the ARIMA model or Seasonal ARIMA based on the data

Now we have optimal model parameters, we can fit a Seasonal ARIMA model to learn the pattern of the series.

The below graph is the comparison of the predicted value against the actual data. However, the predicted value fitted in the model somehow. And it is also following the trend of the actual data.

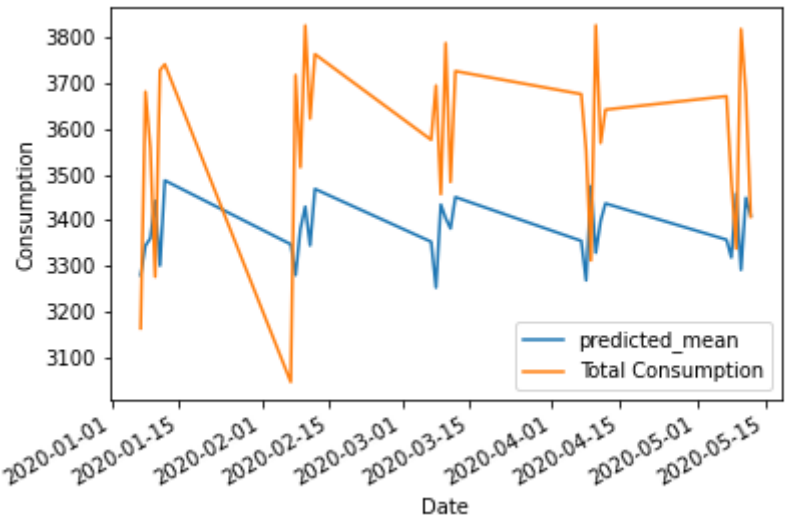


Fig.4. Actual and predicted forecast of the total consumption of electricity.

Measuring the predictive accuracy of the model, MAPE is used i.e. Mean Absolute Percentage Error. The formulae for calculating the MAPE is:

MAPE = (1/n) * Σ(|actual – prediction| / |actual|) * 100

MAPE shows the average difference between the predicted value and the actual value. A lower value indicates a better prediction model.

E. Use the model to make predictions

After completing the model, we will be able to predict the future at this stage. Finally model will get ready to predict.

III. RESULT

Seasonal ARIMA models have been applied to predict the electricity consumption of India. Daily data from the year 2019 to 2020 is used to forecast the consumption for the next 4 months. Numerical results with the Seasonal ARIMA model are presented. Fig. 2-4 shows all the outcomes resulting from the Seasonal ARIMA model. After applying the model, the model is somehow able to forecast the predicted consumption value against the actual consumption value in Fig.4.The proposed model obtained the MAPE of 7.17%. Also, the model can forecast the new consumption values for the next 4 months.

The following graph is the forecasting of electricity consumption of India for next 4 months.

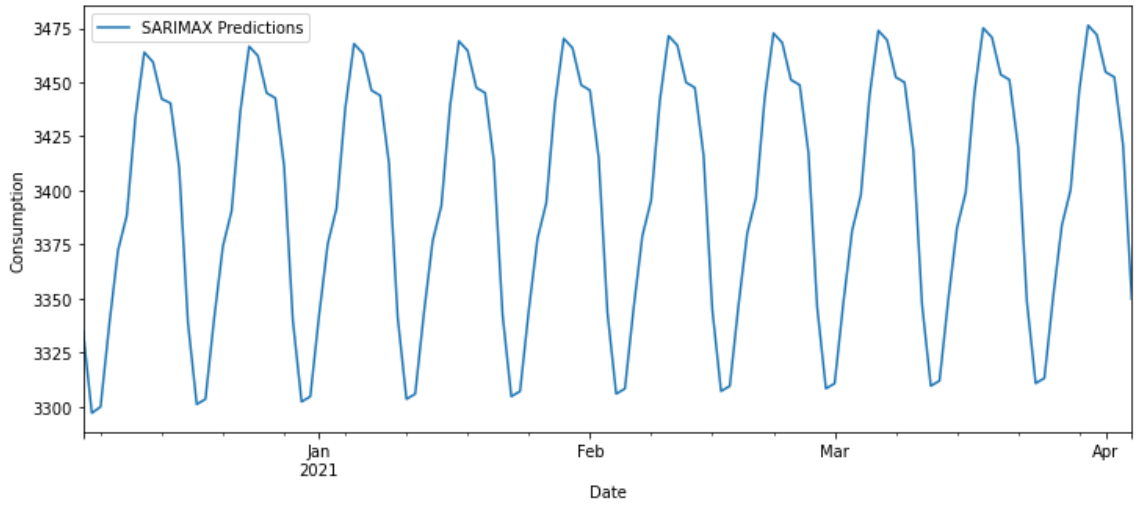


Fig.5. Final forecasting of daily electricity consumption of India for the next 4 months.

IV. CONCLUSION

The main contribution of this review is to discuss the machine learning model used for time series data. This paper proposes a Seasonal ARIMA model to predict the daily consumption of electricity for 4 months. To implement the ARIMA model we need to stationarise a time series data. After making the data stationary, only then a model is constructed. Once the model is constructed, we can use it to make predictions. So after implementing the model, it can forecast the values against actual data and the forecasted value is somehow matching with the actual values. This model is also able to forecast the daily consumption of electricity for the next 4 months. Its MAPE (Mean Absolute Percentage Error) is 7.17%. In ARIMA model, there are so many parameters and because of this, they don't generalize well. So, for better accuracy, we can use another model like the ANN (Artificial Neural Network) model for the prediction of time-series data.

REFERENCES

[1] E. Weiss, "Forecasting commodity prices using ARIMA," *Technical Analysis of Stocks & Commodities, vol. 18, no. 1,* p. 18–19, 2000.

[2] M. L. a. O. C. M. Chinn, "The predictive characteristics of energy futures: Recent evidence for crude oil, natural gas, gasoline and heating oil," [Online]. Available: http://people.ucsc.edu/~chinn/energyfutures.pdf.

[3] C. Morana, "A semiparametric approach to short-term oil price forecasting," *Energy Economics, vol. 23, no. 3,* p. 325–338, May 2001.

[4] K. Buchananan et.al., "Which way the natural gas price: An attempt to predict the direction of natural gas spot price movements using trader positions," *Energy Economics, vol. 23, no. 3,* p. 279–293, May 2001.

[5] F. D. Galiana et.al., "Short-Term load forecasting," *Proc. IEEE vol. 75, no. 12,* p. 1558–1573, Dec 1987.

[6] S. M. Behr et.al., "The time series approach to short term load forecasting," *IEEE Trans. Power Syst., vol. 2,* p. 785–791, Aug 1987.

[7] B. Fosso et.al., "Generation scheduling in a deregulated system. The Norwegian case," *IEEE Trans. Power Syst., vol. 14, no. 1,* p. 75–81, Feb 1999.

[8] S. Hippert et.al., "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans Power Syst., vol. 16,* p. 44–55, Feb 2001.

[9] B. Ramsay et.al., "A neural network-based estimator for electricity spot-pricing with particular reference to weekend and public holidays," *Neurocomputing, vol. 23,* p. 47–57, 1998.

[10] B. R. Szkuta et.al., "Electricity price short-term forecasting using artificial neural networks," *IEEE Trans.Power Syst., vol. 14,* p. 851–857, Aug 1999.

[11] D. Nicolaisen et.al., "Price signal analysis for competitive electric generation companies," in *Proc. Conf. Elect. Utility Deregulation and Restructuring and Power Technologies*, London, U.K, Apr. 4–7, 2000.

[12] Jui-Sheng Chou et.al., "Hybrid Machine Learning System to Forecast Electricity Consumption of Smart Grid-Based Air Conditioners," *IEEE SYSTEMS JOURNAL,* pp. 1-9, 2019.

[13] Paras Mandal et.al., "A Novel Approach to Forecast Electricity Price for PJM Using Neural Network and Similar Days Method," *IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 22, NO. 4,* pp. 2058-2065, November 2007.

[14] P. Basagoit, "Spanish power exchange and information system design concepts, and operating experience," in *Proc. 21st Power Ind.Comput. Applicat. Int. Conf.* , Santa Clara, CA, May 1999.

[15] Z. Alaywan et. al., "Implementation of the California independent system operator," in *Proc. 21st Power Ind. Comput. Applicat. Int.Conf*, Santa Clara, CA, May 1999.

[16] J. E. R. N. F. J. &. C. A. J. Contreras, "ARIMA Models to Predict Next-Day Electricity Prices," *IEEE TRANSACTIONS ON POWER SYSTEMS, VOL. 18, NO. 3,* pp. 1014-1020, AUGUST 2003.

[17] P. Box et.al., Time Series Analysis Forecasting, and Control, Englewood Cliffs, NJ: Prentice-Hall, 1994.

A PREDICTION SYSTEM FOR CONFIRMED VS CURED AND DEATH RATE OF COVID-19

Ankush Kumar[#1], Dhavleesh Rattan[*2]
[#]*Punjabi University, Patiala*
[1]ankushkumar1994@gmail.com
[2]dhavleesh@gmail.com

ABSTRACT— Prediction is a technique that has been used by almost every field of science and technology, health and education, business and stock market to take an attempt to analyse the current situation and predict the futuristic values based on current data. The similar approach has been used in the proposed study where the COVID-19 real-time dataset from Health Ministry Government of India has been used to analyse the confirmed, cured and death cases and hence a predictive view of the future has been given. The presented work reports a 96% R2 score for Random Forest algorithm whereas Linear Regression reported 87% and Polynomial Regression resulted with 55% R2 score in case of confirmed vs death rate whereas 90%, 54% and 97% R2 score for Linear Regression, Polynomial Regression and Random Forest Regression in case of Confirmed vs cured cases . The paper aims to provide an insight of the COVID-19 confirmed vs death rate to help the health care sector get a better view and necessary actions could be taken on time to help individuals survive health.

KEYWORDS— COVID-19, Linear Regression, Support Vector Machine, Polynomial Regression, Random Forest Regression, Decision Tree Regression

INTRODUCTION

Machine Learning has emerged as the most popular tool for predictive analysis of the data given. Being the multi-disciplinary work domain, machine learning is vastly used wherever one can think of data involvement. Since late 2019, the world has been constantly suffering from the COVID-19 disease. First case was reported in Wuhan, China. Since then, the virus has taken many lives and many are in hospitals currently due to its devastating consequences. The virus has no cure till now, just some precautions may help its spread like maintaining social distance, washing hands frequently and personal hygiene [1]. The medical field is constantly looking for the cure whereas the computer science field is trying to figure out the correlation among the data reported so as to get the farsightedness of the virus infected and destroyed individuals. This helps to get better arrangements in the healthcare sector such as it helps to aim at how much supply of vaccines, masks or remedies are needed for the future. A good prediction algorithm could help in this approach and hence the machine learning approaches are considered in this paper.

For COVID-19 related data, Machine Learning models could be categorised into following:

H. Predictive Models

One of the most prominent applications of Machine Learning is predictive analysis [2]. A variety of algorithms which are related to machine learning that can be utilized in the field of diseases forecasting, weather prediction, stock market data forecasting, disease diagnosis and many others. Standard as well as enhanced versions of algorithms like decision tree regression, linear regression, support vector regression, polynomial regression and random forest regression are present in machine learning scope to help researchers explore all possibilities of future analysis of the data known to them. The machine learning models have been used rigorously in the healthcare sector to help the system to predict diseases at an early stage [3]. A wide range of applicability has been seen in forecasting the range of diseases using machine learning models like prediction of coronary artery [4], heart diseases [5] and forecasting of breast cancer in patients [6].

I. Classification Models

Classification means dividing the combined available data into different classes like yes/no, on/off, active/inactive etc. In the applicability of machine learning classification algorithms on COVID-19, the study presented in [7] has deployed a logistic regression type classification approach on clinical data available. For classification of COVID-19 data, the study[8-9] has implemented a random forest model. For crucial diagnosis of COVID-19, various deep learning models [10] have been used to extract features of chest CT images and deploy the learnt features to classify COVID-19 along with decision tree and Ada-boost techniques. The paper published in 2020 [11], the authors created an end to end network for mapping CT images with labels to help the healthcare sector to identify COVID-19 virus in the patients.

J. Clustering Models

As the data in every field is growing rapidly, the type of data is also increasing drastically and taking the forms like audio, video, text, image and many other forms. Many of the datasets available are in unstructured format hence difficult to analyse. There is an urge to have a technique that can help process, manage and conclude huge amount of data in an easy understandable way [12]. The clustering has been very famous in data mining applications and this machine learning technique has been rigorously studied in the past few years for taking optimum benefits out of it in various application areas like text summarization, biometrics data analysis, segmentation etc. [13]. The issues like cluster validity, durability of clustering approach, behaviour of data and selection of the optimum number of clusters have been discussed extensively.

The nature of the data provided is very crucial to understand as this will help getting clearer analysis of data taken. Exploratory Data analysis (EDA) and Inferential Data Analysis (IDA) are the most extensively used analysis techniques. The EDA technique helps getting the best possible view of the data being analysed whereas IDA tells the process by which the information could be extracted out of the data given as input and hence help making predictive analysis easier.

DATASET DETAILS

The real time data from https://www.mohfw.gov.in/ for a time period of and the details regarding the dataset after the feature extraction step are as well. The dataset has 2594 rows and 3 columns. The data is well cleaned using data pre-processing steps, hence no missing values, and no categorical values are there. The data is available in numerical format and further used to apply various regression techniques to get a better analysis. Fig. 1 showing dataset details in various terms.

	count	mean	std	min	25%	50%	75%	max
Cured	2594.0	538.526600	1898.851092	0.0	1.0	17.0	187.75	30108.0
Deaths	2594.0	44.062452	175.716884	0.0	0.0	1.0	13.00	2362.0
Confirmed	2594.0	1449.380108	5005.322018	0.0	8.0	59.0	788.50	70013.0

Fig. 3 Dataset details

The correlation between the columns taken in the dataset can be presented by the below figure.

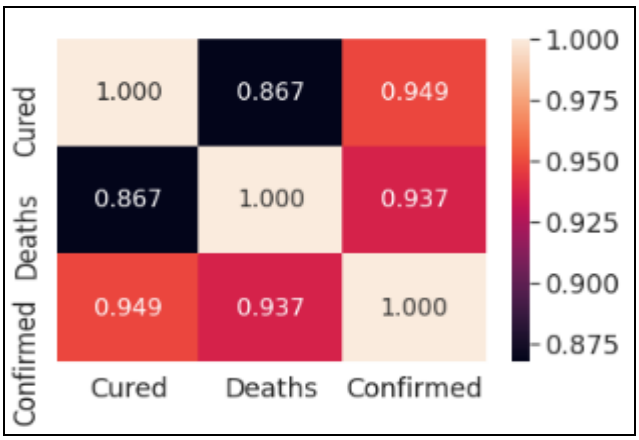


Fig. 2 Correlation matrix

After the data cleaning step, the next ultimate step is to explore the data in a more detailed manner, including what kind of relationship the data among various columns in the dataset have. For that, we can use the Exploratory Data Analysis (EDA) process which helps the researchers finding patterns, relationships and any kind of anomalies in the given dataset. Among various techniques under EDA, the most prominently used method is pairplot also known as scatterplot to get a matrix representation of the columns used in the dataset. Pairplot helps to get the relationship among the various variables available, in our case, its 3 different variables i.e. confirmed, deaths and cured. The below representation we have got after applying EDA on the dataset used.

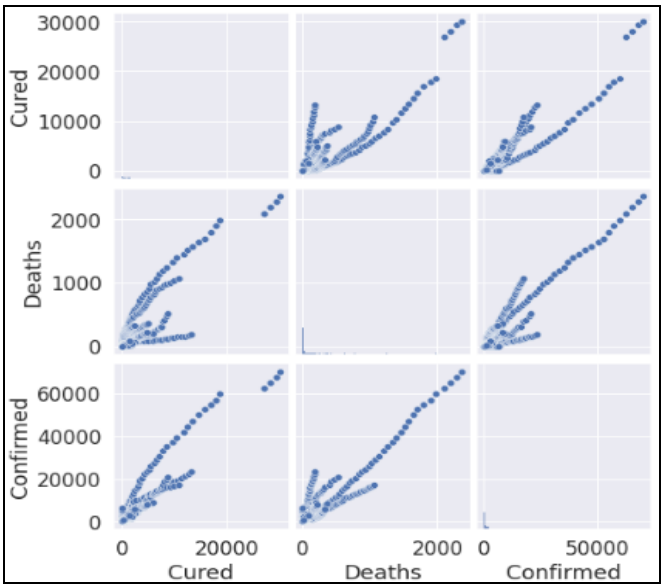


Fig. 3 Exploratory data analysis of COVID-19 data

PROPOSED METHODOLOGY

The proposal has been made for a machine learning model that could work effectively to deliver the accurate results for the prediction of recovery rate as well as death rate of the COVID-19 patients. First of all, the libraries need to be imported including numpy, matplotlib, pandas, seaborn to make the models work correctly. Next usual step is to read the dataset available in the form of a csv file and then dependent and independent variables need to be decided. The need of data preprocessing was there as there were earlier some unnecessary columns like country of origin, states, serial number etc. which was not the point of focus here. The data was checked if there are any missing or null values as well as for any categorical values or scaling is required. Once the data was cleaned, the next immediate step taken was to check if the columns have any correlation among each other. The plotting of the correlation has been depicted in figure 3 using the Exploratory Data Analysis feature of machine learning. Once the data analysis was done, the next step was to split the data into training and testing parts wherein 30% of the data had been reserved for testing purpose and 70% of the data was used to train the model. The real task was now to implement the Linear Regression, Polynomial Regression and Random Forest Regression onto the dataset to see the results for both the scenarios, i.e. confirmed vs death and confirmed vs cured cases. The methodology has been depicted in figure 4 for better understanding.

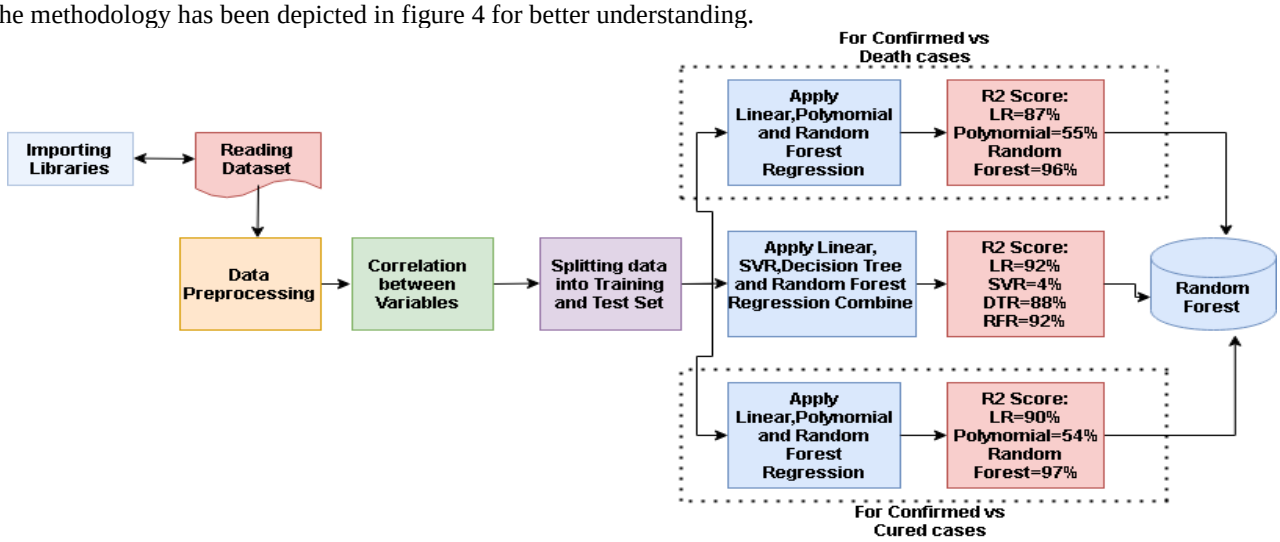


Fig. 4 The proposed methodology

The implementation part has been divided into 3 phases:

K. Confirmed vs Death Cases

This phase has implemented the Linear, Random Forest Regression and Polynomial Regression algorithms individually, to check the accuracy of the given dataset. Linear regression has resulted in 87% of the R2 score whereas Polynomial Regression has given an accuracy of 55% with Random Forest Regression working outstandingly at an accuracy rate of 96%. The figure 5 is representing the algorithms with their plotting using matplotlib library of machine learning.

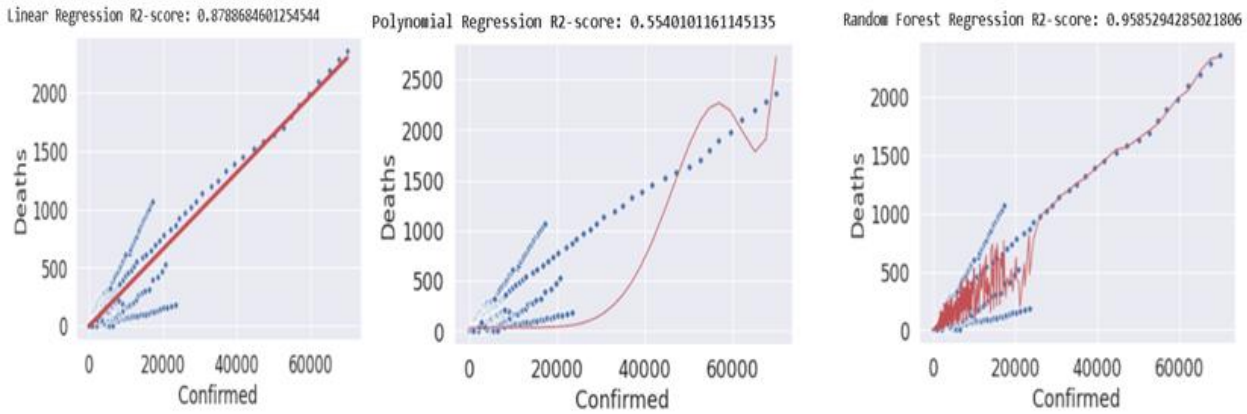


Fig. 5 Confirmed vs Deaths Plotting for Linear, Polynomial and Random Forest Regression

L. Combine Approach

This phase of the approach is opting for a combination of Linear Regression, Decision Tree, Random Forest Regression and Support Vector Regression in one go. The results show that the Linear Regression and Random forest again take the lead to achieve a remarkable R2 score of 92% below which decision tree with 88% and the worst performing was Support Vector with 4% of the R2 score. The system is depicted with the figure 6 below:

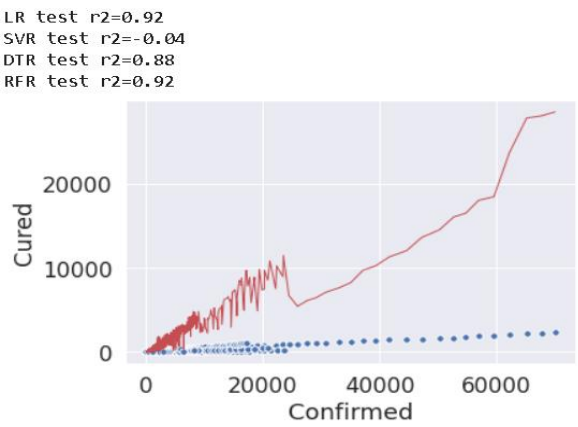


Fig. 6 Combination of Linear, SVM, Decision Tree and Random Forest Regression

M. Confirmed vs Cured Cases

This phase has implemented the Linear, Random Forest Regression and Polynomial individually to check the accuracy of the given dataset for confirmed vs cured cases. Linear regression has resulted in 90% of the R2 score whereas Polynomial Regression has given an accuracy of 54% with Random Forest Regression working outstandingly at an accuracy rate of 97%. The figure 7 is representing the algorithms with their plotting using matplotlib library of machine learning.

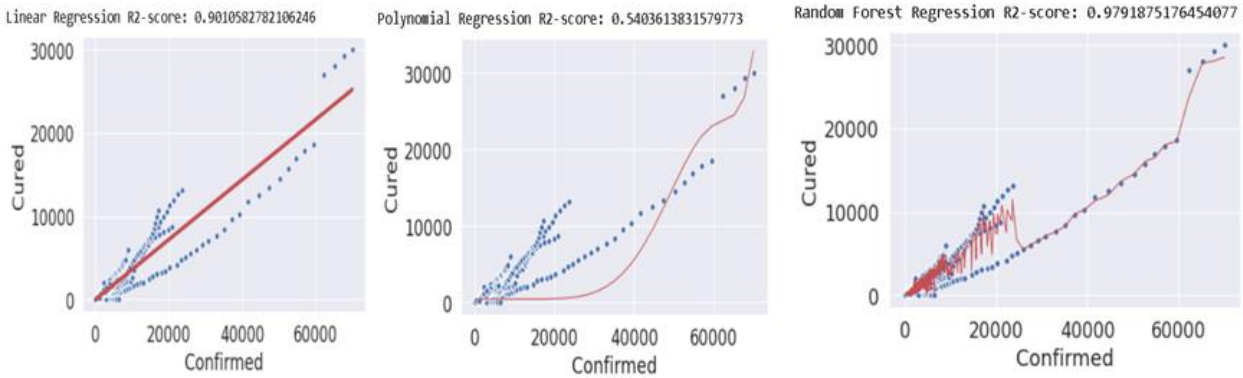


Fig. 7 Confirmed vs Cured Plotting for Linear, Polynomial and Random Forest Regression

CONCLUSIONS

The paper has discussed the regression algorithms on the dataset of COVID-19 patients extracted from the website of the Ministry of Health and Family Welfare. It has been analysed to conclude the most suitable algorithm to fit into the scenario of getting a prediction of COVID-19 recovery rate and death rate in the future based upon the historical data. The overall results conclude that Random forest has outperformed among all the algorithms in prediction of the deaths as well as recovery of the COVID-19 patients. The motive of the study was to help the healthcare sector estimate the need to hospital emergency aids like hospital beds, Oxygen cylinders required, ICUs along with the medicines that could help treat the patients in the best possible way. The study helps the current situation to get the accurate results of the prediction of recovery and deaths that might occur due to the virus and hence an urge of medical aids could be arranged well in advance to combat the situation in a healthy way. The study could be implemented on other diseases prediction in the near future if such a situation may arise. This will help the government and medical field to get a better analysis of the patients' statistics.

REFERENCES

[1] E. Gambhir, R. Jain, A. Gupta, U. Tomer, *Regression Analysis of COVID-19 using Machine Learning Algorithms*, Proceedings of the International Conference on Smart Electronics and Communication, 2020.

[2] G. Bontempi, S. B. Taieb, and Y.-A. Le Borgne, *Machine learning strategies for time series forecasting*, Second European Summer School, eBISS, 2012.

[3] F. E. Harrell Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert, *Regression models for prognostic prediction: Advantages, problems, and suggested solutions*, Cancer Treatment Report, 1985.
[4] P. Lapuerta, S. P. Azen, and L. Labree, *Use of neural networks in predicting the risk of coronary artery disease*, Computers and Biomedical Research, 1995.
[5] K. M. Anderson, P. M. Odell, P. W. Wilson, and W. B. Kannel, *Cardiovascular disease risk profiles*, American Heart Journal, 1991.
[6] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, *Using machine learning algorithms for breast cancer risk prediction and diagnosis*, Procedia Computer Science, 2016.
[7] W. Shi et al., *Deep learning-based quantitative computed tomography model in predicting the severity of COVID-19: A retrospective study in 196 patients*, Annals of Translational Medicine, 2021.
[8] Z. Tang et al., *Severity assessment of coronavirus disease 2019 (COVID-19) using quantitative features from chest CT images*, arXiv, 2020.
[9] F. Shi et al., *Large-scale screening of COVID-19 from community acquired pneumonia using infection size-aware classification*, arXiv, 2020.
[10] S. Wang et al., *A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19)*, MedRxiv, 2020.
[11] X. Xu et al., *Deep learning system to screen coronavirus disease 2019 pneumonia*, arXiv, 2020.
[12] A. K. Jain, *Data clustering: 50 years beyond K-means*, Pattern Recognition Letters, 2009.
[13] Charu C. Aggarwal and Chandan K. Reddy, *Data Clustering: Algorithms and Applications*, Taylor & Francis Group, 2013.
[14] Dheeraj Khera, Williamjeet Singh, *Prediction and Analysis of Injury Severity in Traffic System Using Data Mining Techniques*, International Journal Of Computer Applications, 2015.
[15] Ms. Gagandeep Kaur, Er. Harpreet Kaur, *Prediction of the Cause of Accident and Accident Prone Location on Roads Using Data Mining Techniques*, IEEE, 2020.

A RECENT TRENDS IN IMAGE CONTRAST ENHANCEMENT METHODS: A REVIEW

Jagdeep Singh[1], Er. Rakesh Singh[2], Dr. Navjot Kaur[3]
Department of Computer Science and Engineering, Punjabi University
[1]jagdeep80877@gmail.com
[2]rakesh_ce@pbi.ac.in
[3]navjot@pbi.ac.in

ABSTRACT— Contrast enhancement plays a crucial role in digital image processing. It simply improves the quality and clarity of the image that helps in easy recognization of the content in the image. In the medical field, the image's poor quality and lack of clarity may lead to mistakes in detecting required information from the image. So contrast enhancement methods play an essential role in the medical field by improving image clarity. In this paper, the researcher discusses some of the contrast enhancement methods that have been developed so far and it will help the researchers in understanding and choosing the contrast enhancement method best for their future research.

KEYWORDS— Contrast Enhancement, Contrast Enhancement methods, Histogram Equalization**.**

INTRODUCTION

Digital his images often suffer from problems like overexposure, underexposure, low contrast, noises, etc., which makes the image content hard to recognize. So for this problem, contrast enhancement plays a vital role in improving the image in terms of quality and clarity. It makes the objects in the image recognizable. There are a lot of methods that have been developed for enhancing the contrast. In this paper, the researcher discusses some of these popular methods, i.e., Histogram Equalization (HE), Brightness Preserving Bi-Histogram Equalization (BBHE), Contrast Limited Adaptive Histogram Equalization (CLAHE), and Recursive Mean-Separate Histogram Equalization (RMSHE). These methods enhance the contrast of the image and preserve the brightness, limiting the noises, improving the clarity, so overall improve the quality of the image. These methods play an essential role in the medical field, in which it helps to recognize the organs, tumors, bones, etc. Image processing is also used in sonar image processing, speech recognition, and so forth.

This paper is organized as follows: Section provides the introduction of contrast enhancement. Section II deals with the literature survey, Section presents the contrast enhancement methods, and Section IV concludes this paper.

LITERATURE REVIEW

This section is presenting the research work of some prominent authors from the recent research work done by the researchers in the field of Image Enhancement.

In 2018, Sonali, S Sahu, A K Singh, S P Ghrera, and M Elhoseny proposed a noise removal and contrast enhancement algorithm for fundus image. This algorithm integrates contrast enhancement adaptive histogram equalization (CLAHE) technique and filters to provide an enhanced and de-noised fundus image.

In 2018, L M Satapathy, R K Tripathy, and P Das proposed a new method for contrast enhancement of an image based on the variational mode decomposition (VMD) of image and various histogram equalization approaches. In this the original image is decomposed into sub-images or modes using VMD. The conventional histogram equalization (CHE) and contrast limited adaptive histogram equalization (CLAHE) are applied to mode1 of the original image. A weighting method is used to combine the mode1 image and the original image to get the results. This new method provides better results than the histogram equalization methods that are HE and CLAHE.

In 2018, D Garg, N K Garg, and M Kumar proposed contrast limited adaptive histogram equalization (CLAHE) and percentile methodologies for enhancement of underwater images. These two methodologies are blended to get enhanced image. This method improves the visual appearance and clarity of the image.

In 2018, S Kansal, S Purwar, and R Tripathi proposed maximum entropy bi- histogram equalization method. In this method, maximum entropy of the original image is used, leading to flattening the image histogram. It provides an image with better contrast and clarity.

In 2018, E Reddy, and R Reddy discussed dynamic clipped histogram equalization (DCLHE) method which enhances low contrast image. It selects a clipped level at all the occupied bins and then histogram equalization is performed on clipped histogram which provides an enhanced image. Some methods provide the best results for dark images and some for bright images, but DCLHE gives the best results for all kinds of images.

In 2018, H Mzoughi, I Njeh, M B Slima, and A B Hamida discussed various contrast enhancement methods mainly used for MRI images. This paper compares various contrast enhancement methods to check which technique gives the best enhanced image in detecting tumor in the brain.

In 2019, Y Mousania, and S Karimi proposed an algorithm which uses recursive mean-separate histogram equalization with a fusion of contrast-limited adaptive histogram equalization. This algorithm can improve the brightness and contrast of mammography images and decrease the noise in the images.

In 2019, C Liu, X Sui, X Kuang, Y Liu, G Gu, and Q Chen proposed an adaptive contrast enhancement method based on neighborhood conditional histogram, which improves the quality of thermal infrared images. This method replaced the

clip-redistribution histogram of CLAHE with neighborhood conditional histogram. This method provides better contrast and avoids the over-enhancement of homogenous regions of the image.

In 2019, S K Rupa, M F Yousuf, and S Rahman introduced the mean-separate histogram equalization (MSHE) method. It first finds the most frequent intensity of the image, and then histogram equalization is applied. It provides better contrast and clarity of the image.

In 2019, S F Tan and N A M ISA proposed a modified HE- based contrast enhancement technique for non-uniform illuminated image i.e., exposure region based multi-histogram equalization (ERMHE) capable of enhancing different exposure regions in the image and provides a better quality image.

In 2020, D Vijayalakshmi, M K Nath, and O P Acharya discussed various contrast enhancement methods, mainly on spatial domain. This study examines brightness preservation, entropy preservation, structural information loss, etc. These methods are applied to different datasets and compare all these methods' parameters to check which technique performs best.

In 2020, M Agarwal, G Rani, and V S Dhaka proposed an optimized double threshold weighted constrained histogram equalization method. This method is an integration of Otsu's double threshold, particle swarm optimized weighted constrained model, histogram equalization, adaptive gamma correction, and wiener filtering. This method can effectively detect a tumor in an enhanced MRI image.

In 2020, G Rani and M Agarwal proposed range limited double threshold and weighted histogram equalization with dynamic range stretching (RLDTWHE-DRS). It is an integration of Otsu's double threshold, dynamic range stretching, weighted distribution, adaptive gamma correction, and homomorphic filtering. This method improves the contrast of the image with less visual artifacts.

In 2020, A K Bhandari, S Shahnawazuddin, and A K Meena propose a fuzzy-DCT scheme that includes automatic calculation of the number of parts in which histogram is divided. It was proposed to overcome the problem of calculating the number of parts of histogram division which directly affects the quality of the image. This method provides a clearer and natural enhanced image.

In 2020, W Wang, X Wu, Z Gao, and X Yuan discussed the main methods of low-light image enhancement. This paper discussed various image enhancement methods like gray transformation, histogram equalization, retinex, frequency domain, etc. At last, all these methods' enhanced images' quality parameters are compared to get a better contrast image.

CONTRAST ENHANCEMENT METHODS

There are various methods for image contrast enhancement which the researcher has discussed under this section.

A. Histogram Equalization (HE)

Histogram Equalization is the popular method that helps in improving the contrast of a low contrast image. HE works as spreading the pixels of an image on the dynamic range of gray levels, which ranges from 0 (black) to 255 (white), and equalizes all the pixels that give an enhanced image as a result [1]. It enhances the high frequent gray levels more than other gray levels [5]. The limitation of HE is that it does not preserve the brightness of the image, which results in an over-enhanced image. It degrades the quality of the image and makes the image content hard to recognize.

B. Brightness Preserving Bi- Histogram Equalization (BBHE)

BBHE method was introduced to preserves the mean brightness of the image, which was the limitation of the HE method. BBHE works by dividing the original image into two sub-images based on the mean value of the original image. The histogram of the image ranges from 0 to L-1 (255). The first sub-image contains the values between 0 to mean, and the second sub-image contains the values between mean+1 to L-1. Then BBHE applies histogram equalization method on both these sub-images, and at last, both these sub-images are combined to get an enhanced image. BBHE method's ability to preserve the brightness as well as increasing the contrast makes it the best fit for consumer products [1].

C. Contrast Limited Adaptive Histogram Equalization (CLAHE)

CLAHE method is an extended version of adaptive histogram equalization, which was proposed to overcome the noise problem in the image. It works on the small regions in the image by enhancing the local contrast of each region of the image. It computes multiple histograms of each region of the image. It applies histogram equalization method to each region to produce an enhanced image.

CLAHE avoids over-enhancement of the image and produces a naturally enhanced image with less noise and better quality.

D. Recursive Mean-Separate Histogram Equalization (RMSHE)

RMSHE works like BBHE by dividing the original image into several sub-images on the basis of the mean values of the image. However BBHE method splits the image only once, but RMSHE method splits the image recursively, up to recursive level r (set by the user), generating 2r sub-images. At last, it applies the histogram equalization method on these sub-images to produce equalized sub-images.

RMSHE produces the best results as compared to other above-discussed methods, i.e., HE, BBHE, CLAHE, by preserving the brightness and enhancing the contrast of the image and gives a natural enhanced image.

E. Maximum Entropy Bi- Histogram Equalization

This method enhances the contrast of the image by first calculating the discrete level of the image based on each entropy entry and maximum entropy. Then the image is divided into two sub-images which are then equalized independently and at last combine both of the sub-images to get the resultant image.

It improves the contrast and gives the highest entropy which results in clarity of the image [5].

F. Dynamic Clipped Histogram Equalization (DCLHE)

DCLHE preserves the intensity levels of the low contrast image. It divides the image into small portions, i.e. clipped levels on which histogram equalization is applied to get an enhanced image.

It performs uniform degree enhancement and has the highest entropy which gives a natural enhanced image [12].

G. Exposure Region based Multi-Histogram Equalization (ERMHE)

ERMHE divides the image histogram into sub-histograms and computes threshold of each sub-histogram. And then computes dynamic gray level range. At last, histogram equalization is applied to equalize each sub-histogram which gives an enhanced image [14].

H. Optimized Double Threshold Weighted Constrained Histogram Equalization

It first divides an image based on otsu's double threshold method and applies weighted constrained model to modify probabilities of sub-histograms and then applies HE and automatic gamma correction on each sub-histogram. In the end, it applies wiener filtering to remove the noise and gives an enhanced image with better contrast and quality [17].

I. Range Limited Double Threshold and Weighted Histogram Equalization with Dynamic Range Stretching (RLDTWHE-DRS)

RLDTWHE-DRS first divides based on ostu's double thresholds and after applies dynamic range stretching and weighted distribution model to modify probabilities of each sub-histogram and then HE is applied to equalize each sub-image. At last, gamma correction and homomorphic filtering are applied for global and local contrast enhancement [15].

J. Adaptive Contrast Enhancement based on Neighborhood Conditional Histogram

This method replaces clip-redistribution histogram of CLAHE with the neighborhood conditional histogram to avoid block artifacts. It replaces local mapping function with global mapping function and then both local and global mapping functions are combined to produce an enhanced image [9].

It improves the contrast and quality and avoids over-enhancement of the image.

K. Mean-Separate Histogram Equalization (MSHE)

MSHE enhances image contrast by first calculating the image mode with a weighting factor and a bias, which divide an image into two sub-images on which histogram equalization is applied and then combine both sub-images [13].

It produces a better quality image with improved contrast.

L. Fuzzy-DCT Scheme

It first calculates the number of clusters and then fuzzy c-means clustering method is performed and HE is applied on each cluster individually. Lastly, DCT is applied on the equalized image which produces an enhanced image with improved contrast and preserved brightness [18].

M. Variational Mode Decomposition (VMD) and Histogram Equalization

VMD divides the image into modes. In which HE and CLAHE are applied on the mode to get an equalized image and then weighting method is used to combine both the original image and the equalized image to produce an enhanced image with better contrast [8].

TABLE VI

ANALYSIS OF VARIOUS CONTRAST ENHANCEMENT METHODS.

Authors	**Methods**	**Datasets**	**Descriptions**
Sonali, S Sahu, A K Singh, S P Ghrera, and M Elhoseny [7], 2018	Contrast Limited Adaptive Histogram Equalization (CLAHE)	STARE database	Removes noise and improves contrast in fundus images
M Satapathy, R K Tripathy, and P Das [8], 2018	Variational Mode Decomposition (VMD) and Histogram Equalization Methods	Kodak database	Provides better contrast as compared to the histogram equalization methods

D Garg, N K Garg, and M Kumar [10], 2018	Contrast Limited Adaptive Histogram Equalization (CLAHE) and Percentile methodologies	Sample Images	Provides more visibility and better colors and clarity
S Kansal, S Purwar, and R Tripathi [5], 2018	Maximum Entropy Bi-Histogram Equalization	Test image	Improves image clarity
E Reddy, and R Reddy [12], 2018	Dynamic Clipped Histogram Equalization (DCLHE)	Test images	Preserves entropy, preserves all the gray levels and gives uniform degree enhancement
H Mzoughi, I Njeh, M B Slima, and A B Hamida [11], 2018	Contrast Enhancement Methods	BRATS database	AHE technique provides better contrast for MRI images as compared to other HE methods
Y Mousania, and S Karimi [6], 2019	Recursive Mean-Separate Histogram Equalization (RMSHE) with a fusion of Contrast Limited Adaptive Histogram Equalization (CLAHE)	MIAS database	Improves the image contrast
C Liu, X Sui, X Kuang, Y Liu, G Gu, and Q Chen [9], 2019	Adaptive Contrast Enhancement Method based on Neighborhood Conditional Histogram	Test images	Reduces block artifacts and enhances the local contrast
S K Rupa, M F Yousuf, and S Rahman [13], 2019	Mean-Separate Histogram Equalization (MSHE)	DIP3/e Book Images	Provides better quality image and preserves the brightness
S F Tan and N A M ISA [14], 2019	Exposure Region based Multi-Histogram Equalization (ERMHE)	154 sample images	ERMHE technique provides image with natural appearance, better contrast, less degradation, and detail preservation
D Vijayalakshmi, M K Nath, and O P Acharya [16], 2020	Contrast Enhancement Methods based on Spatial Domain	USC-SIPI	Enhances the contrast and preserves the brightness
M Agarwal, G Rani, and V S Dhaka [17], 2020	Optimized Double Threshold Weighted Constrained Histogram Equalization	CVG-UGR database	Preserves image quality and improves contrast and brightness
G Rani and M Agarwal [15], 2020	Range Limited Double Threshold and Weighted Histogram Equalization with Dynamic Range Stretching (RLDTWHE-DRS)	CVG-UGR database	Enhances the contrast, preserves the brightness and natural appearance
A K Bhandari, S Shahnawazuddin, and A K Meena [18], 2020	Fuzzy-DCT Scheme	Test Images	Improves the contrast and quality of images
W Wang, X Wu, Z Gao, and X Yuan [19], 2020	Image Enhancement Methods	Sample Images	Improves the contrast and image details

CONCLUSION

This paper discusses various contrast enhancement methods that have been proposed so far. The purpose of this review is to analyse different contrast enhancement methods, which will be helpful for the researchers in choosing the contrast enhancement method best for their research. This analysis leaves a future scope of improvement in the contrast enhancement methods, in which further research can be done.

REFERENCES

[16] Kim, Y. (1997). Contrast enhancement using brightness preserving bi-histogram equalization. IEEE Transactions on Consumer Electronics, 43(1), 1-8. doi:10.1109/30.580378

[17] Ziaei, A., Yeganeh, H., Faez, K., & Sargolzaei, S. (2008). A Novel Approach for Contrast Enhancement in Biomedical Images Based on Histogram Equalization. 2008 International Conference on BioMedical Engineering and Informatics. doi:10.1109/bmei.2008.300

[18] Wang, C., & Ye, Z. (2005). Brightness preserving histogram equalization with maximum entropy: A variational perspective. IEEE Transactions on Consumer Electronics, 51(4), 1326-1334. doi:10.1109/tce.2005.1561863

[19] Yadav, G., Maheshwari, S., & Agarwal, A. (2014). Contrast limited adaptive histogram equalization based enhancement for real time video system. International Conference on Advances in Computing, Communications and Informatics (ICACCI). https://doi.org/10.1109/icacci.2014.6968381

[20] Kansal, S., Purwar, S., & Tripathi, R. (2018). Enhancement of Image using Maximum Entropy Bi-Histogram Equalization. 3rd International Conference on Communication and Electronics Systems (ICCES). doi:10.1109/cesys.2018.8724088

[21] Mousania, Y., & Karimi, S. (2019). A Novel Improved Method of RMSHE-Based Technique for Mammography Images Enhancement. Lecture Notes in Electrical Engineering Fundamental Research in Electrical Engineering, 31-42. doi:10.1007/978-981-10-8672-4_3

[22] Sonali, Sahu, S., Singh, A., Ghrera, S.P., Elhoseny, M. (2018). An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE, Optics & Laser Technology, Volume 110, Pages 87-98, ISSN 0030-3992, https://doi.org/10.1016/j.optlastec.2018.06.061.

[23] Satapathy, L. M., Tripathy, R. K., & Das, P. (2018). A Combination of Variational Mode Decomposition and Histogram Equalization for Image Enhancement. National Academy Science Letters, 42(4), 333-336. doi:10.1007/s40009-018-0742-y

[24] Liu, C., Sui, X., Kuang, X., Liu, Y., Gu, G., & Chen, Q. (2019). Adaptive Contrast Enhancement for Infrared Images Based on the Neighborhood Conditional Histogram. Remote Sensing, 11(11), 1381. https://doi.org/10.3390/rs11111381

[25] Garg, D., Garg, N. K., & Kumar, M. (2018). Underwater image enhancement using blending of CLAHE and percentile methodologies. Multimedia Tools and Applications, 77(20), 26545-26561. doi:10.1007/s11042-018-5878-8

[26] Mzoughi, H., Njeh, I., Slima, M. B., & Hamida, A. B. (2018). Histogram equalization-based methods for contrast enhancement of MRI brain Glioma tumor images: Comparative study. 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). doi:10.1109/atsip.2018.8364471

[27] Reddy, E., & Reddy, R. (2018). Dynamic Clipped Histogram Equalization Technique for Enhancing Low Contrast Images. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences, 89(4), 673-698. doi:10.1007/s40010-018-0530-6

[28] Rupa, S. K., Yousuf, M. F., & Rahman, S. (2019). "Contrast Enhancement using Mode-Separate Histogram Equalization," *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pp. 1-6, doi: 10.1109/ICASERT.2019.8934613.

[29] Tan, S. F., & Isa, N. A. M. (2019). "Exposure Based Multi-Histogram Equalization Contrast Enhancement for Non-Uniform Illumination Images," in *IEEE Access*, vol. 7, pp. 70842-70861, doi: 10.1109/ACCESS.2019.2918557.

[30] Rani, G., & Agarwal, M. (2020). Contrast Enhancement Using Optimum Threshold Selection. International Journal of Software Innovation, 8(3), 96-118. doi:10.4018/ijsi.2020070107

[31] Vijayalakshmi, D., Nath, M.K. & Acharya, O.P. (2020). A Comprehensive Survey on Image Contrast Enhancement Methods in Spatial Domain. Sens Imaging 21**,** 40. https://doi.org/10.1007/s11220-020-00305-3

[32] Agarwal, M., Rani, G., Dhaka, V. S. (2020). Optimized contrast enhancement for tumor detection, Int. J. Imaging Syst. Technol., **30**, 687-703. doi: 10.1002/ima.22408

[33] Bhandari, A. K., Shahnawazuddin, S., & Meena, A. K. (2020). A Novel Fuzzy Clustering Based Histogram Model for Image Contrast Enhancement. IEEE Transactions on Fuzzy Systems, (IF 12.029).

[34] Wang, W., Wu, X., Yuan, X., & Gao, Z. (2020). An Experiment-Based Review of Low-Light Image Enhancement Methods. IEEE Access, 8, 87884–87917. https://doi.org/10.1109/access.2020.2992749

SUPERVISED MACHINE LEARNING METHODS: A COMPARATIVE ANALYSIS FOR EPILEPSY SEIZURE DETECTION

Sandeep Singh[1,2*], Harjot Kaur[1]
[1]Department of Computer Science Engineering, Guru Nanak Dev University, Regional Campus, Gurdaspur, India
[2]Department of Computer Science Engineering, SGT University, Gurgaon, India
*E-mail id of corresponding author: er.ss1989@gmail.com

ABSTRACT——Epilepsy is a pathological condition that involves the occurrence of seizures. Seizure detection plays a significant role in diagnosing epilepsy disease. EEG recordings containing human brain activities have been used for diagnosing epilepsy disease. However, manual inspection of vast volume and high rate electroencephalogram (EEG) data is time-consuming, expensive, resource intensive, and error prone. In contrast, automatic analysis of EEG signals enables improving and assisting physicians in diagnosing epilepsy disease by detecting epilepsy seizures. Several machine learning methods have been proposed automatic analysis of EEG signals for epilepsy seizure detection that have reported different results in the literature. Different experimental setup, data sets, and performance metrics make it difficult to find a suitable machine learning classifier for developing an automatic epilepsy seizure detection system.

In this work, we performed a set of comprehensive experiments in a controlled environment using supervised machine learning classifiers and EEG data collected by Neurology and Sleep Centre, New Delhi (NSC-ND). EEG signals are processed to extract time-domain and spectral features. The extracted features are pre-processed by standardizing numeric values to a uniform scale for use with machine learning classifiers. Experiments are conducted by training machine learning classifiers and testing them using a 10-fold cross-validation strategy. Ten independent experiments have been conducted for each machine learning classifiers in this work. Results are presented and analyzed by computing mean and standard deviation. In addition, the mean performance of machine learning classifiers is compared using the most common performance metrics of accuracy, sensitivity, specificity and area under ROC curve.
The reporting results demonstrate that most machine learning classifiers have reported good performance in detecting epilepsy seizures using the benchmark real-time NSC-ND dataset. Gaussian process classifier has reported the best performance among machine learning classifiers in terms of the identified benchmark datasets. It has detected epilepsy seizure up to 88% of accuracy. It is followed by quadratic discriminant analysis (QDA) classified by providing compare table accuracy of 87% in detecting epilepsy seizures. Support vector machine (SVM) classifier with linear and RBF kernels have reported an accuracy of 74% approximately in detecting epilepsy seizures.

Keywords Electroencephalogram (EEG) · Epilepsy · Machine learning · Seizure detection · Seizure detection.

1 INTRODUCTION

Epilepsy is considered the most severe chronic neurological disorder disease. The disease can be diagnosed by analyzing brain signals generated by brain neurons. The brain neurons are interconnected that communicate with human organs and produce various signals in different intensities. Brain signals are generally recorded in electroencephalograms (EEG). No specific reasons for epilepsy seizures or their severity are reported in the domain. Epilepsy disease has distributed throughout the word uniformly [21][4][15].

The main reason behind epilepsy seizures are an electrical signal disturbance in the brain [1][14] due to malformation, lack of oxygen, and low blood sugar level [17][18].

It has been observed that 50 million peoples have been affected by epilepsy disease [1]. Epilepsy disease's prevalence rate is reported to be 1%[19]. In epilepsy disease, the patient experiences multiple seizures that may cause unusual brain activity, sudden breakdown or loss of consciousness. The epilepsy seizure can lasts from few seconds to minutes. Epilepsy seizure can cause different injuries such as burns, fractures etc.[21].

An automatic and intelligent epilepsy seizure detection system can help the physicians an early diagnosis and detection of epilepsy disease [12][13]. Therefore, developing an intelligent epilepsy detection system is necessary for diagnosing the disease by analyzing many EEG data signals.

We compare the performance of different machine learning classifiers using different performance metrics because different classifiers are designed by considering different optimization criteria [8][20]. For instance, support vector machines are developed for minimizing the structural risk hence optimize accuracy [9][6]. Whereas neural networks are being developed for minimizing empirical risk and hence optimize root mean square error. It may be possible that one classifier may show different performance with different performance metrics in the same experimental setup. Major contributions of this work are as below.

1. Processing and decomposing EEG signals for extracting time-domain and spectral features.
2. Extracting time-domain and spectral features from decomposed ECG signals.
3. Pre-processing extracted time-domain and spectral features for processing with machine learning classifiers.

4. Empirical comparison of machine learning classifiers to identify the best performing in detecting epilepsy seizures.

We perform a set of experiments using machine learning methods for detecting epilepsy seizure and a benchmark real-time data set collected by Neurology and Sleep Centre, New Delhi. We conducted ten independent experiments for ten machine learning classifiers and computed their performance using a 10-fold cross-validation strategy. The mean performance of machine learning classifiers is compared to analyze the better performance of machine learning classifiers in detecting epilepsy seizures.

Rest of the paper is organized as follows. Section 2 describes the machine learning methods used in this study. Section 3 explains the methodology followed in this work by describing different stages in detecting epilepsy seizure, benchmark dataset, and experimental setup for conducting comprehensive set of experiments. Section 4 presents the experimental results obtained in this work and provides an analysis of the reporting results. Finally, the paper is concluded and provided clues for extending this work in future in Section 5.

2 MACHINE LEARNING METHODS

This section introduces the machine learning methods used in this study as below. These methods have been successfully in different domains including intrusion detection, computer vision [10].

2.1 k-Nearest Neighbors (KNN)

KNN is a supervised machine learning method for solving classification and regression problems [11]. It works by analysing the distance between input data samples. It is also known as non parametric machine learning methods as it does not learn any explicit model function during the training process. It simply learns all previous instances and predicts the outcome by searching through the training data set for k nearest neighbours of the instance. For the classification task, KNN classifier predicts the majority class among k nearest neighbours. In contrast, it predicts the output based upon average values of K nearest neighbour for regression problems [10]. KNN classifier is also known as instance based learning or memory based learning method and is widely used as lazy learning that delays learning till production is performed. Therefore it reduces the computational overhead till the prediction phase.

2.2 Support vector machine (SVM)

SVM is the most commonly used machine learning model for a classification task that match the input data into higher dimensional feature space to make it linearly separable by using hyperplane [23]. The hyperplane is used to distinguish between different instances and protect their target class in case of the classification problem. It is also used in regression problem, and hyperplanes predict continents value as an output. SVM use kernel function to map input space to higher dimensional nonlinear feature space. The most commonly used are kernel functions are linear kernel function, radial basis kernel function and polynomial kernel function. Different kernels are used for different applications.

2.3 Gaussian Process Classifier (GPC)

GPC is a supervised machine learning algorithm. Gaussian Processes are a generalization of the Gaussian probability distribution. They can be used as the basis for sophisticated non-parametric machine learning algorithms for classification and regression.

GPC is a type of kernel model, like SVMs. But, GPC is capable of predicting highly calibrated class membership probabilities, although the choice and configuration of the kernel used at the heart of the method can be challenging.

2.4 Decision Tree (DT)

DT is a supervised machine learning method that involves building tree shaped graph to predict possible output corresponding to input values. The built tree contains one root element and some internal elements called decision nodes, used to test the input against a learnt expression. The leaf nodes of the tree correspond to the final prediction of the classifier. The decision tree is used to drive decision rules for solving the decision problem by starting at the root node and moving downward to word leaf nodes to predict the target class. Decision tree achieves accuracy for linearly separable data. Constructing an optimal decision tree e is considered an NP complete problem [16]. Many variants have been proposed in decision trees classifiers such as Iterative Dichotomiser 3 (ID3), and C4.5.

2.5 Random Forest (RF)

RF is an ensemble classifier based upon the bagging technique. It uses a decision tree as a base classifier. The bagging technique involves generating multiple base classifiers, and it produces the final output based upon individual predictions of multiple base classifiers. The bagging process helps to reduce the overall variance of the output [16]. Random forest classifier works on building a large number of relatively and correlated decision tree models, which was a committee to predict the final output of the ensemble classifier.

2.6 Neural Network (NN)

NN is a computational model that mimics the biological neural network concerning its structure and functioning. It provides a nonlinear statistical data modelling that involves the multipart association of money input data and output data [19]. Multilayer perception (MLP) is the most popular neural network architecture [2]. It is a feed forward neural network

consisting of different layers of inter connected neurons. It contains three layers: the input layer, hidden layer, and output layer containing different neurons in each layer. Each neuron performs a biased weighted sum of its inputs and applies an activation function to transfer its output to the next layer. MLP can model any arbitrary complexity with a number of layers and the number of units in each layer. During the training process, weights are optimised to obtain minimum error at the output layer [2].

2.7 AdaBoost

AdaBoost is one of the most popular algorithms to construct a strong classifier with linear combination of member classifiers. The member classifiers are selected to minimize the errors in each iteration step during training process. AdaBoost provides very simple and useful method to generate ensemble classifiers. The performance of the ensemble depends on the diversity among the member classifiers as well as the performance of each member classifiers.

2.8 Naive Bayes (NB)

NB classifier is a probabilistic machine learning classifier defined based upon Bayes theorem [5]. It performs the classification task based upon the assumption that features are independent of each other in predicting the target class. It has been proved that NB classifier is very effective for classification task due to its simplicity, even in large sized data sets [5]. It performs well in the case of categorical input variables compared to the numerical variable(s). It suffers from many limitations, including action on dictionary independent features.

2.9 Quadratic discriminant analysis (QDA)

QDA is a variant of linear discriminant analysis (LDA) that allows for non-linear separation of data. It is a generative model. It assumes that each class follow a Gaussian distribution. It estimates an individual covariance matrix for every class of observations. It is useful if there is prior knowledge that individual classes exhibit distinct covariance.

3 Research methodology

This section describes the methodology followed in this work for conducting a comprehensive comparison of machine learning classifiers in detecting epilepsy seizures. It explains different phases of the proposed methodology for extracting the time-domain features and spectral features from recording and converting them into a form compatible with the processing of machine learning classifiers. It also describes the dataset used and identifies the most commonly used performance metrics in comparing machine learning classifiers.

3.1 Methodology

In this work, we followed the methodology depicted in Figure 1, which includes five stages, namely, data collection, feature extraction, pre-processing, classification and performance analysis for conducting a comprehensive comparison of machine learning classifiers. The details are described below.

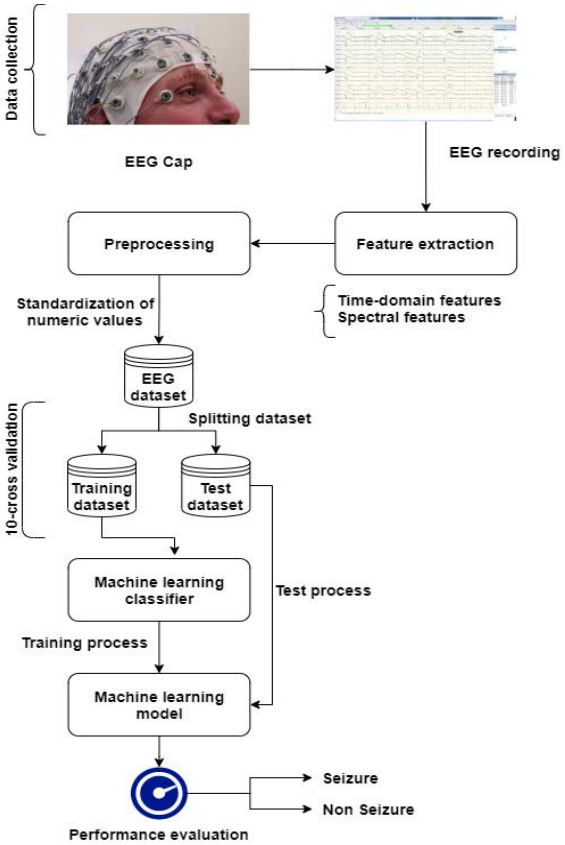


Fig. 1 The proposed framework

3.1.1 Data collection

The experimental data is collected using an EEG cap and other equipment in the form of EEGs. The details of the real-time dataset used in this work are provided in Section 3.2. The collected data is further processed to extract the relevant features in this work.

3.1.2 Feature extraction

In this stage, data signals from EEGs are processed to extract relevant features and arranged in rows and columns. We attempted to keep one intrinsic mode functions (IMFs) and further analyzed it by Hilbert transform to obtain different features.

In this work, we extracted time-domain and spectral features. We extracted features using one mode of EEG signals that is considered as frequency modulated and amplitude modulated signals as features of original EEG signal. The extracted features represent the properties of the spectrum of the signal modes [3]. We extracted nine spectral features and two time-domain features in the set of experiments as described in Table 1.

Table 1 Features extracted from EEG signals

Sr No	Feature	Description
1	Spectral power	The power spectral density (power spectrum) reflects the frequency content of the signal or the distribution of signal power over the frequency
2	Spectral entropy	Spectral entropy (SE) is a measure of signal irregularity, which sums the normalized signal spectral power
3	Spectral peak	EEG power is typically split up into bands that correspond to different spectral peaks related to behavior or cognitive state.
4	Frequency	Frequency associated with spectral peak
5	spectral centroid	Spectral centroid (SC) measures the shape of the spectrum of EEG signals. It is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes.
6	AM bandwidth	Bandwidth parameters
7	FM bandwidth	Bandwidth parameters
8	Hjorth mobility	Mean frequency of the signal and proportional to the variance of its spectrum
9	Hjorth Complexity	estimate of the signals' bandwidth
10	Skewness	Signal distribution's asymmetry
11	Kurtosis	Tails of the distribution yielded by the signal

3.1.3 Pre-processing

It has been observed that most machine learning methods report better performance when input values are preprocessed to a uniform scale. Normalization and standardization are the most commonly used methods for scaling numeric data to a standard range in the pre-processing stage. The normalization process scales numeric values to a range of 0 to 1. In contrast, the standardization process scales each numeric value separately by subtracting the mean and dividing by standard deviation to shift the distribution with a mean of 0 and a standard deviation of 1.

In this work, we use the standardization process to convert extracted features to a uniform scale using the following equations.

$$X\_standardized = \frac{X - mean}{standard\_deviation} \tag{1}$$

3.1.4 Classification

Classification of epilepsy seizure dataset require training of machine learning classifier. The pre-processed data is divided into training data set and test data set. We used a 10-fold cross-validation strategy to train and test the machine learning classifiers in this work. In the 10-fold cross-validation process, the data set is divided into ten parts. Nine parts are used as training data set, and one part of the dataset is used to evaluate the performance of machine learning classifiers.

We evaluated the ten most commonly used supervised machine learning methods in this work: KNN, Linear SVM, RBF SVM, GPC, DT, RF, NN-MLP, AdaBoost, NB, and QDA classifiers.

Each supervised machine learning algorithm is executed for ten independent experiments, and mean values of results are recorded in terms of identified performance metrics.

3.1.5 Performance evaluation

For evaluating the performance of machine learning classifiers and conducting a comprehensive comparison, the performance of each classifier is measured in terms of four metrics, accuracy, sensitivity, specificity and area under ROC curve.

These values are computed from the confusion matrix representing the values of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) defined as below [7].

- True positives (TP): Cases predicted as seizures that are seizures.
- True negatives (TN): Cases predicted as non-seizure that are non-seizure.
- False positives (FP): Cases predicted as seizures that are non-seizures.
- False negatives (FN): Cases predicted as non-seizures that are seizures.

These performance metrics can be computed as per the following equations.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{3}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{4}$$

3.2 Dataset

In this work, we use EEG data set provided by Neurology and Sleep Centre, New Delhi, for evaluating supervised machine learning classifiers. This data set contains segmented EEG recordings of 10 epilepsy patients collected at Neurology and Sleep Centre, New Delhi [3][22]. This data set is collected using the Grass Telefactor Comet AS40 amplification system at 200 Hz, and the gold plated scalp EEG electrodes placed following the international 10-20 electrode placement system. Further, the signals have been processed by a band-pass filter with cut-off frequencies of 0.5 Hz and 70 Hz.

These signals are classified into three categories, namely, preictal, interictal and ictal by expert physicians. There are 50 single-channel recordings in each class of data set. Each recording consists of 1024 samples with a duration of 5.12 s. In this work, we classify the dataset samples into three classes, namely, preictal, interictal and ictal and presented the recorded results.

3.3 Experimental setup

The machine learning algorithms used in this work are implemented in Python language using Scikit library for machine learning methods. We conducted experiments on a machine Intel Core I3-2330M CPU @ 2.20 GHz, 4 GB RAM, and 1TB HDD. We performed classification of the NSC-ND dataset into three classes, ictal, interictal and preictal. Hyper-parameters of the selected machine learning classifiers in this work are presented in Table 2.

Table 2 Hyper-parameters of machine learning classifiers

Classifier	Hypermeter values
KNN	n_neighbors =3 leaf_size =30 p(Power parameter for the Minkowski metric) =2 metric =’minkowski’
SVC (linear)	kernel=linear C=0.025 Degree =3 gamma =scale
SVC (RBF)	kernel=RBF C=1 Degree =3 gamma =2
GPC	kernel = 1.0 * RBF(1.0) optimizer =’fmin_l_bfgs_b’ max_iter_predict=100
DT	criterion=’gini’ max_depth=10 splitter=’best’
RF	Max_depth=5, n_estimators=10 max_features=1 criterion=gini
MLP	alpha=1 max_iter=200 hidden_layer_sizes=100 activation=relu solver=adam learning_rate=constant
AdaBoost	Base_estimator=DecisionTree n_estimators=50 learning_rate=1.0 algorithm=’SAMME.R’
NB	Var_smoothing=1e-09
QDA	Reg_param=0.0, store_covariance=False, tol=0.0001

4 Experimental results and analysis

We conducted ten independent sets of experiments for each using supervised machine learning methods. We employed KNN, Linear SVM, RBF SVM, GPC, DT, RF, NN-MLP, AdaBoost, NB, and QDA classifiers for classifying NSD-ND dataset in three classes. The performance of each classifier is recorded in terms of accuracy, sensitivity, specificity and AUC for 10-independent experiments.

Figures 2 to 5 present the box plots of experimental results in terms of accuracy, sensitivity, specificity, and area under ROC curve for ten independent experiments using a 10-fold cross-validation strategy.

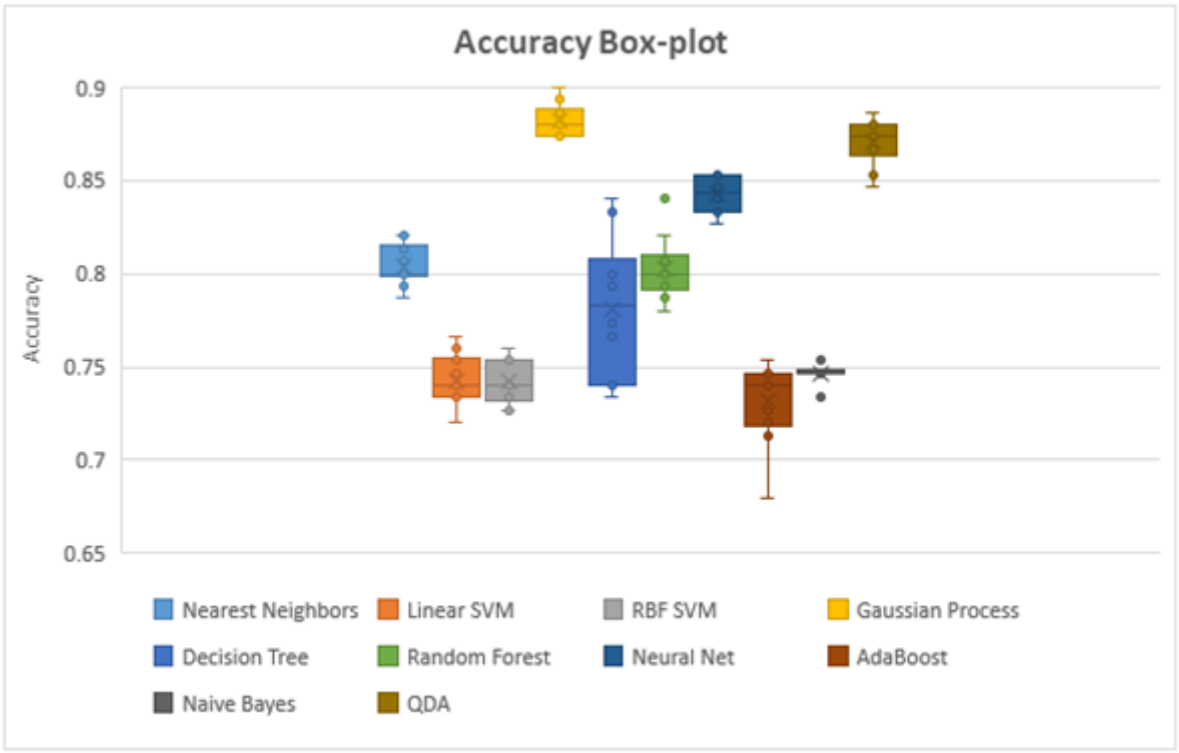


Fig. 2 Box plot of accuracy for ten independent experiments

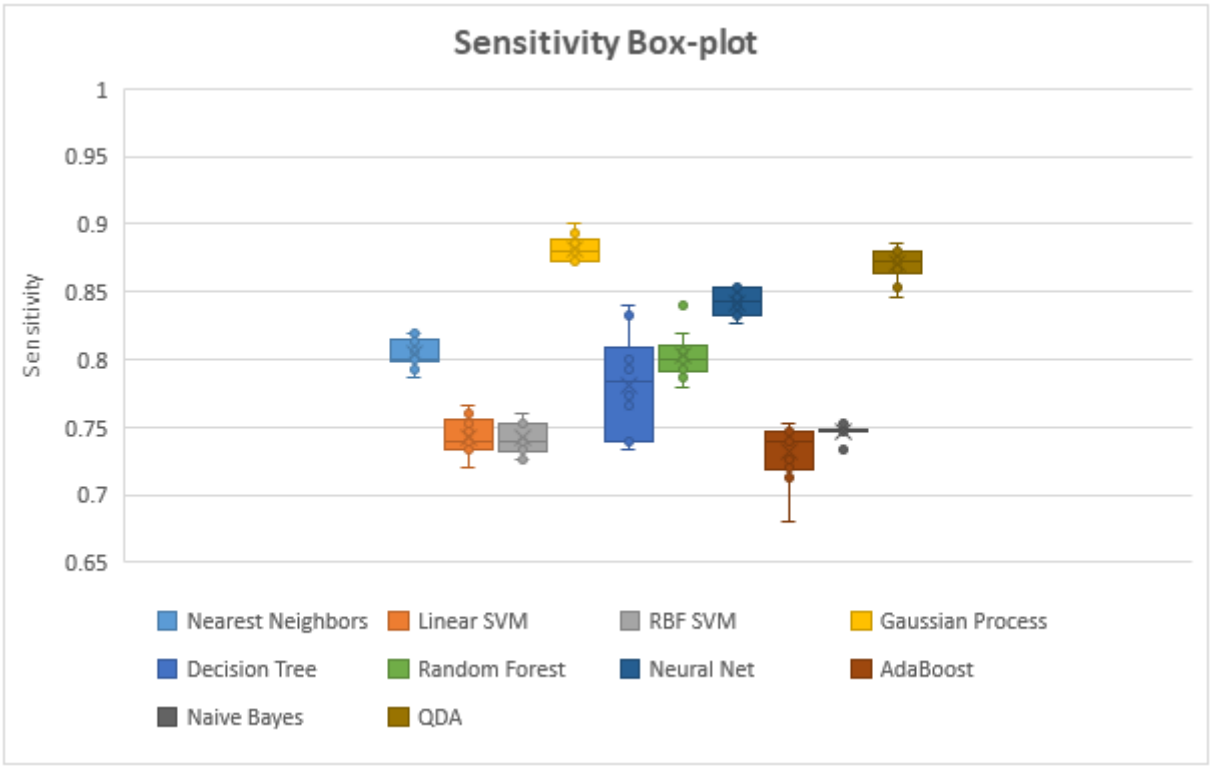


Fig. 3 Box plot of sensitivity for ten independent experiments

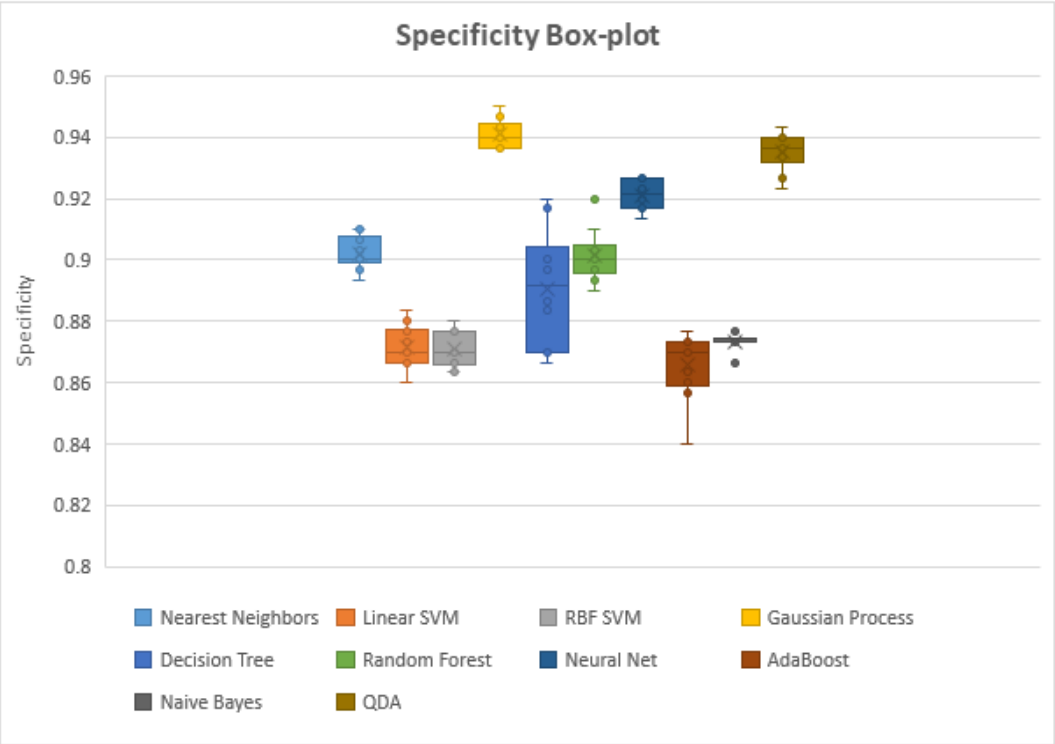


Fig. 4 Box plot of specificity for ten independent experiments

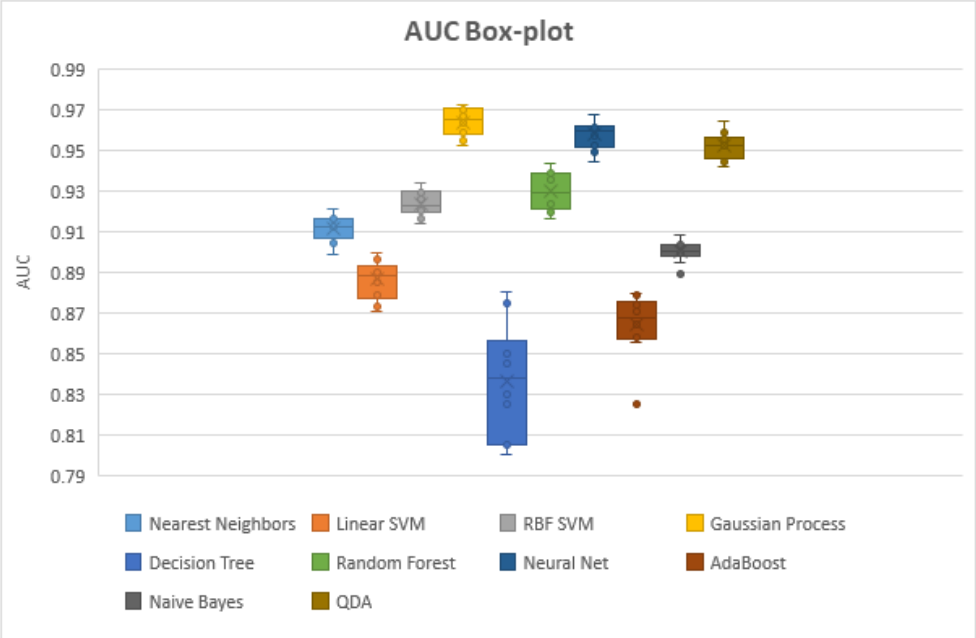


Fig. 5 Box plot of AUC for ten independent experiments

It can be observed from Figures 2 to 5 that most of the classifiers exhibit stable performance in detecting the accuracy of epilepsy seizure detection. However, decision tree classifier reported specific varying results in comparison to other classifiers. Similar performance of decision tree classifiers has also been observed for sensitivity, specificity and AUC as presented in Figures 2 to 5.

We calculated the mean and standard deviation of the recorded metrics in 10 experiments and presented in Table 3 and Figure 6, considering ictal as normal/positive class. Results are presented in the format of (mean ± standard deviation) for different machine learning classifiers.

It can be seen from Table 3 that GPC classifier and QDA provided the best and comparable results of approximately 87% accuracy among these ten classifiers in terms of accuracy. However, linear SVM, RBF SVM, adaBoost and NB algorithm reported comparable performance of approximately 73% accuracy. In terms of sensitivity metric, similar behaviour of Gaussian Process classifier and QDA classifiers have been observed by reporting 88% approximately value for sensitivity for classifying benchmark data set.

In terms of area under ROC, GPC, NN-MLP, and QDA classify as reported equivalent performance by covering an area of approximately 96%.

In terms of area under ROC, GPC, NN-MLP, and QDA classify as reported equivalent performance by covering an area of approximately 96%. In terms of specificity, QDA classified reported performance of approximately 95% with a standard deviation of 0.0060. It is followed by a Gaussian Process classifier bi reporting specificity of 94% with a standard deviation of 0.0042. Linear SVM, RBF SVM, and NB classifiers reported approximately 87% of specificity each.

The mean performance of the three best performing classifiers, GPC, NN-MLP and QDA classifiers, in terms of accuracy, is 88.20%, 84% and 87%, respectively. High values of accuracy, sensitivity, specificity and AUC reported by Gaussian Process classifier demonstrates its superior class separability over other classifiers.

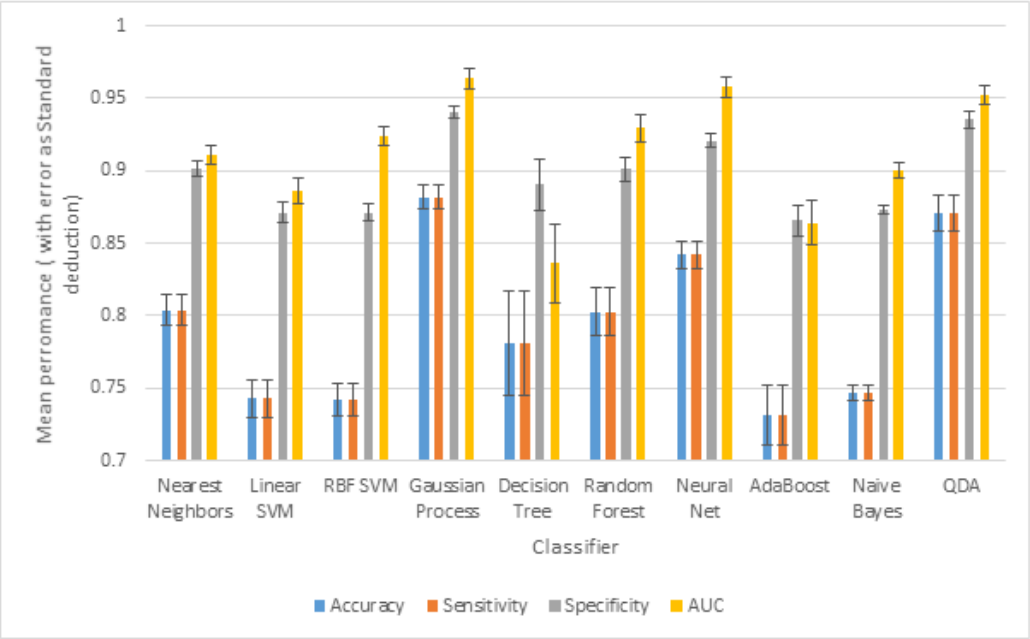


Fig. 6 Comparative analysis of mean performance (error bar as standard deviation) of machine learning classifiers

5 Conclusion

In this comparative empirical analysis of machine learning classifiers for detecting epilepsy seizure, we performed a set of comprehensive experiments in a controlled environment using supervised machine learning classifiers and EEG real-time data collected by Neurology and Sleep Centre, New Delhi. EEG signals are processed to extract time-domain and spectral features. The extracted features are pre-processed by standardizing numeric values to a uniform scale for use with machine learning classifiers. Experiments are conducted by training machine learning classifiers and testing them using a 10-fold cross-validation strategy. Ten independent experiments have been conducted for each machine learning classifiers in this work. Results are presented by computing mean and standard deviation for analysis. In addition, the mean performance of machine learning classifiers is compared using the most common performance metrics of accuracy, sensitivity, specificity and area under ROC curve.

It can be concluded from the reporting results that most classifiers have reported good performance in detecting epilepsy seizures using the benchmark NSC-ND dataset. GPC classifier has reported the best performance among machine learning classifiers in terms of the identified benchmark datasets. It has detected epilepsy seizure up to 88% of accuracy. It is followed by QDA classified by providing compare table accuracy of 87% in detecting epilepsy seizures. SVM classifier with linear and RBF kernel has reported an accuracy of 74% approximately in detecting epilepsy seizures.

We conducted these experiments without decomposing EEG signals into multiple modes. In our future work, we plan to explore signal decomposition methods and feature selection techniques to improve the accuracy of epilepsy seizure detection. These experiments and their results provide promising directions for fellow researchers in applying machine learning classifiers for detecting epilepsy seizures.

Conflict of interest disclosure

The authors declare that there is no conflict of interest regarding the publication of this paper.

Table 3 Mean performance of machine learning classifiers (with standard deviation)

Metric / Classifier	KNN	Linear SVM	RBF SVM	GPC	DT	RF	NN-MLP	AdaBoost	NB	QDA
Accuracy	0.8040 ± 0.0104	0.7427 ± 0.0134	0.7420 ± 0.0116	0.8820 ± 0.0085	0.7813 ± 0.0358	0.8027 ± 0.0164	0.8420 ± 0.0095	0.7313 ± 0.0211	0.7467 ± 0.0052	0.8707 ± 0.0120
Sensitivity	0.8040 ± 0.0104	0.7427 ± 0.0134	0.7420 ± 0.0116	0.8820 ± 0.0085	0.7813 ± 0.0358	0.8027 ± 0.0164	0.8420 ± 0.0095	0.7313 ± 0.0211	0.7467 ± 0.0052	0.8707 ± 0.0120
Specificity	0.9020 ± 0.0052	0.8713 ± 0.0067	0.8710 ± 0.0058	0.9410 ± 0.0042	0.8907 ± 0.0179	0.9013 ± 0.0082	0.9210 ± 0.0047	0.8657 ± 0.0105	0.8733 ± 0.0026	0.9353 ± 0.0060
AUC	0.9111 ± 0.0063	0.8863 ± 0.0090	0.9238 ± 0.0061	0.9638 ± 0.0068	0.8360 ± 0.0268	0.9295 ± 0.0095	0.9577 ± 0.0066	0.8643 ± 0.0151	0.9002 ± 0.0049	0.9522 ± 0.0062

REFERENCES

1. Alarcón, G., Valentín, A.: Introduction to epilepsy. Cambridge University Press (2012)
2. Bishop, C.M., et al.: Neural networks for pattern recognition. Oxford university press (1995)
3. Carvalho, V.R., Moraes, M.F., Braga, A.P., Mendes, E.M.: Evaluating five different adaptive decomposition methods for eeg signal seizure detection and classification. Biomedical Signal Processing and Control 62, 102073 (2020)
4. Chaudhary, U.J., Duncan, J.S., Lemieux, L.: A dialogue with historical concepts of epilepsy from the babylonians to hughlings jackson: persistent beliefs. Epilepsy & Behavior 21(2), 109–114 (2011)
5. Ian, H.W., Eibe, F.: Data mining: Practical machine learning tools and techniques (2005)
6. Kaur, R., Sachdeva, M., Kumar, G.: Study and comparison of feature selection approaches for intrusion detection. Int. J. Comput. Appl 7, 6 (2016)
7. Kumar, G.: Evaluation metrics for intrusion detection systems-a study. Evaluation 2(11), 11–7 (2014)
8. Kumar, G., Kumar, K.: Ai based supervised classifiers: an analysis for intrusion detection. In: Proceedings of the International Conference on Advances in Computing and Artificial Intelligence, pp. 170–174 (2011)
9. Kumar, G., Kumar, K.: An information theoretic approach for feature selection. Security and Communication Networks 5(2), 178–185 (2012)
10. Kumar, G., Kumar, K., Sachdeva, M.: The use of artificial intelligence based techniques for intrusion detection: a review. Artificial Intelligence Review 34(4), 369–387 (2010)
11. Larose, D.T., Larose, C.D.: Discovering knowledge in data: an introduction to data mining, vol. 4. John Wiley & Sons (2014)
12. Peng, H., Lei, C., Zheng, S., Zhao, C., Wu, C., Sun, J., Hu, B.: Automatic epileptic seizure detection via stein kernel-based sparse representation. Computers in Biology and Medicine 132, 104338 (2021)
13. Peng, H., Li, C., Chao, J., Wang, T., Zhao, C., Huo, X., Hu, B.: A novel automatic classification detection for epileptic seizure based on dictionary learning and sparse representation. Neurocomputing (2019)
14. Qu, H., Gotman, J.: Improvement in seizure detection performance by automatic adaptation to the eeg of each patient. Electroencephalography and clinical Neurophysiology 86(2), 79–87 (1993)
15. Reynolds, E.H.: Milestones in epilepsy. Epilepsia 50(3), 338–342 (2009)
16. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics 21(3), 660–674 (1991)
17. Sazgar, M., Young, M.G.: Absolute epilepsy and EEG rotation review. Springer (2019)
18. Schachter, S.C., Shafer, P., Sirven, J.: What causes epilepsy and seizures. Epilepsy Foundation (2013)
19. Shafer, P.O., Sirven, J.I.: Epilepsy statistics. Epilepsy Foundation (2014)
20. Sheena, K.K., Kumar, G.: Analysis of feature selection techniques: A data mining approach. In: IJCA Proceedings on International Conference on Advances in Emerging Technology, pp. 17–21 (2016)
21. Siddiqui, M.K., Morales-Menendez, R., Huang, X., Hussain, N.: A review of epileptic seizure detection using machine learning classifiers. Brain informatics 7, 1–18 (2020)
22. Swami, P., Gandhi, T.K., Panigrahi, B.K., Tripathi, M., Anand, S.: A novel robust diagnostic model to detect seizures in electroencephalography. Expert Systems with Applications 56, 116–130 (2016)
23. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks 10(5), 988–999 (1999)

A REVIEW OF TECHNIQUES AND APPLICATIONS OF SOCIAL MEDIA SENTIMENT ANALYSIS

Pritpal Kaur[#1], Dr. Himanshu Aggarwal[#2], Dr. Harmandeep Singh[#3]
[#]*Computer Science and Engineering Department, Punjabi University*
[1] kaurpritpal94@gmail.com
[2] himashu@pbi.ac.in
[3] harmanjhajj@yahoo.co.in

ABSTRACT— During the last decade, the world has seen immense growth in digitalization. From large scale businesses to whether its market, education, media, etc. every field is getting digitalized. Owing to this, use of social media networking sites has also been intensified. The abundant and prodigious data available on social media is a great extensive source to substantially understand the human behaviour. It paves the way for the researchers to strengthen a new field in NLP called social media analysis or social media sentiment analysis. Sentiment analysis plays a part and parcel role in the internet era due to its extensive range of applications and easy availability of data on social media. A lot of work has been done in the field. This paper is aimed to provide an insight about the various works done in the field of sentiment analysis. Through paper the various applications and techniques of social media analysis are discussed. The most popular social media networks which are being explored by the researchers are also discussed using the study.

KEYWORDS— Sentiment analysis, social media analysis, data mining, machine learning, opinion mining

INTRODUCTION

Social-media is very popular and effective way of expressing one's views or opinions on a topic. The studies suggest that in 2020 about 3.6 billion people are using social media. And in 2018, India alone had about 326 million social media users and it is expected to reach 448 million till 2023. Due to widespread popularity the social media, researchers have been able to capture real time sentiments from large number of individuals [2]. Through social media, users are able to display their emotions and sentiments publicly in the form of electronic media [4]. The social media provide abundance of data but the data available is very unstructured and raw. Designing opinions from this unstructured data is a tedious work indeed. Data mining is the rescue in this case. With the extensive research in data mining field number of algorithms and tools are there to clean unstructured social media data.

Social media provide not only data but a social structure to understand the social relations and human behavior patterns. The individuals on social media are also known as social atoms and the communities on social media are known as social molecules. Analyzing and creating actionable patterns from social media data is called social media analysis which is a knowledge discovery method. Sentiment analysis is performed mainly on three different levels which are document level, sentence level and aspect level [14]. The document level analysis takes the one review or micro-blog as one unit and calculates the sentiment for the whole document. While in sentence level analysis all the sentences in the micro- blog or review are analyzed individually and aspect level analysis is used to study the different aspects of a single piece of text [24]. Most of the studies focus on the document level analysis of the data. Sentiment analysis is useful not only for industries but can help the customers in their shopping decisions [14].

The social networking sites like twitter provide not only the message or tweets but also information like timestamps, number of replies or re-tweets, geospatial information which is very useful to understand the individuals and their opinions about various topics. The micro-blogs analysis can be Timestamps provided with social media data which gives the information that when the data is created. This timestamp has enabled the researchers to study the change of emotions or opinions of people about a certain topic over a time period. This method is known as time series analysis which is a branch of sentiment analysis [18].

Sentiment analysis plays major role in the internet era due to its extensive range of applications and easy availability of data on social media. Inspiration behind sentiment analysis is that it provides people's opinion about the product, which helps to improve the product quality [3]. The social media provide us with variety of data like Text, images, videos, etc but researchers mainly focus on the text data. Sentiment analysis identifies opinions and sentiments expressed in text [2]. For doing sentiment analysis of data there are mainly two methods or techniques, Lexicon approach and machine learning. The lexicon based approach follows tokenization followed by comparing the words collected with the pre identified emotional words in the database. The overall score can be computed as positive, negative or neutral. On the other hand machine learning based approach uses a train and test data set [11]. Both these approaches of sentiment analysis has its own pros or cons. Sentiment analysis is divided into three different levels which are sentence level, document level and feature level [26]. A comparison of social media monitoring tools conducted in October 2014 by Ideya Ltd3 shows that there are at least 245 tools for social media monitoring available, of which 197 are paid, with the remainder free or using a free model [27]. This paper focuses to provide an overview of the sentiment analysis applications, and brief review about techniques used for the sentiment analysis by the researchers during latest times. Research has been undertaken to give reader insight into latest Methods and applications of the field. The study is undertaken to understand the various techniques used by the researchers in the field of sentiment analysis of social media. The study focuses to draw a comparison between many

techniques used in the field of sentiment analysis. With help of this study various application fields of sentiment analysis are explored. This will be helpful to understand the importance of sentiment analysis in the real world.

METHODOLOGY

The main goal of this paper is to provide the understanding of techniques used in sentiment analysis of social media data with respect to the application context of the various researches. The study shows various ways and applications in which social media data is utilized. The review of various techniques and tools used during social media sentiment analysis and data sources used is given. The basic framework or methodology followed for a review paper is as follows:

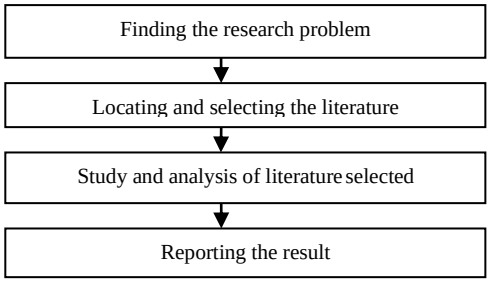


Fig. 1: steps carried out to complete the review

The clear understanding of the objectives of the review to be undertaken is very important. The main objective of this review is to provide understanding of sentiment analysis of social media data. The techniques and applications of the sentiment analysis

The selection of literature to be reviewed is crucial job for the literature review. Total 25 papers are selected from Science Direct and IEEE Xplore database to meet the research goals for this review paper.

SENTIMENT ANALYSIS

N. methods

After reading the selected literature we have come across two techniques which are used for sentiment analysis of data Dictionary based approach and machine learning approach. There are many works focus on the Emotional Dictionary and rules. Machine learning is the most popular method in Sentiment Analysis, for example, some researchers combining multiple machine learning methods with different feature selection methods [19]. Out of 25 reviewed papers 6 papers implemented the dictionary based approach and 17 used machine learning methods and 2 used the combination of both methods for analysis of social media data.

1) Lexicon Approach:

In lexicon-based approach, sentiment classification is performed using a sentiment dictionary, a collection of lexical units accompanied with their sentiment orientation, where lexical units may be words or phrases and Sentiment Analysis Techniques and sentiment orientation may be coarse classes (e.g., positive, negative), fine-grained classes (e.g., varying from very positive to very negative) or real values in an interval such as [−1, +1] [28].

While using dictionary based approach SentiWord net is a standard dictionary used by most researchers today for sentiment analysis [1]. Other than this [2] uses VADER's dictionary, VADER computes sentiment for each word and generates compound scores for the sentence by summing the sentiment score of each word. A sentiment score ranges from -1 (most extreme negative) and +1 (most extreme positive). In [3] Rapid-Miner AYLIEN tool is used which is a text analysis extension. SentiStrength is a lexicon-based sentiment analysis tool that classifies tweets on an 11- point scale ranging from −5 (negative) to 0 (neutral) and 5 (positive) [10]. The key feature of lexicon based approach is that it does not require any training for algorithm which is very crucial in supervised machine learning. It is an unsupervised method. It is good and gives fair results, but in case of social media data the unstructured and unconventional language and slangs makes the lexicon approach a not so reliable method.

2) Machine learning:

Currently machine learning approach is most used method by the researchers for sentiment analysis of social media. The Machine learning based approach uses a train and test data set. The classifier can be trained with the training data set (e.g. classified tweets) and the test data will be given as input to it. It will give the desired result such as positive or negative [11]. There are number of machine learning algorithms available which can be used for Data analysis and sentiment analysis of social media. It is observed that SVM, ANN and Naïve bayes are the most used algorithms.

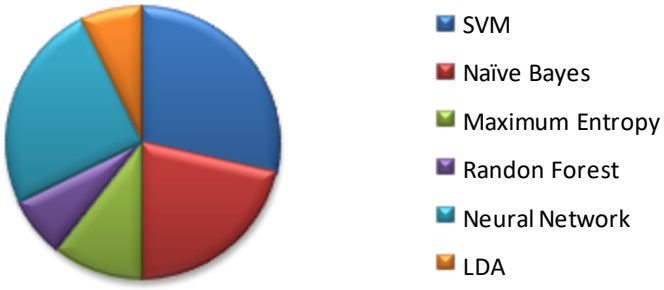


Fig. 2: most used machine learning algorithms in studies reviewed

2.1) Supervised machine learning:

The supervised machine learning algorithms predicts the results based on labeled data provided as training set. Supervised machine learning is most utilized method by the researchers for sentiment analysis. The various supervised machine learning algorithms are used by the researchers. Supervised learning is an important technique for solving classification problems [5].

SVM: SVM was first proposed by Cortes and Vapnik in 1995, it has showed many unique advantages in solving small sample, nonlinear and high dimensional pattern recognition problems [19]. Support vector machine analyzes the data, define the decision boundaries and uses the kernels for computation which are performed in input space [5]. SVM is discriminative method and is formally defined by differentiating the hyperplane. It takes as input the labeled training data and outputs the optimal hyperplane to classify the data [9]. SVM also supports both classification and regression. Which make it useful for statistical learning and it helps recognizing the factors precisely that needs to be taken into account to understand it successfully [5]. SVM multi-class classification is not very good at determining the quantity of sentiment change; it is still very good at determining the direction of sentiment change [18].

Naïve bayes: Naive bayes Classifier is the Bayesian classification learning method, which is the statistical method used for classification of opinions. It assumes a model to be probabilistic model, and it captures the uncertainty about the model. It is used to solve the diagnostics and predictive problems by the researchers. The algorithm is named after the scientist Thomas Bayes, who proposed the algorithm [9]. It has been used because of its simplicity in both during training and classifying stage [5]. In naive bayes classifiers, every feature is taken into account to determine which label should be assigned to a given input value [15]. A model called the tree augmented naive Bayes (TAN), allows each node to have at the most one parent node in addition to the class variable node. The resulting DAG is a tree, with $n - 1$ edges [17]. In paper [5] it is observed that naïve byes technique gives better result than the maximum entropy and SVM when being subjected to unigram model than using it alone.

Maximum Entropy: This classifier uses a model similar to naïve Bayes classifier, except it uses search techniques to find set of parameters to maximize the performance of the classifier rather than using probabilities to set model's parameters [15]. In Maximum entropy defied entropy is maximized on the conditional probability distribution. It even handles overlap feature and is same as logistic regression which finds distribution over classes [5].

Neural Networks: Neural Networks such as a convolutional neural network (CNNs) and recurrent neural system (RNNs) have practiced the better outcomes for social media sentiment analysis [7]. Recurrent Neural Networks (RNN) is popular model that gives promising and good results in many NLP tasks [19]. They work tremendously well on a large variety of problems, and are now widely used [19]. D-CNN is the improved version of CNN, which introduces another hyper-parameter to the receptive layer. D-CNN raises the overall network's performance by increasing the size of the receptive field to capture additional information by placing zeros in the filter components [7]. Bayesian network (BN) is a directed acyclic graph (DAG), where the nodes represent discrete random variables and the probabilistic relations amongst them are represented by the edges of the graph [17]. In paper [7] it is observed that the utilization of more hidden layers does not result in efficient improvement. In [29] authors have developed a new algorithm using the CNN and RNN which is named UCRNN (User attributed Convolutional and Recurrent Neural Network). The hybrid system uses the CNN for feature extraction from user attributes and RNN for word position information. The hybrid approach gives more accuracy and precision compared to other algorithms.

Table 2 shows the comparison of accuracy recorded by researchers for supervised learning methods. It is seen in the table that most of the models give fair results and above 70% accuracy.

The training data set can be increased to improve the feature vector related sentence identification process [5].

TABLE 1
COMPARISON OF PERFORMANCE OF VARIOUS SUPERVISED MODELS USED DURING STUDIES BY RESEARCHERS

Papers ▼ / **Accuracy recorded →**	**SVM**	**Maximum entropy**	**Neural Networks**	**Naïve bayes**
Geetika Gautam, Divakar yadav [5]	85.5	83.8	-	85.5
Muhammad Alam, Fazeel Abid, Cong Guangpei , L.V. Yunrong [7]	-	-	72.34	-
Priti Sharma, A.K. Sharma [9]	71.3	-	83.9	-
R. A. S. C. Jayasanka, M. D. T. Madhushani, E. R. Marcus, I. A. A. U. Aberathne and S. C. Premaratne [15]	-	77.99	-	78.06
Gonzalo A. Ruz, Pablo A. Henríquez, Aldo Mascareño [17]	81.2	-	76.4	74.2
Le T. Nguyen, Pang Wu, William Chan, Wei Peng, Ying Zhang [18]	74.93	-	-	-
Hongyang Xu, Hui Lu, Guowei Yang, Cong Zhang [19]	86	-	89	-

Table 2 shows the comparison of accuracy recorded by researchers for supervised learning methods. It is seen in the table that most of the models give fair results and above 70% accuracy.

The training data set can be increased to improve the feature vector related sentence identification process [5].

2.2) Unsupervised machine learning: In unsupervised machine learning the algorithms are trained using unlabeled data. An unsupervised machine learning technology, LDA uses the bag-of-words method to identify the information about topic hidden in document sets or corpora. In LDA, a document is an unordered set of words. A document may have many topics, and each word in the document is created by one of these topics. LDA can distribute the topic of each document in the document set in the form of probability distribution [16]. Unsupervised machine learning is mainly used for clustering. It is used for identifying hidden patterns from unlabeled data but results from unsupervised machine learning methods are not considered very reliable for classification problems.

TABLE 7
ADVANTAGES AND DISADVANTAGES OF DIFFERENT TECHNIQUES OF SENTIMENT ANALYSIS

	Advantages	**Disadvantages**
Dictionary approach	➢ Labeling of data is not required ➢ training of data not required ➢ easy to implement ➢ low computation time	➢ not reliable method
Supervised ML	➢ reliable method ➢ gives more accurate predictions ➢ Good for classification problems ➢ Variety of algorithms are available	➢ Labeled data required for training ➢ Computation time is more, as training of model takes time
Unsupervised ML	➢ Labeling of data is not required ➢ Good for hidden pattern discovery	➢ not reliable method ➢ Training of model is time consuming ➢ Not good for classification problems

O. Applications

There are number applications for which sentiment analysis of social media data can be done from health care, education to product marketing etc.

1) Marketing and Business:

In [10]study offers a better understanding of customers' opinions towards online retailing and provides insight into what customers are really thinking about by analyzing their opinions as expressed on Twitter. Stock price forecasting is very important in the planning of business activity, in [12] authors proposed a novel feature 'topic-sentiment' to improve the performance of stock market prediction. In [18] the author is comparing the sentiment changes over multiple topics (iPhone, Android and Blackberry). In [21] authors have implemented the emotional analysis for Chinese public opinion

texts about P2P network lending on Sina Weibo social networking platform. In [6] authors constructed a linear regression model for predicting box-office revenues of movies in advance of their release. The model can be generalized for predicting the revenues of products based on social media.

2) Politics:

In [13], Tweets were collected from the period of Jan 2019 to March 2019. Using that tweets, sentiment analysis was performed to analyse the opinion polarity of the people during general elections held in India. Sentiment analysis can also be used for improvement of education methods. The paper [23] presents a methodology, called IOM-NN (Iterative Opinion Mining using Neural Networks), for discovering the polarization of social media users during election campaigns characterized by the competition of political factions. The proposed model was tested using two case studies to analyse the polarization of a huge number of Twitter users during the 2018 Italian general election and the 2016 US presidential election. In [2] authors evaluate the feasibility of using the social media platform Twitter to monitor negative social discussions about Mexicans and Hispanics on Twitter during the 2016 United States presidential election, as this was a time when negativity towards this minority groups is observed the most.

3) Emergency or critical events:

Analysis of emotions during crisis is an important but a quite complicated task. Critical events are those when individually or socially people experience events which surpass the threshold. This creates a state of panic and mixed reactions from people. Sentiment analysis can be very helpful during the emergencies, which can be used to know the perspective of people and their states of minds during though times. For example in study [25] the authors focuses on finding emotional reactions of users during the COVID-19 outbreak by exploring the tweets. A random sample of 18,000 tweets is collected from twitter. Tweets are classified into positive and negative sentiments along with eight emotions like anger, anticipation, disgust, fear, joy, sadness, surprise, trust. The emotion of fear among the people was observed as the most dominating trait in tweets explored.

4) Health care and education:

The lexical and statistical analysis can act as a surveillance tool for health care data analysis [11]. In [22] the study utilized data obtained through convenience sampling from a Student Union Facebook forum in a research-intensive university in New Zealand. The central question that guided the discussion was: ‘‘do you think lectures should be recorded?’’ The results of this study indicate that students maintained a shared view that lecture recordings should be available to them.

5) Government policies:

The paper [8] proposes a sentimental analysis of twitter data related to the demonetization event in India. The aim was to analyse public opinion on the topic of demonetization after six months of the event. In [1] authors capture polarity of the sentiments captured from twitter data for Case Study of Digital India mission.

After considering all the papers it is observed that Business and marketing and politics are the hot topics of sentiment analysis. Out of 25 Papers 9 have utilized the research for marketing related tasks like product reviews, stock movement prediction, box-office collection prediction of movies, etc. On the other hand education and medical or health care are the least explored using sentiment analysis of social media.

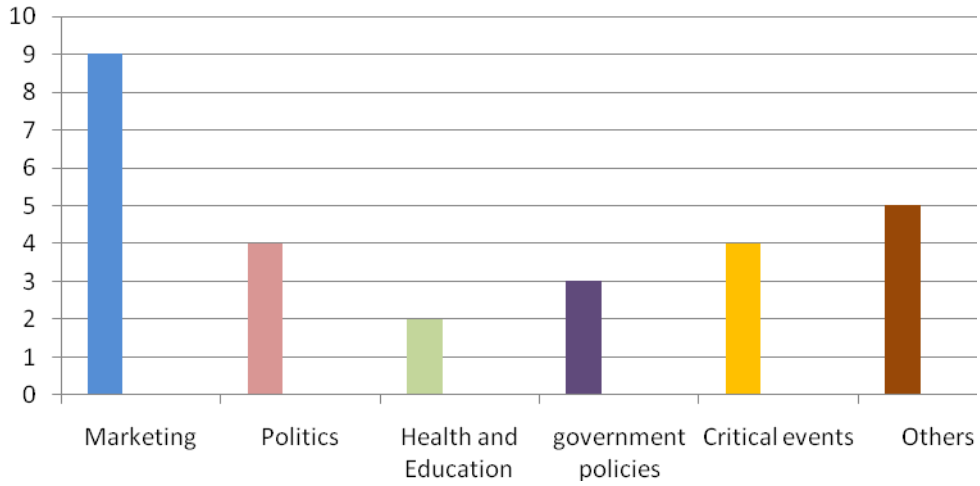


Fig. 3: The application areas explored by the researchers

P. Datasets

In the process of sentiment analysis dataset acquisition is the first step. Getting the right type of data for the job is an important thing. In social media sentiment analysis datasets are collected from social media platforms but there are number of social media platforms. Social information services or social media can be categorized into four types based on their application and usage: Content communities (Youtube, Instagram), Social networking (Facebook, LinkedIn), Blogs

(Reddit, Quora) and Micro-blogs (Twitter, Tumblr) [26]. When it comes to sentiment analysis researchers mainly focuses on text data. Blogs and micro-blogs are considered more reliable in this case. Twitter mining is very popular these days, as it provides the important information which is used and applied in various fields. It is one of the major research areas. By using the various public APIs various tweets can be collected and analyzed for research purpose. Through authenticated requests twitter APIs are established [13]. Other than twitter Facebook, yahoo finance message boeard, news blogs And Chinese Sino Weibo social networking site is used for the study.

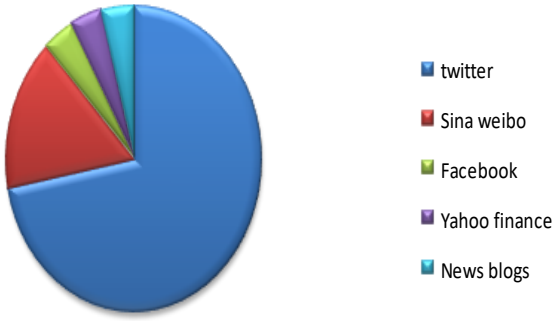


Fig. 4: the different sources used by the researchers for datasets collection

CONCLUSIONS

The aim of this study is to provide a literature review of social media sentiment analysis. This will provide an insight to the readers about the different techniques of sentiment analysis of social media analysis. There are main two techniques used for sentiment analysis, lexicon approach and machine learning approach. It is observed during the study that machine learning approach is most used by the researchers. In machine learning approach, supervised machine learning is widely used. It is also observed that SVM, Naïve-bayes and neural networks are some most preferred algorithm of supervised method and these algorithms are able to give fair and satisfactory results. In the lexicon based approach Sentiword net dictionary is widely used by the researchers. Selection of technique for sentiment analysis is mainly depends upon type of data being used. The performance of algorithm is also dependent on preprocessing and features selected from the data. When it comes to data sources it is observed that Twitter networking site is most trusted by the researchers. Because accessing the data of Twitter is easy with large number of users. The application area explored by the researchers during studies involves marketing and product reviews, politics, health care and education. The business and marketing field utilized the social media data to the maximum. It has also been observed that time series of data is analyzed by the researchers very much, which is a good method to analyze the sentiment change over time.

REFERENCES

[1] S. M. Prerna Mishra , Dr. Ranjana Rajnish, Dr.Pankaj Kumar, “Sentiment Analysis of Twitter Data:Case Study on Digital India”, International conference on Information Technology 2016

[2] Yulin Hswen, Qiuyuan Qin, David R. Williams, K. Viswanath, S.V. Subramanian , John S. Brownstein, “Online negative sentiment towards Mexicans and Hispanics and impact on mental well-being: A time-series analysis of social media data during the 2016 United States presidential election”, Heliyon 6 2020

[3] Vallikannu Ramanathan, T. Meyyappan, “Twitter Text Mining for Sentiment Analysis on People’s Feedback about Oman Tourism”, 4th MEC International Conference on Big Data and Smart City 2019

[4] Dilesh Tanna, Manasi Dudhane, Prof. Kiran Deshpande, Prof. Neha Deshmukh, Amrut Sardar, “Sentiment Analysis on Social Media for Emotion Classification”, International Conference on Intelligent Computing and Control Systems 2020

[5] Geetika Gautam, Divakar yadav, “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis” IEEE 2016

[6] Sitaram Asur, Bernardo A. Huberman “Predicting the Future With Social Media”, L.P 2020

[7] Muhammad Alam, Fazeel Abid, Cong Guangpei , L.V. Yunrong, “Social media sentiment analysis through parallel dilated convolutional neural network for smart city applications”, Computer Communication 154 2020

[8] Niharika Kumar,” Sentiment Analysis of Twitter Messages: Demonetization a Use Case”, International Conference on Computational Systems and Information Technology for Sustainable Solutions 2017

[9] Priti Sharma, A.K. Sharma, “Experimental investigation of automated system for twitter sentiment analysis to predict the public emotions using machine learning algorithms”, Materials Today: Proceedings 2020

[10] Noor Farizah Ibrahima, Xiaojun Wang, “Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media”, Computers in Human Behavior 96 2019

[11] R. Meena, Dr. V. Thulasi Bai, “Study on Machine learning based Social Media and Sentiment analysis for medical data applications”, Third International Conference on I-SMAC 2019

[12] Thien Hai Nguyena, Kiyoaki Shirai , Julien Velcin, “Sentiment analysis on social media for stock movement prediction”, Expert Systems With Applications 42 2015

[13] Ankita Sharma, Udayan Ghose, "Sentimental Analysis of Twitter Data with respect to General Elections in India", Procedia Computer Science 2020

[14] K.Rajendra Prasad, C.Raghavendra, Padakandla Vyshnav, "intelligent system for visualized data analytics a review", International Journal of Pure and Applied Mathematics 2017

[15] R. A. S. C. Jayasanka, M. D. T. Madhushani, E. R. Marcus, I. A. A. U. Aberathne and S. C. Premaratne, "Sentiment Analysis for Social Media", 2013

[16] Bangren Zhua , Xinqi Zhenga, Haiyan Liuc, Jiayang Li, Peipei Wang, "Analysis of spatiotemporal characteristics of big data on social media sentiment with COVID-19 epidemic topics", Chaos, Solitons and Fractals 140 2020

[17] Gonzalo A. Ruz, Pablo A. Henríquez, Aldo Mascareño, "Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers", Future Generation Computer Systems 106 2020

[18] Le T. Nguyen, Pang Wu, William Chan, Wei Peng, Ying Zhang," Predicting Collective Sentiment Dynamics from Time-series Social Media", wisdom 2012

[19] Hongyang Xu, Hui Lu, Guowei Yang, Cong Zhang "Sentiment Analysis of Chinese Version Using SVM & RNN", 6th International Conference on Information Engineering Proceedings 2017

[20] Firoj Fattulal Shahare, "Sentiment Analysis for the News Data Based on the social Media", International Conference on Intelligent Computing and Control Systems 2017

[21] Lei Li, Yabin Wu, Yuwei Zhang, And Tianyuan Zhao," Time+User Dual Attention Based Sentiment Prediction for Multiple Social Network Texts With Time Series", IEEE access 2019

[22] Larian M. Nkomo , Ifeanyi G. Ndukwe , And Ben Kei Daniel, "Social Network and Sentiment Analysis: Investigation of Students' Perspectives on Lecture Recording", IEEE access 2020

[23] Loris Belcastro, Riccardo Cantini, Fabrizio Marozzo, Domenico Talia And Paolo Trunfio, "Learning Political Polarization on Social Media Using Neural Networks", IEEE access 2020

[24] Mondher Bouazizi And Tomoaki Ohtsuki, "Multi-Class Sentiment Analysis in Twitter: What If Classification Is Not the Answer", IEEE access 2020

[25] Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, And Rakhi Batra , "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets", IEEE access 2020

[26] Zulfadzli Drus, Haliyana Khalid, " Sentiment Analysis in Social Media and Its Application: Systematic Literature Review", The Fifth Information Systems International Conference 2019

[27] Diana Maynard, Ian Roberts, Mark A. Greenwood, Dominic Rout, Kalina Bontcheva, "A framework for real-time semantic social media analysis", Web Semantics: Science, Services and Agents on the World Wide Web 2017

[28] Foteini S. Dolianiti, Dimitrios Iakovakis, Sofia B. Dias, Sofia Hadjileontiadou, José A. Diniz, and Leontios Hadjileontiadis, "Sentiment Analysis Techniques and Applications in Education: A Survey", Springer Nature Switzerland AG 2019

EPILEPTIC SEIZURE DETECTION IN EEG SIGNAL USING ADAPTIVE MODE DECOMPOSITION METHODS

Sandeep Singh[1,2*], Harjot Kaur[1]
[1]Department of Computer Science Engineering, Guru Nanak Dev University, Regional Campus, Gurdaspur, India
[2]Department of Computer Science Engineering, SGT University, Gurgaon, India
*E-mail id of corresponding author: er.ss1989@gmail.com

ABSTRACT- Epilepsy seizure detection plays a significant role in diagnosing the disease and can prevent significant injuries such as fractures and burns. Recently, many automatic epilepsy seizure detection methods have been developed based upon EEG signal analysis. Most methods involve the decomposition of EEG signals in different modes, extracting features from decomposed modes and classifying signals using machine learning methods. Many adaptive mode decomposition methods have recently been proposed, addressing the limitations of conventional Fourier-based methods when dealing with non-linear and non-stationary data. Empirical mode decomposition (EMD), empirical wavelet transform (EWT) and variational mode decomposition (VMD) are the most commonly used adaptive mode decomposition methods.

This work analyses the effectiveness of adaptive mode decomposition methods for classifying epilepsy seizures by analyzing EEG signals. EEGs signals have been decomposed into different modes in the form of IMFs using the adaptive decomposition method, EMD, EWT and VMD. Various spectral and time-domain features are extracted from decomposed modes of EEG signals. A neural network, an efficient and effective classifier, is used for classifying epilepsy seizures based upon the extracted spectral and time-domain features. The performance of the neural network classifier is analyzed for different adaptive mode decomposition methods with different segmented modes in terms of accuracy, sensitivity, specificity and area under the ROC curve (AUC). The reporting results demonstrate the effectiveness of the adaptive mode decomposition methods in successfully classifying epilepsy seizures and normal EEG signal. It can be concluded that EMD method provides its best results in 2-mode decomposition, whereas VMD provide the best results in the 8-mode decomposition of EEG signals. EMD method results in classification accuracy of neural network up to 88%, whereas VMD method reported an accuracy of 87% in classifying epilepsy seizure EEG signal from standard EEG signals. The reporting results demonstrate using the adaptive mode decomposition methods as a promising direction for developing an automatic epilepsy seizure detection system based upon spectral and time-domain features.

Keywords- Electroencephalogram (EEG) · Epilepsy · Machine learning · Seizure detection · Empirical mode decomposition · Empirical wavelet transform · Intrinsic mode functions · Variational mode decomposition.

1 Introduction

Epilepsy is a neurological disorder and chronic disease that impacts the consciousness and behaviour of the subject. The leading cause behind epilepsy seizure dysfunction of neural activities of the brain. It has been observed that more than 50 million peoples have epilepsy [16]. However, the origin of epilepsy seizures is still unanswered. Therefore, several methods have been developed for detecting epilepsy seizure for its diagnosis [4][14]. Electroencephalogram (EEG) that records brain activities in electrical signals has been widely used to detect epilepsy seizures. It is used as a clinical tool for assessing the neurological status of the suspected subjects [15]. It contains the recordings of the brain's electrical activities produced by nerves.

Traditionally, EEGs are analyzed by neurological experts for detecting epilepsy seizures. However, manual inspection of many multi-channel EEG signals is very cumbersome, time-consuming, error-prone. Therefore, an automatic system for detecting epilepsy seizures by analyzing EEG signals can help neurological experts greatly. Several methods have been developed for detecting epilepsy events automatically from EEGs. Most methods involve extracting features and pattern classification. Conventionally, spike amplitude and frequency were used to extract time-domain features for detecting epilepsy seizures. However, most conventional methods are unable to extract time-domain features from EEGs in the presence of noise. The methods like Fourier transforms are generally not recommended for analyzing frequency component due to non-stationary components present in EEGs [8]. Some methods have been proposed based on time-frequency methods like Lyapunov exponents [9].

Recently, adaptive methods have been proposed for decomposing signals, such as empirical mode decomposition (EMD) and its variants [8], variational mode decomposition (VMD). These adaptive methods are effective methods for complicated signal analysis. These methods are data-driven and posterior. Adaptive methods have been successfully implemented in different domains, such as fault diagnosis [5].

This work analyses the effectiveness of adaptive mode decomposition methods for classifying epilepsy seizures by analyzing EEG signals of a benchmark real-time data set collected by Neurology and Sleep Centre, New Delhi (NSC_ND). EEGs signals have been decomposed into different modes in the form of IMFs using the adaptive decomposition method, EMD, EWT and VMD. Various spectral and time-domain features are extracted from decomposed modes of EEG signals.

A neural network, an efficient and effective classifier, is used for classifying epilepsy seizures based upon the extracted spectral and time-domain features. The performance of the neural network classifier is analyzed for different adaptive mode decomposition methods with different segmented modes in terms of accuracy, sensitivity, specificity and area under the ROC curve (AUC). We conducted ten independent experiments for EMD, EWT and VMD methods, each with three modes using neural network classifier and computed their performance using a 10-fold cross-validation strategy. The mean performance of neural network classifier is compared to analyze the performance of adaptive mode decomposition methods in detecting epilepsy seizures. Major contributions of this work are as below.

1. Processing and decomposing EEG signals for extracting spectral and time-domain features using adaptive mode decomposition methods.
2. Extracting spectral and time-domain features from decomposed ECG signals.
3. Pre-processing extracted spectral and time-domain features for processing with neural network classifier.
4. Empirical comparison of adaptive mode decomposition methods using neural network to demonstrate their performance in detecting epilepsy seizures.

Rest of the paper is organized as follows. Section 2 describes the adaptive mode decomposition methods used in this study. Section 3 explains the methodology followed in this work by describing different stages in detecting epilepsy seizure, benchmark dataset, and experimental setup for conducting comprehensive set of experiments. Section 4 presents the experimental results obtained in this work and provides an analysis of the reporting results. Finally, the paper is concluded in future in Section 5.

2 Adaptive decomposition methods

Adaptive mode decomposition methods are practical for analyzing complicated and multi-channel signals such as the brain's EEGs. These methods are data-driven and posterior methods for decomposing the multi-channel EEG signals [5]. These methods do not require any prior knowledge about the signals and impose any condition on signal representation in different domains such as time and frequency. Consequently, these methods can extract oscillation modes of mono-component nature representing oscillation properties from an arbitrary signal. These methods can represent oscillation properties as a superposition of several mono components. The ability to decompose the signal into mono-component help to estimate the instantaneous frequency and amplitude accurately. These parameters lead to frequency decomposition and time variability of the signal. Adaptive mode decomposition methods are very adaptive to complicated and morphological contents that enable their suitability for harmonic, impulsive and modulated components. These methods can extract dynamic features of the system by analyzing the amplitude and frequency of the resultant mono-component of the signal. Several adaptive mode decomposition methods have been proposed, such as EMD and its variants, EWT and VMD. In this work, we focus on EMD, EWT, and VMD to decompose EEG signals for detecting epilepsy seizure. These methods are described briefly in the following subsections.

2.1 Empirical mode decomposition (EMD)

EMD is an adaptive and data dependent method for decomposing signal without any stationary and linearity condition. It decomposers the non-linear and non-stationary signal into a sum of intrinsic mode functions (IMFs) satisfying conditions of same or with one difference between the number of extrema and zero crossings; and define the mean value envelope using local maxima and envelope of local minima of zero. EMD method is very effective in mono-component decomposition. However, it suffers from the limitation of lack of mathematical Foundation [5]. It suspects to mode mixing under singularity. It is unstable under noise interference and overfitting and underfitting problem because of cubic spline interpolation. The details of EMD of Method can be further explored in [6][5].

2.2 Empirical wavelet transform (EWT)

Gilles [7] developed an empirical wavelet transform method with a solid mathematical foundation. This method involves building an adaptive wavelet that can extract amplitude modulated and frequency modulated components of a signal. This method is proposed by the motivation of the concept such that such constituent AM-FM components have a compact support Fourier spectrum. This method is similar to Fourier spectrum segmentation for separating different modes and apply filter according to the detected Fourier support. This method does not require following a specific approach such as dyadic discretization for computing dilation factor [5]. It detects dilation factor as per characteristics of signal Fourier spectrum empirically. This method is similar to the wavelet transformation method in separating empirical mode in a frequency order from low to high. But bandwidth is not dyadic as the frequency band is segmented empirically.

EWT method addresses the problems with EMD method by adopting wavelet transform and designing appropriate wavelet filter banks that enable decomposition of a signal into a predetermined number of modes [2]. This method has been successfully applied in different domains such as EEG signal analysis [12], decomposing seismic activities [13], and representing time-frequency representation of non-stationary signals [1].

2.3 Variational mode decomposition (VMD)

VMD is a non-recursive decomposition method that decomposes a multi-component signal into constituent amplitude modulated and frequency modulated components in the presence of noise [5][3]. Extract constituent amplitude modulated and frequency modulated components of a multi-component signal dynamically and simultaneously. This method involves defining IMFs as explicit amplitude modulated and frequency modulated models and associating these models' parameter

to the bandwidth of IMFs. This parameter is determined by minimizing bandwidth as per the narrow-band property of IMFs.

This method involves the conversion of real value multi-component signals into the discrete number of sub-signals initially [11]. These sub-signals have specific sparsity traits of bandwidth in the spectral domain. Gaussian smoothness function is applied to each mode of bandwidth. This method has several advantages over other mode decomposition method. The primary advantages include rationale and noise robustness theoretically.

3 Research method and material

This study analyses different adaptive decomposition methods for detecting epilepsy seizures using EEGs based on benchmark real-time epilepsy dataset and the most commonly used neural network classifier with the input of extracted spectral and time-domain features from EEG signals.

This section describes the methodology followed in this work for conducting a comprehensive comparison of adaptive mode decomposition methods in detecting epilepsy seizures using neural network. It explains different phases of the proposed methodology for extracting the time-domain features and spectral features from EEG recordings and converting them into a form compatible with the processing of neural network classifier. It also describes the dataset used and identifies the most commonly used performance metrics in comparing performance of neural network classifier.

3.1 Methodology

In this work, we followed the methodology depicted in Figure 1, which includes five stages, namely, data collection, feature extraction, pre-processing, classification and performance analysis for conducting a comprehensive comparison of adaptive mode decomposition methods using neural network classifiers. The details are described below.

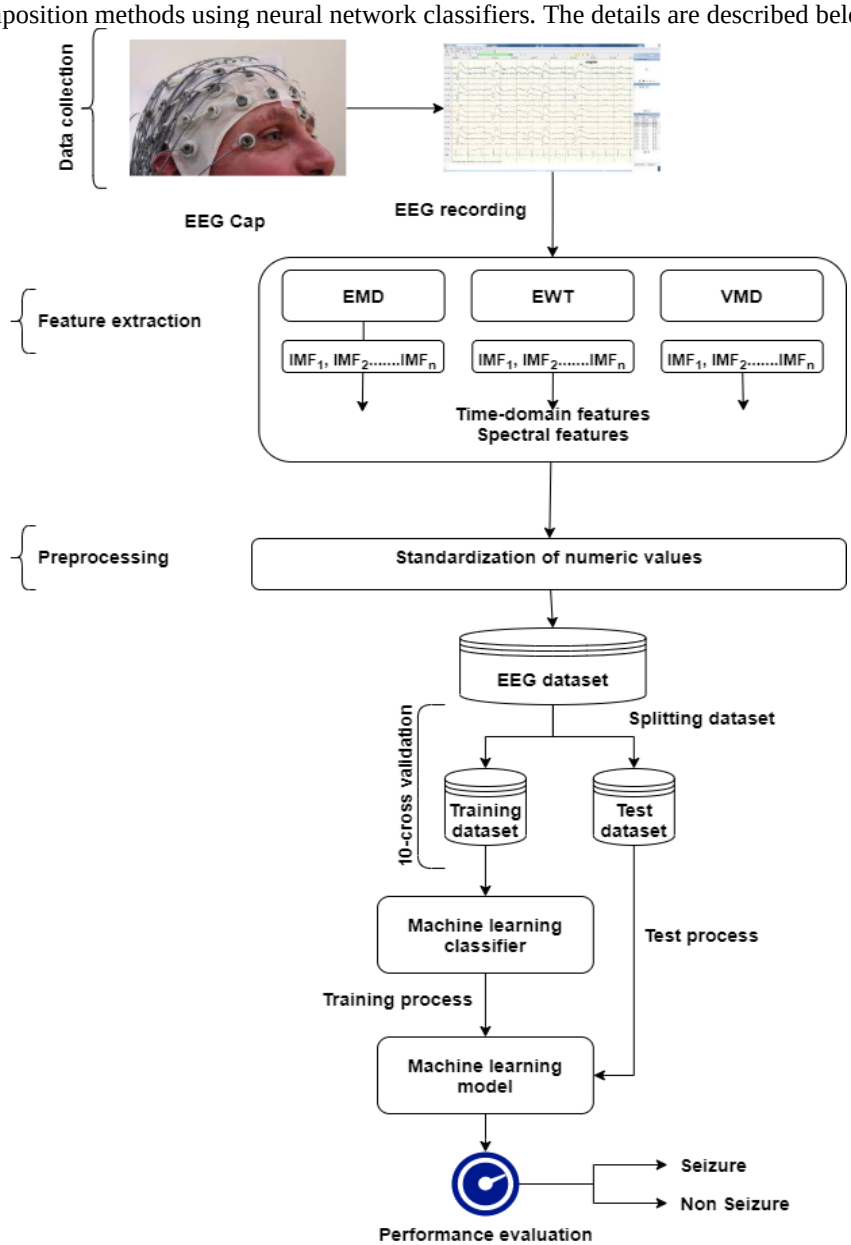


Fig. 1 The proposed framework

3.1.1 Data collection

The experimental data is collected using an EEG cap and other equipment in the form of EEGs. The details of the real-time dataset used in this work are provided in Section 3.2. The collected data is further processed to extract the relevant features in this work.

3.1.2 Feature extraction

In this stage, data signals from EEGs are processed to extract relevant features and arranged in rows and columns. We applied EMD, EWT and VMD as adaptive mode decomposition methods for analyzing EEG signals in 2,4, and 8 modes. The decomposed signals are further analyzed by Hilbert transform to obtain different features.

In this work, we extracted spectral and time-domain features for each adaptive mode decomposition method in different modes. We extracted features using one mode of EEG signals that is considered as frequency modulated and amplitude modulated signals as features of in different modes. The extracted features represent the properties of the spectrum of different signal modes [2]. We extracted nine spectral features and two time-domain features in the set of experiments as described in Table 1.

Table 1 Features extracted from EEG signals

Sr No	Feature	Description
1	Spectral power	The power spectral density (power spectrum) reflects the frequency content of the signal or the distribution of signal power over the frequency
2	Spectral entropy	Spectral entropy (SE) is a measure of signal irregularity, which sums the normalized signal spectral power
3	Spectral peak	EEG power is typically split up into bands that correspond to different spectral peaks related to behavior or cognitive state.
4	Frequency	Frequency associated with spectral peak
5	Spectral centroid	Spectral centroid (SC) measures the shape of the spectrum of EEG signals. It is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes.
6	AM bandwidth	Bandwidth parameters
7	FM bandwidth	Bandwidth parameters
8	Hjorth mobility	Mean frequency of the signal and proportional to the variance of its spectrum
9	Hjorth Complexity	estimate of the signals' bandwidth
10	Skewness	Signal distribution's asymmetry
11	Kurtosis	Tails of the distribution yielded by the signal

3.1.3 Pre-processing

It has been observed that most machine learning methods report better performance when input values are preprocessed to a uniform scale. Normalization and standardization are the most commonly used methods for scaling numeric data to a standard range in the pre-processing stage. The normalization process scales numeric values to a range of 0 to 1. In contrast, the standardization process scales each numeric value separately by subtracting the mean and dividing by standard deviation to shift the distribution with a mean of 0 and a standard deviation of 1.

In this work, we use the standardization process to convert extracted features to a uniform scale using the following equations.

$$X\_standardized = \frac{X - mean}{standard\_deviation} \tag{1}$$

3.1.4 Classification

Classification of epilepsy seizure dataset require training of neural network classifier. The pre-processed data is divided into training data set and test data set. We used a 10-fold cross-validation strategy to train and test the neural network classifier in this work. In the 10-fold cross-validation process, the data set is divided into ten parts. Nine parts are used as training data set, and one part of the dataset is used to evaluate the performance of machine learning classifiers.

We evaluated the neural network classifier with EMD, EWT and VMD adaptive mode decomposition methods in 2, 4, and 8 decomposed modes. Each experiment is repeated ten times, and mean values of results are recorded in terms of identified performance metrics.

3.1.5 Performance evaluation

For evaluating the performance of neural network classifier and conducting a comprehensive comparison, the performance of neural network classifier is measured in terms of four metrics, accuracy, sensitivity, specificity and area under ROC curve (AUC).

These values are computed from the confusion matrix representing the values of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) defined as below [10].

- True positives (TP): Cases predicted as seizures that are seizures.
- True negatives (TN): Cases predicted as non-seizure that are non-seizure.
- False positives (FP): Cases predicted as seizures that are non-seizures.
- False negatives (FN): Cases predicted as non-seizures that are seizures.

These performance metrics can be computed as per the following equations.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{2}$$

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{3}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{4}$$

3.2 Dataset

In this work, we use EEG data set provided by Neurology and Sleep Centre, New Delhi (NSC_ND), for evaluating neural network classifier with adaptive mode decomposition methods. This data set contains segmented EEG recordings of 10 epilepsy patients collected at Neurology and Sleep Centre, New Delhi (NSC_ND) [2][17]. This data set is collected using the Grass Telefactor Comet AS40 amplification system at 200 Hz, and the gold plated scalp EEG electrodes placed following the international 10-20 electrode placement system. Further, the signals have been processed by a band-pass filter with cut-off frequencies of 0.5 Hz and 70 Hz.

These signals are classified into three categories, namely, preictal, interictal and ictal by expert physicians. There are 50 single-channel recordings in each class of data set. Each recording consists of 1024 samples with a duration of 5.12 s. In this work, we classify the dataset samples into three classes, namely, preictal, interictal and ictal and presented the recorded results.

3.3 Experimental setup

The neural network classifier used in this work is implemented in Python language using Scikit library for machine learning methods along with adaptive methods described in [2]. We conducted experiments on a machine Intel Core I3-2330M CPU @ 2.20 GHz, 4 GB RAM, and 1TB HDD. We performed classification of the NSC-ND dataset into three classes, ictal, interictal and preictal. Hyper-parameters of neural network classifier in this work are presented in Table 2.

Table 2 Hyper-parameters of neural network classifiers

Classifier	Hyper parameter values
MLP	alpha=1 max_iter=200 hidden_layer_sizes=100 activation=relu solver=adam learning_rate=constant

4 Experimental results and analysis

We conducted ten independent sets of experiments using different adaptive mode decomposition methods and neural network classifier. We extracted spectral and time-domain features using different adaptive mode decomposition methods by decomposing benchmark real-time EEG signals into 2, 4 and 8-modes.

We employed a neural network classifier to detect epilepsy seizures based upon the extracted spectral and time-domain features from EEG signals of the benchmark NSC_ND dataset. Performance of neural network classifier for each adaptive mode decomposition method is recorded in terms of accuracy, sensitivity, specificity and AUC for ten independent experiments using 10-fold cross-validation strategy.

Figures 2 to 4 present the box plots of experimental results in terms of accuracy, sensitivity, specificity, and area under ROC curve for ten independent experiments using a 10-fold cross-validation strategy using EMD, EWT and VMD mode decomposition methods.

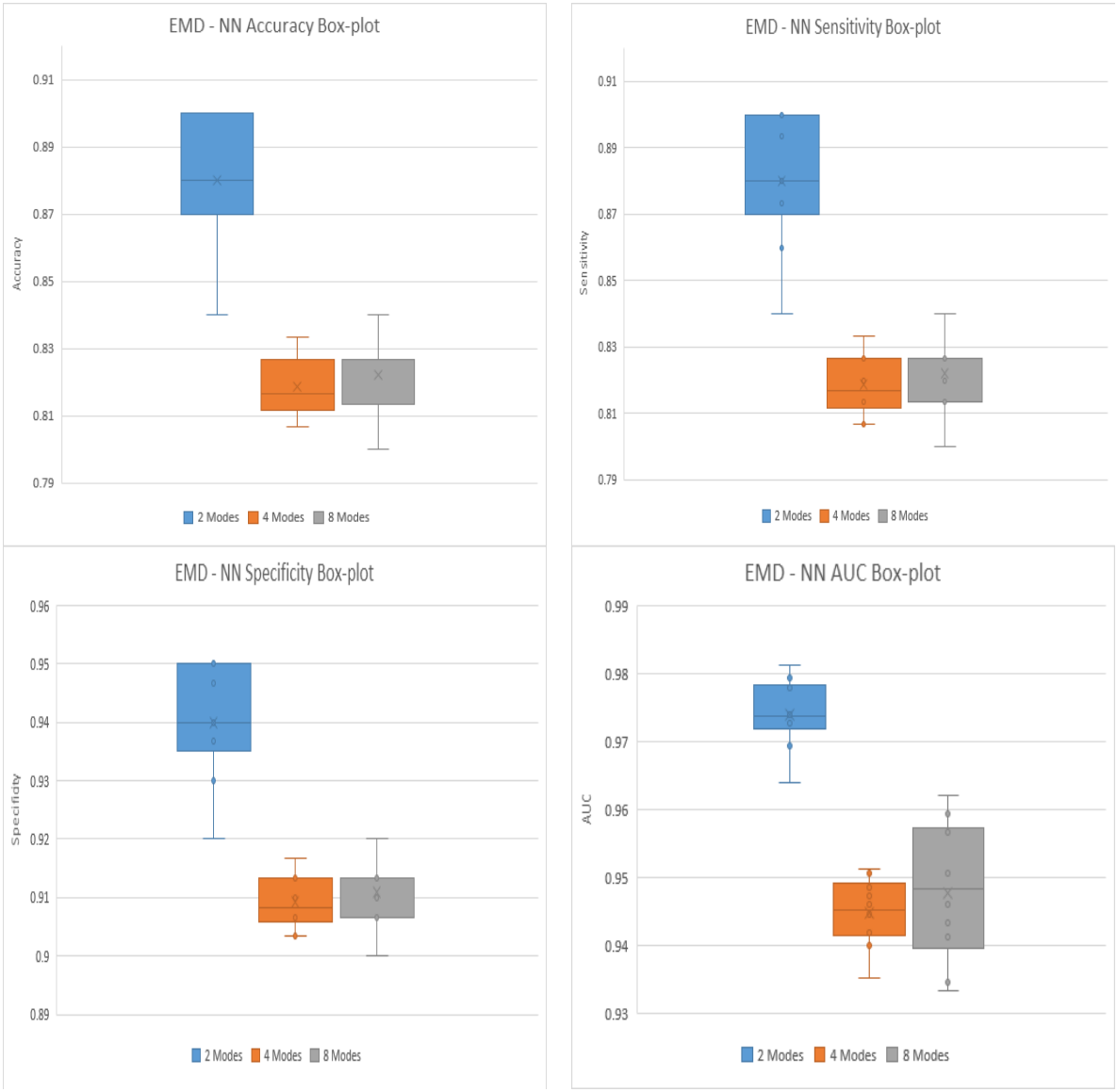
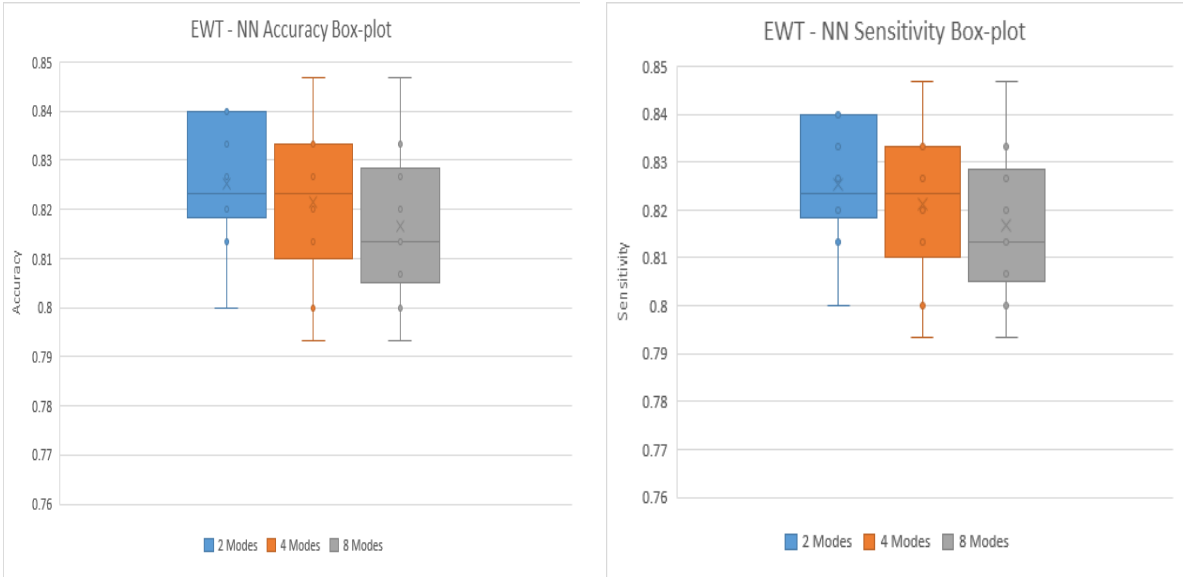


Fig. 2 Box-plots of accuracy of neural network using EMD method



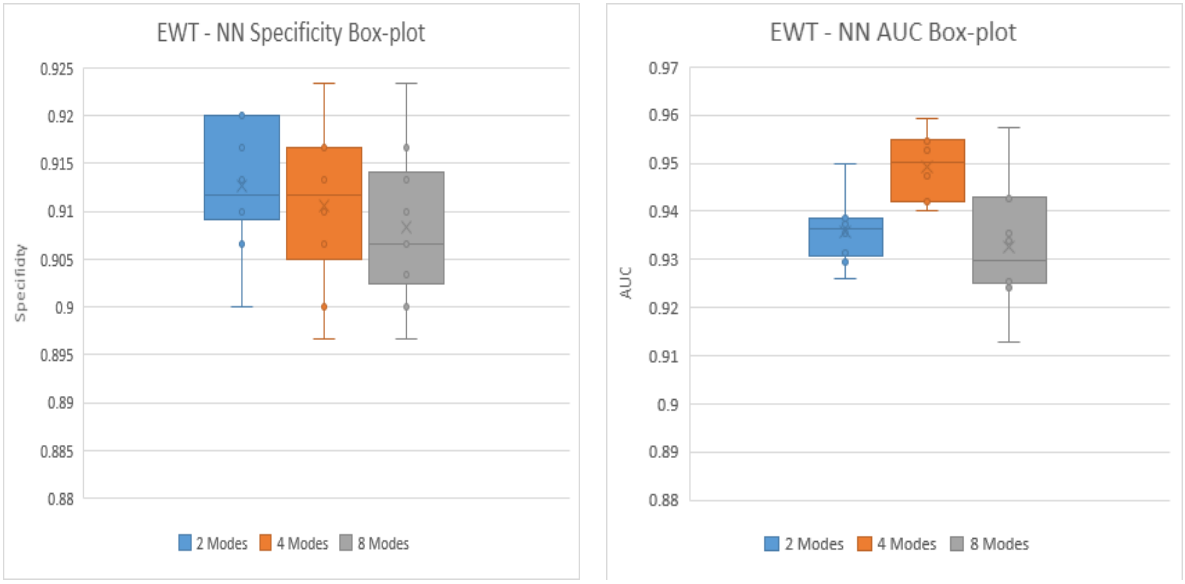


Fig. 3 Box-plots of accuracy of neural network using EWT method

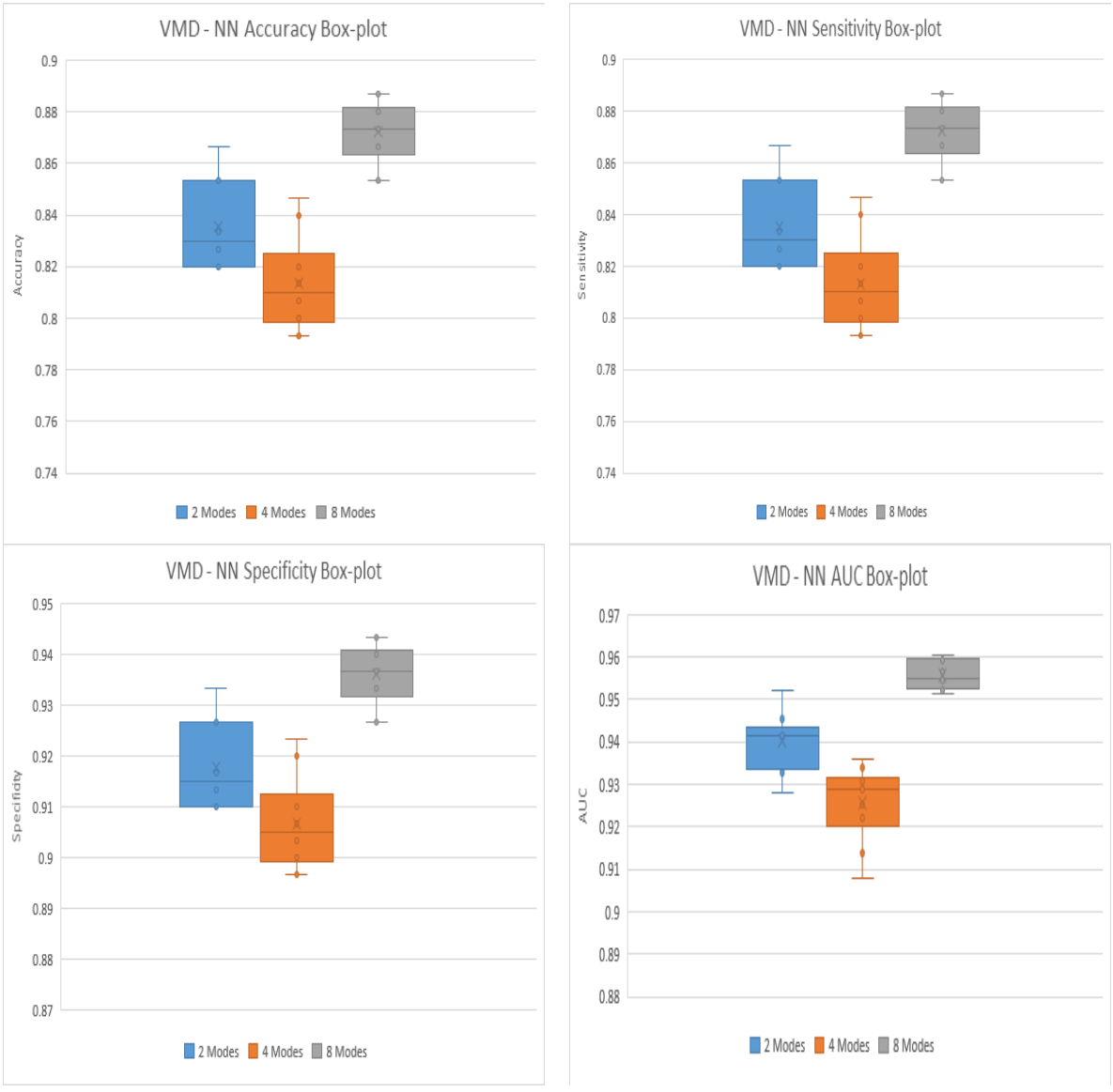


Fig. 4 Box-plots of accuracy of neural network using VMD method

It can be observed from Figure 2 neural network classifier produced stable results for 4 and 8-modes EMD method in comparison to 2-modes in terms of accuracy, sensitivity and specificity. However, in the AUC metric, the neural network classifier gives more stable results in 8-mode decomposition than the other modes of EMD method.

In the EWT method, neural network classifier results are unstable in ten independent experiments for different EWT modes in accuracy, sensitivity, specificity, and AUC metrics as presented in Figure 3.

It can be observed from Figure 4 that neural network classifier results based upon VMD method are more stable than other adaptive mode decomposition methods. VMD method produced are stable results for more number of modes in comparison to less number of signal decomposition modes.

Table 3 to 6 represent a comprehensive comparison of mean values for accuracy, sensitivity, specificity and AUC with standard deviation provided by neural network classifier in ten independent sets of experiments based upon different decomposition modes of ECG signals using EMD, EWT, and VMD mode decomposition methods.

Table 3 Mean accuracy performance of neural network classifier (± standard deviation)

Method/Modes	2-modes	4-modes	8-modes
EMD	0.88 ± 0.0196	0.8187 ± 0.0093	0.822 ± 0.0109
EWT	0.8253 ± 0.0133	0.8213 ± 0.0166	0.8167 ± 0.0121
VMD	0.8353 ± 0.0166	0.8133 ± 0.0181	0.872 ± 0.0121

Table 4 Mean sensitivity performance of neural network classifier (± standard deviation)

Method/Modes	2-modes	4-modes	8-modes
EMD	0.88 ± 0.0196	0.8187 ± 0.0093	0.822 ± 0.0109
EWT	0.8253 ± 0.0133	0.8213 ± 0.0166	0.8167 ± 0.0158
VMD	0.8353 ± 0.0166	0.8133 ± 0.0181	0.872 ± 0.0121

Table 5 Mean specificity performance of neural network classifier (± standard deviation)

Method/Modes	2-modes	4-modes	8-modes
EMD	0.94 ± 0.0098	0.9093 ± 0.0047	0.911 ± 0.0055
EWT	0.9127 ± 0.0066	0.9107 ± 0.0083	0.9083 ± 0.0079
VMD	0.9177 ± 0.0083	0.9067 ± 0.009	0.936 ± 0.006

Table 6 Mean AUC performance of neural network classifier (± standard deviation)

Method/Modes	2-modes	4-modes	8-modes
EMD	0.9739 ± 0.005	0.9449 ± 0.005	0.9478 ± 0.0099
EWT	0.9356 ± 0.0066	0.9494 ± 0.0066	0.9325 ± 0.0128
VMD	0.9402 ± 0.0069	0.9257 ± 0.0088	0.9558 ± 0.0035

Figure 5 (a)-(d) provides a comparative analysis of accuracy, sensitivity, specificity and AUC. It can be concluded from Figure 5 (a) that EMD method provides an accuracy of 88% approximately for 2-mode decomposition. For 4-mode decomposition, EMD and EWT provided comparable accuracy results of approximately 82%. VMD method produced accurate results for or 8-mode decomposition of EEG signals. It can be concluded from Figure 5 that EMD method is suitable for a low number of modes for signal decomposition, whereas VMD method provides high accuracy with high decomposition modes.

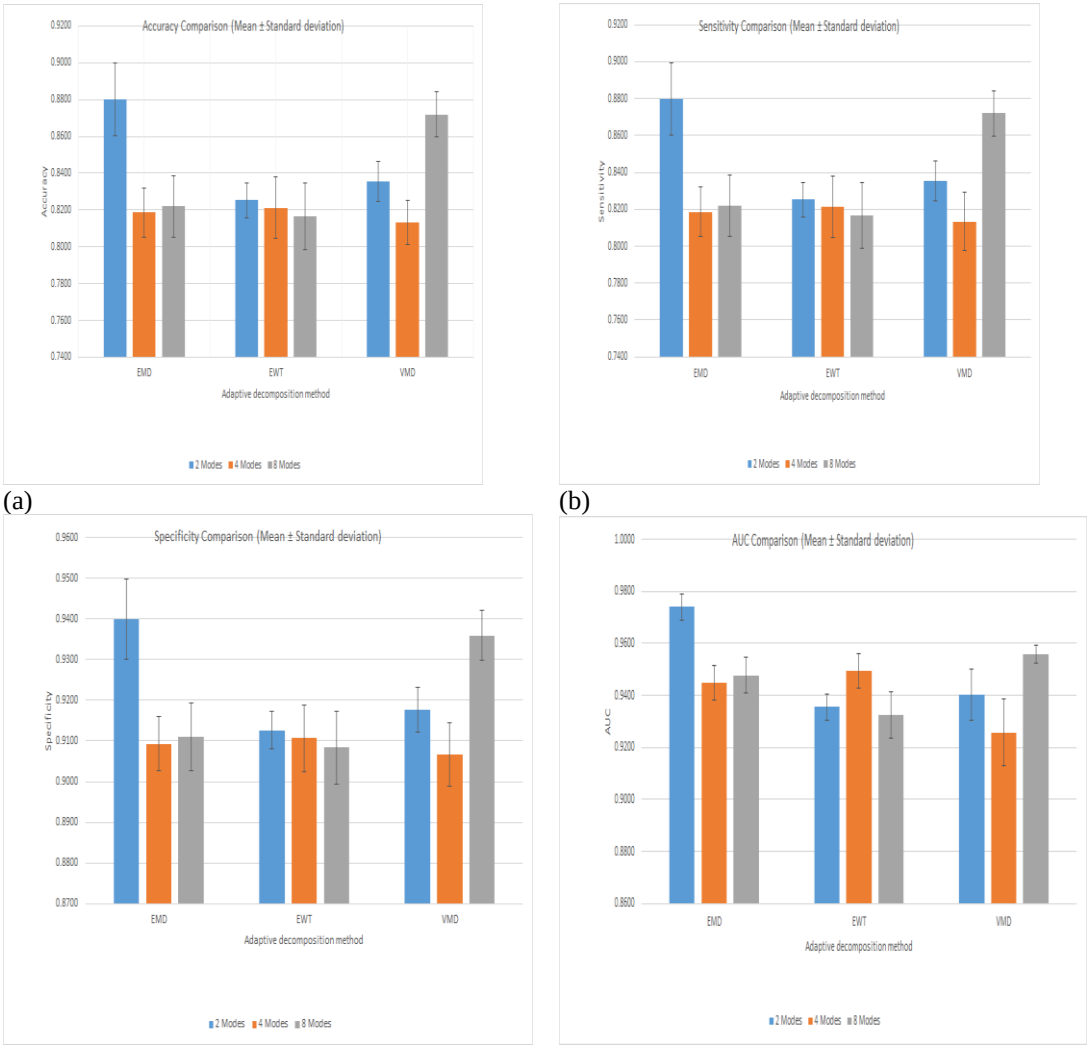


Fig. 5 Mean performance comparison of adaptive mode decomposition methods(± standard deviation)

Similar results have also been observed for sensitivity and specificity metric based comparison as depicted in Table 4 & Figure 5 (b) And Table 5 & Figure 5 (c), respectively. EMD method provided neural network classifier result of sensitivity up to 88% by decomposing EEG in 2-modes. VMD method provided a sensitivity result of 87% approximately for 8-mode decomposition of EEG signals.

The performance of neural network classifier in terms of specificity metric have been observed to be 94% using 2-mode decomposition of EMD method as presented in Table 5 and Figure 5 (c) . Whereas, the similar performance of VMD method has been observed using neural network classifier in terms of specificity up to 93.6% for 8-mode decomposition of EEG signals.

In terms of AUC, it can be observed from Table 6 and Figure 5 (d) that EMD with a 2-mode decomposition of EEG signals resulted in the performance of neural network classifier to be 97.4% approximately. At the same time, 95.6% of AUC performance is recorded for the neural network using a mode decomposition of ECG signal by VMD method.

It can be observed from Table 6 and Figure Figure 5 (d) that EWT method reported AUC for neural network classifier approximately 93% for different decomposition modes.

Reporting results reflect the overall extraction of promising features using the adaptive decomposition methods, specifically by EMD and VMD from EEG signals. The high value of accuracy, sensitivity, specificity and AUC demonstrate better class separability of the extracted spectral and time-domain features using adaptive mode decomposition methods. EMD method reaches the best accurate results with 2-mode signal decomposition, whereas VMD provides the best accuracy results for 8-mode decomposition of EEG signals. The extracted spectral and time-domain features using adaptive decomposition method in different modes enables classification of standard and epilepsy seizure signals accurately using neural network classifier.

5 Conclusion

This work analyses the effectiveness of adaptive mode decomposition methods for classifying epilepsy seizures by analyzing EEG signals. EEGs signals have been decomposed into different modes in the form of IMFs using the adaptive mode decomposition methods, EMD, EWT and VMD. Various spectral and time-domain features have been extracted

from decomposed modes of EEG signals. A neural network, an efficient and effective classifier, is used for classifying epilepsy seizures from the extracted spectral and time-domain features. The performance of the neural network classifier is analyzed for different adaptive mode decomposition method with different segmented modes in terms of accuracy, sensitivity, specificity and AUC metrics.

This work compares the effectiveness of adaptive mode decomposition methods for detecting epilepsy seizures by analyzing EEG signals. The ECG signals are decomposed into different modes in the form of IMFs. Various spectral and time-domain features are extracted. The extracted features are used to train a neural network classifier. The trained neural network model is tested using 10-fold cross-validation strategies for different adaptive modakam position methods using 2, 4 and 8-mode signal decomposition methods. The neural network classifier's performance is recorded and compared in terms of accuracy, sensitivity, specificity, and AUC.

The reporting results demonstrate the effectiveness of the adaptive mode decomposition method in successfully classifying epilepsy seizures from the seizure-free signals. It is concluded that EMD method provides its best reserves in 2-mode decomposition, whereas VMD provide the best reserves in 8-mode decomposition. EMD method resulted in classification accuracy of neural network up to 88%, whereas VMD method performed an accuracy of 87% in classifying epilepsy seizure EEG signal from standard EEG signals. The reporting results demonstrate using the adaptive mode decomposition methods as a promising direction for developing an automatic epilepsy seizure detection system based upon spectral and time-domain features.

Conflict of interest disclosure

The authors declare that there is no conflict of interest regarding the publication of this paper.

References

1. Bhattacharyya, A., Singh, L., Pachori, R.B.: Fourier–bessel series expansion based empirical wavelet transform for analysis of non-stationary signals. Digital Signal Processing **78**, 185–196 (2018)
2. Carvalho, V.R., Moraes, M.F., Braga, A.P., Mendes, E.M.: Evaluating five different adaptive decomposition methods for eeg signal seizure detection and classification. Biomedical Signal Processing and Control **62**, 102073 (2020)
3. Dragomiretskiy, K., Zosso, D.: Variational mode decomposition. IEEE transactions on signal processing **62**(3), 531–544 (2013)
4. Fasil, O., Rajesh, R.: Time-domain exponential energy for epileptic eeg signal classification. Neuroscience letters **694**, 1–8 (2019)
5. Feng, Z., Zhang, D., Zuo, M.J.: Adaptive mode decomposition methods and their applications in signal analysis for machinery fault diagnosis: a review with examples. IEEE access **5**, 24301–24331 (2017)
6. Flandrin, P., Rilling, G., Goncalves, P.: Empirical mode decomposition as a filter bank. IEEE signal processing letters **11**(2), 112–114 (2004)
7. Gilles, J., Tran, G., Osher, S.: 2d empirical transforms. wavelets, ridgelets, and curvelets revisited. SIAM Journal on Imaging Sciences **7**(1), 157–186 (2014)
8. Im, C.H.: Computational EEG Analysis. Springer (2018)
9. Khamis, H., Mohamed, A., Simpson, S.: Frequency–moment signatures: a method for automated seizure detection from scalp eeg. Clinical Neurophysiology **124**(12), 2317–2327 (2013)
10. Kumar, G.: Evaluation metrics for intrusion detection systems-a study. Evaluation **2**(11), 11–7 (2014)
11. Kumar, M.R., Rao, Y.S.: Epileptic seizures classification in eeg signal based on semantic features and variational mode decomposition. Cluster Computing **22**(6), 13521–13531 (2019)
12. Kumar, R., Saini, I.: Empirical wavelet transform based ecg signal compression. IETE journal of research **60**(6), 423–431 (2014)
13. Liu, W., Cao, S., Chen, Y.: Seismic time–frequency analysis via empirical wavelet transform. IEEE Geoscience and Remote Sensing Letters **13**(1), 28–32 (2015)
14. Mahmoodian, N., Boese, A., Friebe, M., Haddadnia, J.: Epileptic seizure detection using cross-bispectrum of electroencephalogram signal. seizure **66**, 4–11 (2019)
15. Misiunas, A.V.M., Meškauskas, T., Samaitienė, R.: Algorithm for automatic eeg classification according to the epilepsy type: Benign focal childhood epilepsy and structural focal epilepsy. Biomedical signal processing and control **48**, 118–127 (2019)
16. Savage, N.: Epidemiology: the complexities of epilepsy. Nature **511**(7508), S2–S3 (2014)
17. Swami, P., Gandhi, T.K., Panigrahi, B.K., Tripathi, M., Anand, S.: A novel robust diagnostic model to detect seizures in electroencephalography. Expert Systems with Applications **56**, 116–130 (2016)

MASSIVE DOWNFALL IN $PM_{2.5}$ AND PM_{10} IN DURING PANDEMIC IN PUNJAB

Bachandeep Singh Bhathal[1], Dr. Gaurav Gupta[2], Dr. Brahmaleen K. Sidhu[2]
[1]Research Scholar, Department of Computer Science and Engineering, Punjabi University, Patiala
[2]Assistant Professor, Department of Computer Science and Engineering, Punjabi University, Patiala

ABSTRACT: During this pandemic, all the industries, transport other sources of pollution are standstill, which were the major contributor to pollute the air quality. Air quality during the COVID-19 pandemic January-June 2020 and before the COVID-19 January-June 2019 is compared. The air pollution quality is compared on the basis of particulate matter $PM_{2.5}$ and particulate matter PM_{10} which are major causes of air pollution. Other factors like visibility of objects at long distances and easiness in breathing specifically in urban areas are also observed. The paper highlights the comparison of air pollution before and during the pandemic time of Ludhiana city (hub of industry). The data is collected, processed and analyzed. A noticeable change has been found in $PM_{2.5}$ and PM_{10} during the mentioned period. According to the survey in the local region, health issues are reduced which were due to air pollution. In this paper, future direction is also provided that can be followed after pandemic effects.

KEYWORDS: COVID-19, $PM_{2.5}$, PM_{10}, Air Pollution

1. INTRODUCTION

The human society has been significantly affected due to the COVID-19 pandemic, with which the health care, social relationships, and financial structures are also affected. A worldwide response that incorporates terminations of associations and social removing has shaped top-notch neighborhood outcomes. The health impacts of pandemic COVID-19 remains the uppermost priority, it is however unknown but the pandemic might also additionally moreover have a touching on numerous factors, especially the threat of pollutants. Air pollutants message is a partner essential and chronic chance problem for respiratory and metabolism health results (Shaddick et al., 2018). But ambient air pollutants are affected through huge disorders in the behavior of pandemic COVID-19 and will furnish essential clues related to health and control of air pollutants emissions (Burnett et al., 2018).

In maiden evaluations, a wilt in human-made ambient air pollution has been examined in the nations which is responding to the pandemic COVID19.

The satellite derived concentrations of various pollutants in northern India from March 2020 which is as low as 20 years ago has been found by NASA (Patel et al., 2020). After the implementation of the lockdown it has been observed that not only the air quality has been improved in the various cities, but this thing has also headed to the decrease in the usual temperature also in this time (Khanna, 2020). From March 2020 onwards. It has been reported that $PM_{2.5}$ and PM_{10} shows a decline of 14% to 30% as compared to 2019. In any case, whereas distantly detected air contamination tiers deliver a superb gauge of expansive shows, there's a characteristic incentive for validation of contamination styles the use of in-situ estimations (Bechle et al., 2013).

The Ground primarily based total estimations communicate to the standard terrific stages for poison fixations and are the tool for body consistency. The calculable fixations should be applied to come to a decision air contamination changes, chiefly the situation hearty checking systems that exist.

Our survey has explored the present impact of the pandemic COVID19 on Ludhiana city pollution in India using Punjab Pollution Control Board air pollution network. We hypothesize a drastic decrease in the $PM_{2.5}$ and PM_{10} fine particulate matters during COVID-19 pandemic lockdown in the consequent to reduced public and non-public traffic along with the industrial business.

2. Material and methodology

The ambient air pollution has been acquired along with the measurements in Ludhiana, India for $PM_{2.5}$ and PM_{10} from January 1-June 31st 2019 to January 1-June 31st 2020 through Online Environmental Pollution Monitoring (PPCB, 2020). The matched pollution data from the dates has been taken from the central pollution control board. Each monitor was assigned with the daily $PM_{2.5}$ and PM_{10} 24-h mean values. Monitors were restricted in the Ludhiana city only and have been also restricted to the $PM_{2.5}$ and PM_{10} concentrations only to make a consistent comparison.

We have classified our data in two groups: the COVID-19 pandemic period spans January 1-June 30th, 2020 with the time the pandemic has attacked the world level and had been affecting the various industries and vehicle movement. The pre-COVID-19 period data from January 1-June 30th, 2019. The nation worldwide has been imposed on the 24th March 2020 and which started affecting the business and the vehicle moments. The Ludhiana city which is also called as hub of the industry has been also affected with the same scenario, the industries along with the businesses has been came to close.

We have calculated the summary statistics for both PM_{10} and $PM_{2.5}$ before pandemic COVID-19 and during pandemic COVID-19 periods. The data has been taken in the daily concentration and then has been converted to the monthly average concentration by using arithmetic mean formula i.e. "**Arithmetic Mean = (1/N) * (x1 + x2 + … + xN)"** (Brownlee, 2019). Comparison has been done between the before COVID-19 and during COVID-19 pandemic by two-sided t-test. We have acquired both the data, which is absolute differences in pollution as well as the percentage change in pollution from before and during pandemic. Data along with the software's which has been used in this research are publically available.

3. Result

We have illustrated the pollutant differences of $PM_{2.5}$ and PM_{10} concentrations during January-June 2019 to January-June 2020 in table 1, which is before and during the pandemic COVID-19. The change in the pollution for $PM_{2.5}$ were arithmetically significant and the reduction of up to 58% in total value. The statistically decrease in the number of $PM_{2.5}$ concentrations can be seen in the regards to when the lockdown was imposed and the non-essential businesses were closed. The similar difference and decline in PM_{10} can be also seen in the regards to the same as of 47%.

Fig 1 and Fig 2 provides a visual comparison between the both time-spans for Ludhiana city. The Measurements of $PM_{2.5}$ and PM_{10} reveled drastic drop in the concentrations. However the mixed responses can be seen in the early January to March but later when lockdown was imposed nationwide the huge change can be seen in the concentration. AOD visual comparison can be seen in Fig 3 over India from 2019 to 2020. The concentrations of aerosols are near to surface then the optical depth can be 1 or above and results to the hazy conditions, whereas if thickness of the same is less than 0.1 onto the whole atmospheric then it is considered as clean (Patel et al., 2020).

4. Conclusion

Our findings has presented the significance that the measured pollution in the ambient air has been drastically decreased in the Ludhiana city during pandemic COVID19 which also include a decrease of 58% in $PM_{2.5}$ and 47% in PM_{10}. The contribution of multiple non-transport sources, such as the industries, biomass burning and food industries are included in this. Moreover these reduction can be seen in the month of the April 2020 which was the total lockdown period for small businesses and the large industries also. There is a possibility that the pandemic COVID19 continues, which can result to the broader decline in the Particulate $Matter_{2.5}$ and Particulate $Matter_{10}$ concentrations.

Table 1
Ambient Air Pollution concentration during current and before timeframes of Ludhiana city for daily concentration of $PM_{2.5}$ and PM_{10}.

Month	Before COVID-2019 (Pandemic) (January-June 2019) Mean (sd)	During/Current COVID-2019 (Pandemic) (January-June 2020) Mean (sd)	Before and Current Means Difference	Change in %
PM_{10} µg/m^3				
January	142.24	85.99	56.25	39.54%
February	147.54	99.44	48.1	32.60%
March	87.4	60.24	27.16	31.07%
April	82.03	43.5	38.53	46.97%
May	83.85	76.19	7.66	9.13%
June	101.24	86.33	14.91	14.72%
$PM_{2.5}$ µg/m^3				
January	49.66	44.6	5.06	10.18%
February	48.7	45.69	3.01	6.18%
March	50.19	30.72	19.47	38.79%
April	45.24	19.03	26.21	57.93%
May	51.49	36.74	14.75	28.64%
June	46.44	32.61	13.83	29.78%

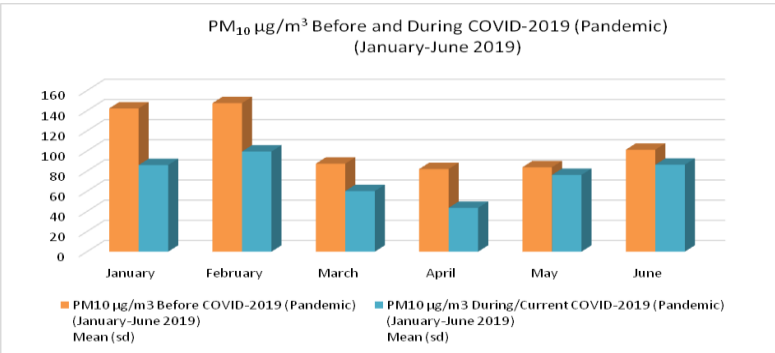


Figure 1: During (January-July 2020) and Before (January-July 2019) Ludhiana city concentration of PM_{10} in µg/m^3

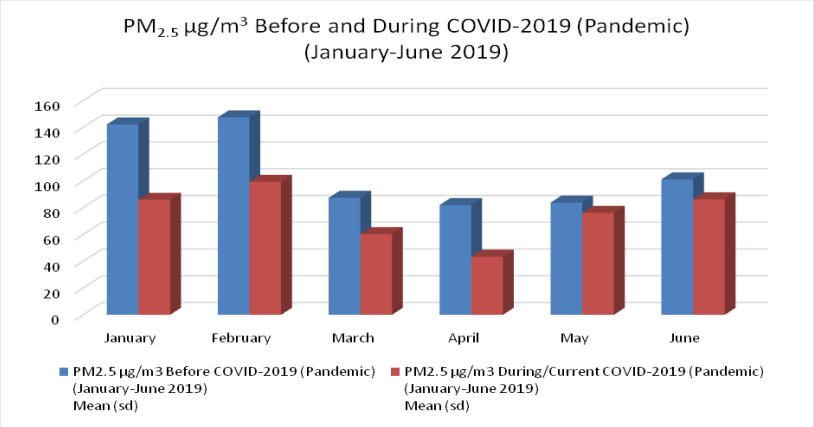


Figure 2: During (January-July 2020) and Before (January-July 2019) Ludhiana city concentration of Particulate Matter$_{2.5}$ in µg/m^3

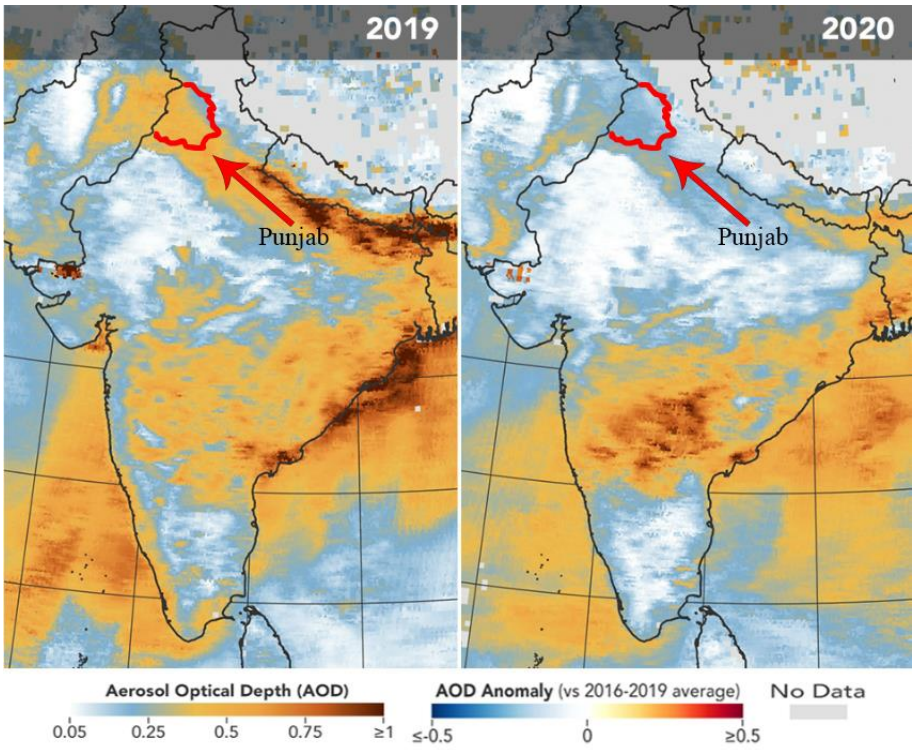


Figure 3: The visual comparison of AOD over India from 2019 to 2020.

The health related consequences has been occurred due to the pandemic COVID19. 1 µg/m^3 can increase the exposure of 15% for $PM_{2.5}$ and PM_{10} and increase in the mortality (Wu et al., 2020). The increase in the pollution in the ambient air increase in the fatality rates which are varying from the severe acute respiratory syndrome (SARS) in China (Cui et al., 2003). These research also indicate a position for ambient air pollutants to worsen pandemic COVID19 and have an effect on the pointy disparities found amongst patients. Our findings has highlighted the importance of the persisted ambient air excellent enforcement to well shield the public.

References

Agirre-Basurko, E., G. Ibarra-Berastegi, and I. Madariaga. 2006. "Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area." *Environmental Modelling and Software* 430-446.

Ardiansyah, Ahmad Yusuf, Riyanarto Sarno, and Oxsy Giandi. 2018. "Rain Detection System for Estimate Weather Level Using Mamdani Fuzzy Inference System." *International Conference on Information and Communications Technology*. 848-854.

CPCB. 2009. November 18. Accessed May 2021. https://cpcb.nic.in/displaypdf.php?id=aG9tZS9haXItcG9sbHV0aW9uL1JlY3ZlZC1OYXRpb25hbC5wZGY=.

Freeman, Brian S., Graham Taylor, Bahram Gharabaghi, and Jesse Thé. 2018. "Forecasting air quality time series using deep learning." *Journal of the Air & Waste Management Association.*

Gorai, Amit Kumar, Pramila Goyal, and Kanchan Kanchan. 2015. "A Review on Air Quality Indexing System." *Asian Journal of Atmospheric Environment* 101-113.

Gouveia, N., and T. Fletcher. 2000. "Time series analysis of air pollution and mortality: effects by cause, age and socioeconomic status." *Journal of Epidemiology and Community Health* 750-755.

Iqbal, Kashif, Muhammad Adnan Khan, and Areej Fatima. 2018. "Intelligent Transportation System (ITS) for Smart-Cities using Mamdani Fuzzy Inference System." *International Journal of Advanced Computer Science and Applications* 94-105.

Kuanr, Madhusree, Bikram Kesari Rath, and Sachi Nandan Mohanty. 2018. "Crop Recommender System for the Farmers using Mamdani Fuzzy Inference Model." *International Journal of Engineering & Technology* 277-280.

Mamdani, Ebrahim H. 1977. "Application of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis." *IEEE Transactions on Computers* 1182-1191. doi:10.1109/TC.1977.1674779.

Mamlook, Rustum, Omar Badran, and Emad Abdulhadi. 2009. "A fuzzy inference model for short-term load forecasting." *Energy Policy* 1239-1248.

Pourjavad, Ehsan, and Arash Shahin. 2018. "The Application of Mamdani Fuzzy Inference System in Evaluating Green Supply Chain Management Performance." *International Journal of Fuzzy Systems, Springer* 901–912.

PPCB. 2020. *Pollution, Monitoring Online Environmental.* http://www.envsaindia.com/aqdm/show_map.php.

Rahman, N. H. A., M. H. Lee, M. T. Latif, and S. Suhartono. 2013. "Forecasting of air pollution index with artificial neural network." *Jurnal Teknologi* 63.

Rojas, I., O. Valenzuela, M. Anguita, and A. Priet. 1998. "Analysis of the operators involved in the definition of the implication functions and in the fuzzy inference process." *International Journal of Approximate Reasoning* 367-389.

Wang, Hsiao-Fan, and Ruey-Chyn Tsaur. 2000. "Insight of a fuzzy regression model." *Fuzzy Sets and Systems* 355-369. doi:https://doi.org/10.1016/S0165-0114(97)00375-8.

Yuan, Z., X. Zhou, T. Yang, J. Tamerius, and R. Mantilla. 2017. "Predicting traffic accidents through heterogeneous urban data: A case study." *Proceedings of the 6th International Workshop on Urban Computing (UrbComp 2017), Halifax, NS, Canada.*

Zhu, Xiaodong, Zhiqiu Huang, Shuqun Yang, and Guohua Shen. 2007. "Fuzzy Implication Methods in Fuzzy Logic." *IEEE Computer Society* 154–158. doi:10.1109/FSKD.2007.327.

FEATURE SELECTION FROM E-COMMERCE DATA FOR CUSTOMER CHURN PREDICTION USING DATA MINING

Seema[#1], Gaurav Gupta[#2]
[1]Research Scholar, [2]Assistant Professor
[#] *Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India*
[1]seemabaghla@pbi.ac.in
[2]gaurav.shakti@gmail.com

ABSTRACT — E-commerce is an electronic activity that helps to sell and buy goods and services using the internet involving the transfer of money as well as information for facilitating services and transactions. In the current scenario, each e-business company is striving to increase the number of customers while keeping their customers. Adding new customers to any business is very costly and time-consuming. It is really important for them to predict the customers on the verge of churn. For this, they need an efficient churn prediction model for the purpose of identifying potential churning customers. If companies come to know priorly, which customers may stop or reduce the use of goods and services offered by them, they can plan suitable remedial actions to retain them. Customer churning is influenced by many complex features and large data, the prediction of customer churn manually is very difficult or almost impossible. In this paper, feature selection from pre-processed e-commerce data for the prediction of customer churn using data mining has been attempted. The neighborhood component analysis (NCA) technique has been applied for feature selection from the benchmarked Brazilian e-commerce dataset. Before applying NCA, the dataset is pre-processed by applying various pre-processing techniques such as missing value imputation, data cleaning, data normalization, and data balancing, etc. The methodology, GUI design, results, conclusion, and future scope are presented. In future, authors intend to work develop customer chum prediction data mining classification models in e-commerce based on the selected features using machine learning and deep learning approaches.

KEYWORDS- Brazilian E-commerce dataset, feature, cleaning, balancing, normalization, feature selection, missing values, feature selection.

INTRODUCTION

Data mining is the process of storing, extracting, editing and deleting data in databases [1]. The basic aim of data mining is to extract the information from a dataset and transform it into a usable form without any abnormalities. This is an iterative process aimed to apply data mining algorithms such as classification, regression clustering, etc. on various data in order to extract useful, novel and valid patterns [1]. Fig. 1 shows knowledge data discovery process in data mining.

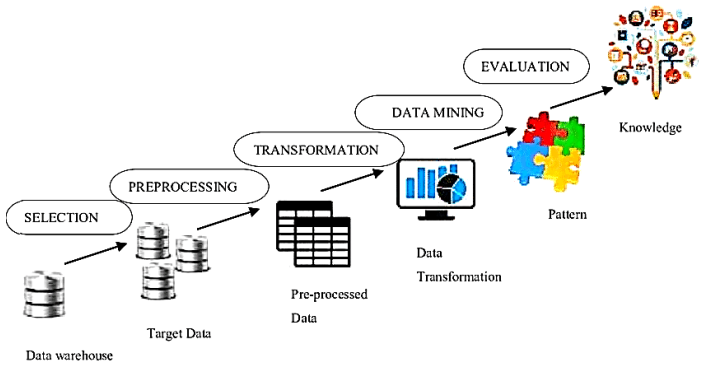


Fig. 1: Knowledge discovery process steps [1]

Various steps of the knowledge discovery process are explained below [1, 2].

1. Data Collection [1, 2]: Collection of accurate data is very important and the basis for building the good data mining models [1]. Data collection involved collection of accurate, accessible and important data. The efficiency of the data can be established by level of integration it can support and how accurately the models can leant from it. Abnormal and inadequate data can lead to generation of unsuccessful and useless data models [2]. Hence, it is really important to collect best data free from abnormalities. It is equally important to omit any personal data to avoid biasness. During collection of data, the storage patterns must also be defined for better data storage [2]. Further, the dataset should be enriched and improved by adding more data and attributes in order to judge its impact on knowledge discovery [2].
2. Data Selection [1, 2]: During this phase, appropriate data is selected from the database for further processing. It involves selection of appropriate data source, type, method and instrument so that the desired research objectives can be achieved and research questions could be answered effectively [2]. The selection of suitable data has significant impact on the model performance. The selection of relevant dataset usually depends on the type of analysis and model to be designed. The adequacy, cost and availability of data play a key role in an efficient model development [2].

There some key issues such as finding relevant data sources, data type and adequate data for answering research questions, selecting data sample, appropriate data collection instrument, and their compatibility with each other etc [2].

3. Data pre-processing [1, 2]: Normally real world data contains vague information, noise and missing values that can cause degradation of data quality [2]. Real world data may also have inconsistency, incompleteness, lack of certain attributes. Due to this, the results are generally of low quality after data mining which lead to development of useless models. It is of upmost importance to improve the data quality before data mining and model development. This can be done using data pre-processing [2, 3]. Data pre-processing involves transformation of the collected data into usable and understandable form relevant to the application. It is a process upgrading data quality by correcting various abnormalities such as missing data, vague data, noise, imbalance, for further analysis data mining, and model development [1]. With pre-processing, collected data is converted in usable form for further processing and analysis. It improves the reliability of data. Data pre-processing includes data cleaning, imputing missing data values, removing noise and outliers using suitable data mining techniques [2].
4. Data Transformation [2]: Data transformation is used to generate data in suitable format. It involves feature selection & extraction, dimension reduction, as well as attribute transformation. The pre-processed data is transformed from one format to another to make it appropriate and suitable for use of data analysis effectively using data mining [2]. Some strategies in data transformation include normalization, smoothing, generalization, aggregation; attribute construction etc [2]. Normalization has been used to scale different kind of data having different attributes in the range of [0, 1]. To remove any noise present, smoothing is used. Generalization is used to convert low level data into high level data. The data is stored, presented and aggregated in the format required by an application using aggression [2]. Finally, for improved data mining, new attributes are constructed using existing attributes in attribute construction [2].
5. Data Mining [2]: Data mining, an important step of KDD process, is used to convert raw data into useful information. It involves analysis of patterns of data to draw meaningful inferences from the data [2]. It involves automatic discovery of data patterns, predictive model development, and further suggesting some meaningful action etc. The tasks in data mining are [2] (i) anomaly detection for identifying anomaly in data (ii) association rule mining for searching the relationships between data items and features (iii) clustering for discovering structures in given data (iv) classification for generalizing the data (v) regression for estimating relation between datasets and further, modelling of the data and (vi) summarization for representing the data in compact form.
6. Data Interpretation/Evaluation [2]: This is used to evaluate and interpret the data mining results in order to arrive at relevant conclusion [2]. In this process, inferences and conclusions are drawn on the basis of analysis results. Data interpretation and evaluation is used to assess data patterns, comprehensibility, utility, and reliability of the developed model [2].

LITERATURE REVIEW

Berger and Kompan [3] developed a neural network model for prediction of user based on its web surfing for real data of an online company. Data cleaning, normalization and balancing were applied in order to improve the dataset. Authors proved that prediction accuracy and precision of the proposed model was better as compared to the existing baseline models. Authors suggested that proposed approach further can be used for other applications. Bahari and Eloyidom [4] developed a NN based customer behavior prediction model in banking for a UCI dataset of Portuguese bank. The dataset pre-processed for any abnormality and missing values. Authors concluded that that developed model had performed better in comparison to existing model based on accuracy, TP rate and specificity. Authors further advised that improved model with higher sensitivity can be developed for other applications. Kumar and Yadav [5] developed an ANN based churn prediction model for food delivery customers. Authors scrapped data from Mouthshut.com and further data pre-processed data. The validation parameters such as accuracy, TP rate, TN rate, FP rate & FN rate are taken for study. The reviews of the customers were understood by sentimental analysis. Authors concluded that model would be very useful to make intelligent decisions related to customer churn & retention.

Khede et al. [6] demonstrated the applications of machine-learning technique churn prediction by developing classification models using machine leaning. Authors developed customer churn prediction model for a telecom dataset. The pre-processing of the data was done by imputing missing data and remove abnormality in the data. AUC, ROC, accuracy and confusion matrix was calculated for validation of the developed model. Authors compared the performance of different churn prediction classification models using machine learning techniques and suggested that improved models can be developed for better accuracy and prediction. Jha et al. [7] proposed k-means clustering based classification model using logistic regression to predict the customer's churning behaviour. The dataset was pre-processed to impute missing values, remove noise and correct data imbalance. Parameters such as customer lifetime value (CLV), accuracy & recency, frequency and monitory (RFM) were taken for model validation. Authors concluded that the developed model will be useful to predict the customer loyalty in an effective way. It was suggested that improved model will be very helpful with good prediction accuracy. Ahmed et al. [8] proposed a telecom customer churn prediction model using four data mining techniques namely decision tree, random forest, gradient boosted machine tree and extreme gradient boosting. Average method was used for missing value imputation and sample selection was done by random sampling. Authors suggested that same model can be used for churn prediction in e-commerce and other applications. The model can also be improved for better performance.

Vijaya and Sivasankar [9] proposed a PSO based customer churn prediction model in telecom for UCI French Orange telecom dataset. Missing value imputation, data cleaning and data balancing was done during pre-processing. The performance of PSO-FSSA based model was found better as compared to other existing models. Yu et al. [10] proposed a SVM based churn prediction model in e-commerce. Data cleaning was done to remove NAN values and missing values were imputed using average method. The imbalance in the dataset was removed using data balancing. Parameters viz. coverage rate, hit rate, accuracy, and lift coefficient were used for model validation and comparison. The results of developed model were compared with neural network and decision tree based model. It was suggested that developed model is highly recommended for use to predict churn for highly imbalance data of multiple e-commerce sites. Pondel et al [11] developed deep learning ANN based model for customers' churn prediction in e-commerce in retail sector. Authors concluded that accurate prediction of customer churn and further actions can result in higher customer retention of churning customers. Authors further concluded that the paper fills a research gap and contributes to the existing literature via the developed model as it worked well with good accuracy, precision and recall. They added that the prediction model development in e-commerce is extremely challenging and needed in the current scenario. Kavya et al. [14] developed SVM based classification model for classification of breast mammograms to classify them as abnormal and normal. Authors used neighborhood component analysis for selection of features for said classification based on tamura and statistical features from the breast mammograms. Authors proved that SVM classifier with quadratic kernel has a good accuracy. They suggested that the proposed models can be used for other applications too.

RESEARCH GAPS

For any type of research study, data is very important study irrespective of its type. Large datasets can have problems such as noise, inconsistence and incompleteness. Noisy irrelevant, redundant, unreliable and inaccurate data generally lead to generation of absurd results [1, 2, 4]. Consequently, further development of models based on this data will be just waste of resources. So, the raw data is required to pre-process to remove problems such as missing attributes, missing values, outliers, duplicate and wrong data [1, 2, 4, 5]. Data pre-processing is a data mining method to convert raw data into useful and clean information [2]. Data pre-processing includes data cleaning, normalization, balancing, feature extraction, feature selection, data importing, missing value treatment, noise removal [2, 4, 6]. The work is motivated by the fact that development of prediction models on the basis of error free database will result into efficient, relevant and accurate results using data mining [4, 6].

METHODOLOGY

C. Data pre-processing [1, 2]

The following paragraphs explain various data pre-processing methodology steps.

1) Data Cleaning [2]

This is normally the starting step in data pre-processing. Data cleaning involves removing noisy, absurd and inappropriate data, imputing missing values and removing outliers. The following paragraphs explain the data cleaning steps.

Missing Values Treatment [2]

- Imputing Missing Values: The missing values in the data are imputed with suitable value calculated using some imputation technique such as mean, median, mode, KNN etc [2]. In this paper, mean method of imputing missing values is used.
- Dropping/Ignoring an observation/value: Sometimes, it can be wise either to drop or ignore an observation that contains missing data than imputing it [2]. Normally, dropping and ignoring some unknown data does not affect much especially for large datasets. This saves lot of time by avoiding calculation and imputation [2].
- Imputing Predicted Value: It is very important to impute accurate values of the missing parameter [2]. It has been seen that some algorithms such as regression, KNN, Bayesian formulation, decision tree etc. can be used to predict most appropriate, accurate and probable value of the missing parameters and then impute the missing data with calculated value [2].

Removing Noise and Outliers [1, 2, 4]

- Any type of error resulted from unwanted variation in an attribute is called noise. The outliers in data can also be counted towards noise. The noise need to be removed from the data to generate better prediction results. Methods such as visualization and informative methods used to remove are called smoothing methods.
- Binning can also be used to remove noisy data from the dataset. The data is smoothen by sorting the noisy data and outliers based on its neighbor data values. After clustering, similar values of the dataset, some values still remain outside the cluster, these ate termed as outliers and need to be removed.
- Smoothing of the dataset can be done by detecting and removing outliers. Binning method are also used to remove outliers.

2) *Data Transformation [2, 4]*

It is the process of using an appropriate strategy to change the form of the data to make it appropriate in order to extract important information from it. It include aggregation, feature extraction, feature selection, discretization, attribute construction and dimension reduction.

3) *Data Reduction [2, 4]*

Sometimes, it is not feasible to locate and find the required data from a single database. In order to create required dataset, collection of data from various sources may be needed. The data is further merged to form a single dataset called data integration. During data integration process, redundancy occurs may occur which can be due to duplicity of data or addition of unwanted data. Data reduction is the process of removing duplicate data or unwanted data. Data reduction involves reducing the volume of data using some methodology without affect analysis results. Techniques such as editing, sorting, collating, scaling, encoding, clustering, sampling etc are used for data reduction.

C. Features selection

Neighborhood components analysis (NCA) is used for feature selection from the input database. This method is a non-parametric supervised learning method for classifying multivariate data into classes. It is used to select features that improve prediction accuracy of classification algorithms [14]. NCA generally reduces the features dimensionality by discarding irrelevant and duplicate features which in turn improves the algorithm performance [14]. The steps involved in NCA are as follows [14].

1. The dataset was divided into training, testing and validation datasets.
2. Lambda value (λ) was tuned in, and NCA was trained using each fold in training set for every lambda value.
3. The minimum average loss value was computed, and best lambda value corresponding to minimum loss was calculated and the corresponding feature weight for each feature was estimated and compared with the threshold value.
4. The features having feature weight more than the threshold value were extracted, and selected features were trained with Adam deep learning classifier and performance of trained classifiers were evaluated.

In order to improve the accuracy of the developed model, the NCA has been used to select the best features with comparatively higher weights. NCA is used to select the best weighted 12 out of all given features. Fig. 2 shows the feature weight of the features selected using NCA.

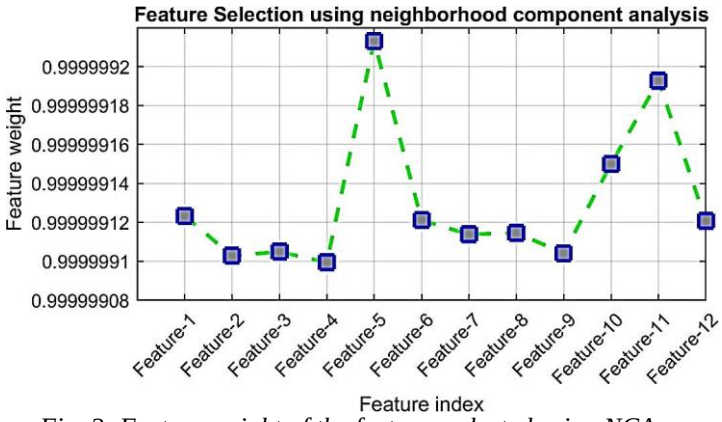


Fig. 2: Feature weight of the features selected using NCA

IMPLEMENTATION

The present implementation has been performed on Intel core i3 1.9GHz Computer having 8GB RAM using MATLAB® software. The Brazilian e-commerce dataset [13] has been used and techniques such as data cleaning, normalization, balancing, feature extraction, feature selection, data importing, missing value treatment, and noise removal are applied on this dataset. The Brazilian dataset is a real time e-commerce public dataset of customer orders and purchases at Olist store. The dataset contains data of approx. one lakh customers of Brazil who purchased online from 2016 to 2018. The dataset contains a number of features viz. order delivery status, payment status, location, city, product type etc. as well as customer reviews.

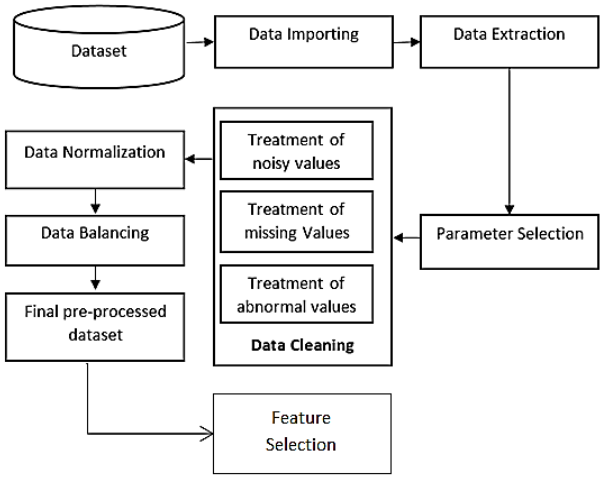


Fig. 3: Proposed Methodology

Flowchart in fig. 3 explains the proposed methodology used in the present work. Data pre-processing operations such as data extraction, data importing parameter selection, data cleaning, missing value computation & data normalization are performed on the dataset.

The steps of the methodology are explained in the following paragraphs.

1. Data Importing: Using this process, the data from the dataset is imported in the form of structures and arrays in the MATLAB environment. TO import the data, MATLAB toolbox has been used. The imported data contains all the attributes of the dataset.
2. Data Extraction: Using this process, the input data types are used to convert data into usable form. The imported data is converted in arrays and matrix form to make it useable for MATLAB toolbox.
3. Parameter Selection: Using this process, the features from the dataset are selected and exported to the MATLAB system. In the Brazilian dataset, there are a number of features in the dataset. The customer unique ID is taken as index value, which is used to import data from dataset into machine learning system.
4. Data Cleaning: In this process, the data is cleaned for any abnormalities, Nan or absurd values. In present implementation, the rows with NaN or absurd values are removed from the dataset. The missing values are imputed using mean method. After extracting the features to be used for prediction, data normalization is also done due to presence of various features.
5. Data Balancing: In this process, the pre-processed dataset is balanced. Resampling technique is used to generate the training dataset after balancing.
6. Feature Selection: In this process, NCA algorithm is applied for selecting the most significant and high weight feature which can give accurate prediction results after the churn prediction model is developed.

RESULTS AND DISCUSSIONS

Fig. 4 shows the GUI of the developed pre-processing approach.

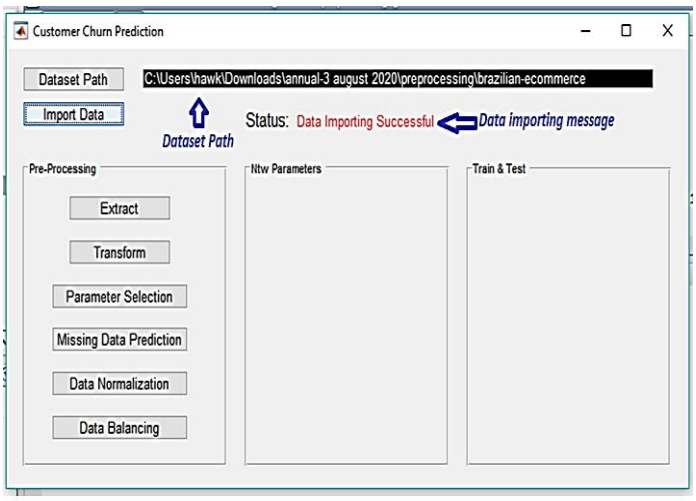


Fig. 4: GUI of the developed pre-processing approach

The developed system works in the following steps.

1. Initially, the dataset path is provided to the developed system. A window is prompted when the dataset path button is pressed. Dataset path is then provided to the system. By this way, the input dataset is loaded in the MATLAB system.

2. By clicking the import data button, the dataset is further imported in the MATLAB system.
3. Extraction is further done after importing the dataset.
4. Further, feature selection is done by pressing on the parameter selection button. Using this process, all the parameters are selected and extracted in the MTALAB system.
5. Thereafter, missing data imputation performed using the mean method by clicking on the missing data prediction button.
6. Further, data balancing has been carried to eliminate the imbalance in the dataset.
7. Finally, NCA algorithm is applied for selecting the most significant and high weight feature which can give accurate prediction results after the churn prediction model is developed.

CONCLUSIONS AND FUTURE SCOPE

In this paper, feature selection method for an e-commerce dataset has been attempted. The feature selection was done from the pre-processed data for purpose of development of customer churn prediction using data mining. The neighborhood component analysis technique was applied for feature selection from the benchmarked Brazilian e-commerce dataset. Before applying NCA, the dataset was pre-processed for any missing value imputation, data cleaning, data normalization and data balancing etc. Important features based on their weight are selected to develop customer chum prediction data mining classification models in e-commerce using machine learning and deep learning approaches. In future, authors intend to work for development of customer churn prediction model using data mining based on the features selected.

References

[1]. I. A. A. Sabri, M. Man, W. A. W. A. Bakar and A. N. M. Rose. Web Data Extraction Approach for Deep Web using WEIDJ. Procedia Computer Science, Vol. 163, pp. 417-426, 2019.
[2]. P.N. Tan, M. Steinbach and V. Kumar. Data Mining: Concepts and Techniques, 2nd ed., The Morgan Kaufmann Publisher, Elsevier Inc., 2016.
[3]. P. Berger and M. Kompan. User modelling for churn prediction in E-commerce. In Proc. Intl. Conf. on Intelligent Systems, pp. 1-6, 2019.
[4]. F. Bahari and M. S. Elayidom. An efficient CRM-Data mining framework for the prediction of customer behavior. Procedia Computer Science, Vol. 46, pp. 725 – 731, 2015.
[5]. H. Kumar and R. K. Yadav. Rule-based customer churn prediction model using artificial neural network based and rough set theory. In: M. Pant, T. Sharma, O. Verma, R. Singla and A. Sikander (eds) Soft Computing: Theories and Applications. Advances in Intelligent Systems and Computing, Vol. 1053, pp. 97-108, 2020.
[6]. A. Khede, A. Pipliya and V. Malviya. A novel approach for predicting customer churn in telecom sector. In: R. Shukla, J. Agrawal, S. Sharma, N. Chaudhari and K. Shukla (eds) Social Networking and Computational Intelligence. Lecture Notes in Networks and Systems, Vol. 100, pp. 295-304, 2020.
[7]. N. Jha, D. Parekh, M. Mouhoub and V. Makkar. Customer segmentation and churn prediction in online retail. In: C. Goutte, and X. Zhu (Eds.) Canadian Conference on Artificial Intelligence, Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence, Vol. 12109, pp. 328-334, 2020.
[8]. A K. Ahmad, A. Jafar and K. Aljoumaa. Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, Vol. 6 (28), pp.1-24, 2019.
[9]. J. Vijaya and E. Sivasankar. An efficient system for customer churn prediction through particle swarm optimization based feature selection model with simulated annealing. Cluster Computing, pp.1-12, 2017.
[10]. X. Yu, S. Guo, J. Guo and X. Huang. An extended support vector machine forecasting framework for customer churn in e-commerce. Expert Systems with Applications, Vol. 38, pp.1425–1430, 2011.
[11]. M. Pondel, M. Wuczyński, W. Gryncewicz, Ł. Łysik, M. Hernes, A. Rot and Agata Kozina. Deep learning for customer churn prediction in ecommerce decision support. Intl. Conf. on Business Information Systems, pp.3-12, 2021.
[12]. G. Bunaccors. Machine Learning Algorithms. 1st ed., The Packt Publishers, Birmingham UK, 2017.
[13]. Olist (2020) Brazilian E-Commerce Public Dataset [Online]. Available: https://www.kaggle.com/olistbr/brazilian-ecommerce.
[14]. N. Kavya, N. Sriraam, N. Usha, D. Sharath, B. Hiremath, M. Menaka and B. Venkatraman. Feature Selection Using Neighborhood Component Analysis with Support Vector Machine for Classification of Breast Mammograms. In: V. Bindhu, J. Chen, J. Tavares (eds) International Conference on Communication, Computing and Electronics Systems. Lecture Notes in Electrical Engineering, 637, pp.253-260, 2020.

CONVERGENCE ANALYSIS OF A 3- NODE SDN CLUSTER

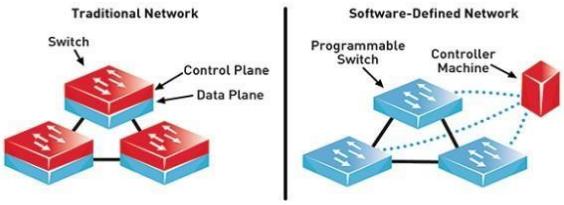
Avtar Singh, Navjot Kaur, Harpreet Kaur
Punjabi University, Patiala,Punjab, India
Pu.avtarsingh@gmail.com, navjot_anttal@yahoo.co.in, khasria.harpreet@gmail.com

ABSTRACT: SDN is revolutionizing the network industry with plethora of benefits it brings along with it. There are various SDN related technologies like SD-WAN,SD-Access, SD-Security etc. There are emerging technologies like 5G and Cloud based data centers which are using SDN based networks to large extent. With SDN, networks are mainly controlled using a centralized controller, so there is a need for them to be highly available and secure. Clustering is a solution which can be used to bring reliability and availability. In this paper, we have made a 3-node controller cluster using Linux Foundation's Opendaylight. We have deliberately shutdown the primary controller or leader and convergence time is checked on the basis of per-packet using Wireshark packet analyzer, that provides per packet analysis.

KEYWORDS – SDN, ODL, Akka, Shards, Module-Shards, ODL Clustering, EastBound, WestBound

INTRODUCTION - SDN mainly works on the basis of centralized controller based network. This centralized architecture where a controller acts like a brain and all other devices acts as data plane or muscles of the network and are used to take data from one device to other from source to destination is not actually new. In Wireless Networks, Access Points are integrated with the centralized Wireless controller and from that wireless controller, all the configuration policies are pushed. SDN is also like this, but in much broader term, as in the SDN, total control of how the packets should be transmitted over the network is made by the SDN controller[6] running a protocol named as Openflow between the controller and the white box switches. Figure showing a difference between traditional networking architecture compared with the SDN architecture is shown below:-

Figure 1.- Traditional Networking vs SDN [7]



Traditional Routers and Switches have their own control and data plane. SDN brings the paradigm[17] shift and decoupled the data plane from control plane. With SDN, all the control plane[16] work is done by a controller and all the other devices acts as a data plane and takes instructions from the controller. [8] In a small to medium sized network, a single controller can work or for redundancy, multiple controllers can be used. In a large scale environment, redundant controllers are a must. Another good reason to migrate from traditional networks to SDN is the cost factor. For example, a company that has 500 traditional network devices(routers and switches) with control plane and data plane capability can cost much more than another company using two controllers and all other white box switches with just data plane ability taking control plane instructions from controllers. It enables simple programmatic control in the networking.

SDN ARCHITECTURE - SDN was developed to facilitate innovation and enable simple programmatic control of the network data-path. A figure showing SDN Architecture is shown below :-

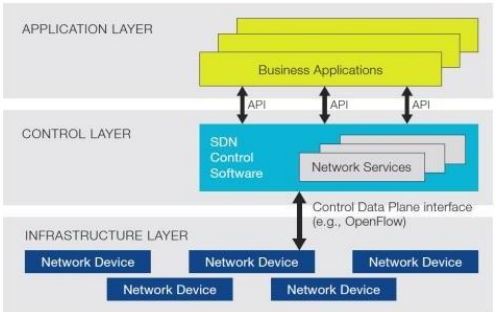


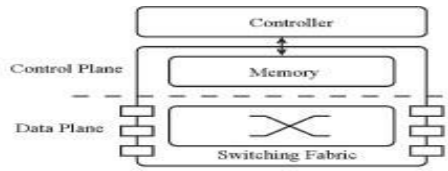
Figure 2 - Software Defined Networking Architecture [9]

As above figure shows, SDN reference model mainly consists of three layers i.e. Application, Control and Infrastructure Layer.

Infrastructure Layer -This is the bottom layer of SDN and it contains switching devices which mainly act as a data plane in the network and these network devices are controlled by a controller. All the packet processing and network management is

done by the controller which is acting as the brain in the SDN layer architecture. This layer has white-box programmable switches installed which only listens to the instructions provided to them from the controller and paths towards destinations are decided on the basis of controllers. Some of the popular vendors in this kind of switches market are Pica8, Big Switch Networks, Dell etc. Apart from them some big network vendors also provides the SDN based Switches like Cisco, Juniper, Arista, HP etc. Devices that comes under this layer are connected using different types of media which can be copper, fiber optics and wireless etc.One of the most important part of this layer is Switching Devices. Below is the switching device model in Software Defined Networking :-

Figure 3 - SDN Architecture [9]



Above figure shows the architectural design of an SDN based Switch that consists of two logical components i.e. Control Plane and Data Plane. Control Plane defines the path and control the flows and provide instructions to data plane. Control Plane is the brain while data plane is the muscle. Data Plane gets the instructions [7-8] from the controller or control plane and does actual packet forwarding. In data plane processor plays an important role in packet forwarding.Various example of Network Processors are Broadcoms XLP Processor Family, Intels XScale Processor that uses ARM Architecture, Netronomes NFP Series that uses ARM Architecture etc. **Control Layer**- Control Layer acts as a bridge between Application and Infrastructure Layer with the help of two interfaces i.e. North-Bound and South-Bound Interface. North-Bound Interface makes interaction of control layer with the application layer, it gives service access points in different forms like an Application Programming Interface(API)[11-12]. The applications which are running under SDN can access network information status gathered from the switches using this API. It is possible that control layer consists of multiple SDN controllers for redundancy in case of large network domains and then the east-west communication interface is used to sync different controllers with each other so that one can act as primary and other one as backup and make the decision making process according to the network design needs. South-Bound interface is used to make communication between controller and bare-metal or white-box switches.OpenFlow is one of the most used and popular SDN protocols, which makes interactions between Controller and Data Plane Switches possible.

Application Layer - Application Layer contains various SDN applications which were present to meet the user's requirements. It is connected with the control layer using the North-Bound interface[14]. A programmable platform is provided by the control layer. The applications running in this layer access the devices at the infrastructure layer[15]. Some of the example that comes under this layer are like server load balancing, network virtualization etc.

ODL Clustering and Convergence This section is about the results and screenshots taken from SDN Clustering where we have used three controllers connected with each other in VMWare Workstation Hypervisor and connected using VMNet 0 Bridge Interface. As all the controllers are connected using a single VMNet 0 interface, they behave like they are connected in a Local Area Network. Interface with which I have bridged the Controllers is my Wifi Interface and I have assigned following IP addresses to the Controllers:

OpenDayLight Controller A – 192.168.43.138 OpenDayLight Controller B – 192.168.43.245 OpenDayLight Controller C – 192.168.43.140

As shown figure 4, three controllers are connected with each other using a Layer 2 Switch and all having IP addresses of the same subnet and are able to communicate with each other. All are connected in the data center and used as a clustered controller in the data center that helps in redundancy as in case the leader goes down, another controller can become leader and makes the decision making. We made a 3-Node Cluster which provides better cluster redundancy than a 2- Node cluster. We can configure clustering by configuring akka and module-shards wither by using a shell script or in manual manner. Below is the call to script we can use by running it under the extracted directory of odl distribution

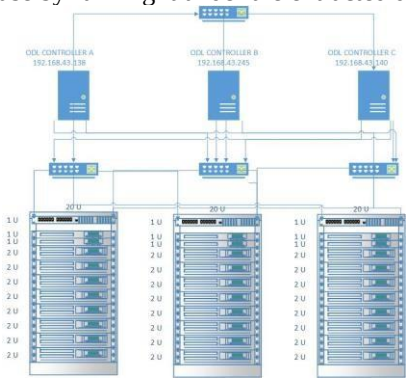


Figure – 4 Node ODL Clustering for DC

bin/configure_cluster.sh 2 192.168.43.138 192.168.43.245 192.168.43.140

Command shown above configures member 2 i.e. 192.168.43.245(ODL B) of a cluster which is made of 192.168.43.138 192.168.43.245 192.168.43.140

After configuring of files for clustering, we can test the outputs if our clustering has started to work by using different inputs. Below is the output in JSON format under the Yangman GUI module page, where Local Shards will be shown in JSON output that shows which shards are operational:


```
{
    "request": {
        "mbean": "org.opendaylight.controller:Category=ShardManager,name=shard-manager-
operational,type=DistributedOperationalDatastore",
        "type": "read"
    },
    "value": {
        "LocalShards": [
            "member-1-shard-default-operational",
            "member-1-shard-topology-operational",
            "member-1-shard-entity-ownership-operational",
            "member-1-shard-inventory-operational",
            "member-1-shard-toaster-operational"
        ],
        "SyncStatus": true,
        "MemberName": "member-1"
```


Figure 5 – Yangman JSON output showing mbeans and operational shards.

Above output clearly shows all the shards are operational and Sync Status is also true in the JSON output shown under Yangman by asking for a JSON output using a GET request for mbeans and operational shards of a local member.

It shows the STATUS as 200 Ok, which means that request is successful. If the Code is 404, then it means that the page requested is not found and jolokia needs to be installed apart from clustering. If Code is 503, it means that the Internal Server Error is found, which is needed to be resolved in fast manner. Jolokia is needed to be installed in order to check if the leader is selected or not, who is the leader? Or if the peers have been selected or not etc. Below figure shows the output that we have received for OpenDayLight Controller A having IP address 192.168.43.138 after configuring the cluster :

Figure 6 – ODL Clustering output of Jolokia Package for ODL A

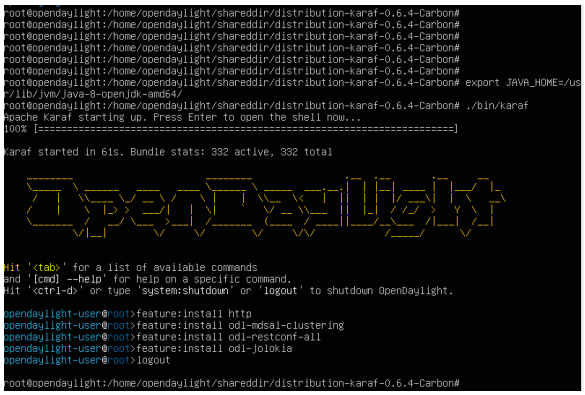
As shown above in the output of ODL A, it is clear that ODL A is Member 1 and is selected as Leader of the Cluster and the shards are replicated with ODL B and ODL C. It also shows the various ODL controllers which are acting as its peers and shares the shards. Another output of Controller B is shown below:

Figure 7 – ODL Clustering output of Jolokia Package for ODL B

Output above shows that ODL Controller B has also selected ODL A as its leader of the cluster and the peers of ODL B are ODL A and ODL C. Sync Status is also True in the key:value pair of the output. Another important part of the output is HTTP Status 200, which means file is working fine.

Figure 8 – ODL Clustering output of Jolokia Package for ODL C

Last output shown above shows the key:value pairs of ODL Controller C which is shown under jolokia requesting the parameters of json and html. It also shares ODL A as the leader of the cluster and peers of this controller are ODL A and ODL B. Status is also 200 which means HTTP OK. All the shards are replicated with ODL A and ODL B and last sync status is true in this jolokia output of ODL Controller C.



As having a cluster in the network brings redundancy in the controllers, so in case leader controller goes down, another peer automatically takes over the role of the leader and cluster remains in the working condition with a difference that a three node cluster becomes a two node cluster. Cluster in ODL does not have the capability of preemption, that means if the leader in the cluster goes down, and peer becomes new leader, and after some time, leader that had gone down comes back up, it does not become the leader straightway and other cluster members needed to be logged off in order to make them

leader. We have ODL A working as a leader and when we logged off the ODL A controller, ODL B automatically becomes the leader as it has the second highest up time after ODL Controller A. Figure below shows the ODL Controller A logged out:

Figure 9 - Logging out of ODL Controller A in order to check if Clustering is working.

In the above output, we have logged out of Cluster and also from ODL Controller A. If everything goes right, then Clustering will not break and ODL B should become cluster which it did as shown in the below figure:

Figure 10 – ODL Convergence in Cluster after the primary node of cluster i.e. Leader goes down.
Above output clearly shows that cluster is in working condition and ODL B controller becomes the new leader. It shows that it has two peers i.e. ODL A and ODL C and also the sync status is true. Cluster is now working properly and leadership and shards replication is working fine. Time it takes to converge from one controller to another controller for leadership change is also under a second, which is quite fast.
When we compared this with Suh. Dongeun[1],where they have a convergence time of around 42seconds for a 1-25 switch network connected with the clustered controller network with HP VAN SDN Controller used by them, we used OpenDayLight Clustering and our results are much better with or without clustering, i.e.

No.of Switch es	OpenDaylight Clustering(Converg nce time in seconds)	HP VAN SDN Controller and ONOS Clustering(Converge nce time in seconds)	Without Clustering – ODL(Convergen cetime in seconds)	Without Clustering – HP VAN SDN and ONOS(Converge nce time in seconds)
30	15	41	35	40
40	17	49	37	43
50	20	53	39	48

Table 1– Comparison and Result Table SDN Convergence – With and Without Clustering

Above table clearly shows that our results with ODL clustering and without clustering are much better than using HP VAN SDN Controller or ONOS clustering or without clustering. When comparing 30, 40 and 50 switch topology and convergence times, we have a much better convergence and reduction in delay with just 15, 17 and 20 seconds with 3-node clustering and 35, 37 and 39 for without clustering. We have reduced the convergence time to more than half of the convergence time[1]. Below graphs also depicts the convergence and flow table setup time with or without clustering.

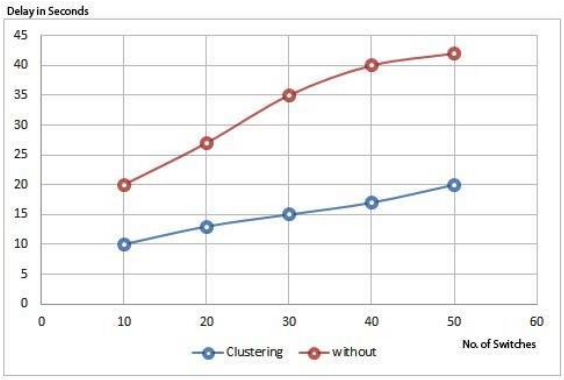


Figure 11 – Convergence and Flow table setup with and without ODL Clustering

CONCLUSION

SDN brings the new era in the networking and telecom industry with SDN and SD-WAN. It has shifted the network industry by changing the way network infrastructure works with controller based network with a single controller controlling the full network in a Data Center and also controller controls the control plane in SD-WAN network which is mainly used for Data Center interconnections and for multiple enterprise sites connected with each other. Problem mainly arises when we have a single controller which goes down due to some illegal activity by hackers like a DDoS attack or some hardware of any other issue. If this happens, it can destroy the network. Clustering of controllers is used in medium to large scale networks and their communication is mainly done at EastBound and WestBound interfaces with both these interfaces are connected with each other and sharing the necessary data with each other. Major files which controls the ODL clustering is akka and module-shards, which can be configured either via shell script or manually. A leader is selected from the cluster which shares the shards between all the cluster seed nodes. In case the cluster leader goes down, another peer becomes the leader and the flow table setup takes around 15 seconds when our topology have 30 interconnected switches, 17 seconds when we have 40 interconnected switches and 20 seconds in case of 50 interconnected switches. Convergence time is becomes slightly higher with more number of switches added in the controller but it's the starting era of SDN and continuous research will help it improve its convergence time be more better.

FUTURE SCOPE

SDN is still in its early stages and companies have started to migrate and adopt this new paradigm shift of network industry. Different type of solutions be it hybrid or purely SDN based solutions are coming out and the best thing about SDN based networks is that they are much more flexible as compared with traditional networks. New research issues are coming issues are coming out regularly as its relatively new technology and new solutions are also coming out. In the future, my research can be extended for much reduced convergence and flow table setup time and also with the load balancing at the control-to-infrastructure layer with both SDN and SD-WAN over the cloud networks which can be inside a data center and interconnecting multiple data centers.

References

1. Abdelaziz A, Fong AT, Gani A, Garba U, Khan S, Akhunzada A, et al.(2017) Distributed controller clustering in software defined networks. PLoS ONE 12(4):e0174715.
2. Suh. Dongeun, Jang. Seokwon, Han. Sol, Pack. Sangheon, Kim. Taehong," On performance of OpenDaylight clustering" IEEE, pp.407-410, June 2016.
3. Taehong Kim, Seong-Gon Choi, Jungho Myung, Chang-Gyu Lim," Load balancing on distributed datastore in OpenDaylight SDN controller cluster"IEEE, 2017.
4. D. Kreutz, F. Ramos, P. Verissimo, C. Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey,"Proceedings of the IEEE, vol. 103, no. 1, pp. 14-76, January 2015.
5. Alexander Gelberger, Niv Yemini, Ran Giladi," Performance Analysis of Software- Defined Networking (SDN)"IEEE,2013.
6. Scott Shenker, Martin Casado, Teemu Koponen, Nick McKeown, et al. The future of networking, and the past of protocols. *Open Networking Summit*, 20:1–30, 2011.
7. Open Networking Foundation. Software-defined networking: The new norm for networks. *ONF White Paper*, 2:2–6, 2012.
8. Software defined networking. Open Networking Foundation.
9. Software defined networking, big switch networks.
10. Software defined networking, microsoft.
11. Traditional networking vs sdn.
12. Radia Perlman, Anoop Ghanwani, Donald Eastlake 3rd, Dinesh Dutt, and Silvano Gai. Routing bridges (rbridges): Base protocol specification. 2011.
13. Sdn architecture.
14. Software defined networks. Cisco.

15. Wenfeng Xia, Yonggang Wen, Chuan Heng Foh, Dusit Niyato, and Haiyong Xie. A survey on software-defined
16. networking. *IEEE Communications Surveys & Tutorials*, 17(1):27–51, 2015.
17. Mr. Sachin Acharya T Mr. Nithin Kumar, Ms. Nidhi K N. A survey on sdn: An unprecedented approach in networking. *IJECS Volume-5 Issue-2*, 2016.
18. IDC Predictions. Competing on the 3rd
19. platform. *IDC-Thu, 29 Nov 2012*, 2013.

A REVIEW ON MAMMOGRAPHY BASED APPROACHES FOR BREASTCANCER DETECTION AND DIAGNOSIS: CADX SYSTEM

Navneet Kaur[a], Lakhwinder Kaur[b], Sikander Singh Cheema[c]
[a,b,c]*Department of Computer Science and Engineering,* [a,b,c]*Punjabi University, Patiala, India.*
[a]*navneetmavi88@gmail.com*
[b]*mahal2k8@gmail.com*
[c]*cheemasikander8@gmail.com*

ABSTRACT- This paper discusses the preprocessing and segmentation methods used in past few years. These methods have evolved in past time vastly for breast cancer detection in mammographic images. Preprocessing is generally removal of unwanted things from the image and segmentation is grouping of pixels to form a recognizable patternin an image to have informative advantages out of it for the user. In this paper, we have tried to study numerous methods of preprocessing like image filters and enhancement methods. Also for segmentation methods like region based, edge based and thresholding based are discussed. Also comparison of major sectors in the above methods on the basis of spatial information, region continuity, speed, automaticityand accuracy is done.

KEYWORDS: *Breast Cancer, Mammography, CADx system, Preprocessing, Segmentation*

I. INTRODUCTION

Breast Cancer is the leading cause of deaths of women in all over the world and it happens to more than 8% of women in their lifetime [30]. The combination ofmedical science and technology contributes a lot to human health and their quality of life. It is very common among women while rare among men. Breast cancer commonly affects women more than 40 years of age however younger women can alsobe affected especially with genetic predisposition (a genetic characteristic that influences the development of an individual organism under the influence of environmental conditions). It arises from the breast tissues mostly from the ductal carcinoma (the inner lining of milk ducts) or less frequently from the lobular carcinoma (the lobulesthat supply milk to the ducts). The risk factors for breast cancer are age, genetics, obesity, family history or late pregnancy [21].Due to the factors related to cost and professional experience, in the last two decades computer systems to support detection and diagnosis have been developed in order to assist experts in early detection of abnormalities in their initial stages. Despite the large number of researches on computer-aided systems, there is still a need for improved computerized methods [10]. The most crucial parameter for newly diagnosed breast cancer is staging and prognosis (chance of recovery). Staging is done at the time of diagnosis and for this, TNM parameters are considered that includes tumour size (parameter T), whether or not the tumour has spread to auxiliary lymph nodes (parameter N), and whether or not the tumour has spread to a more distant part of the body (parameter M). At first stage, the mass is lesser than 2 cm in diameter and there is no lymph node; At second stage, the mass diameter is between 2cm to 5 cm, and/or there are lymph nodes under the armpit. At third stage, the mass diameter is larger than 5 cm and there are lymph nodes under the armpit. And at the fourth stage, the mass may be of any size and the lymph nodes in the armpit are often impacted; the cancer spreadsto other parts of the body.

A. Male Breast Cancer

Breast cancer in men does not occur very often, less than 1% of all breast cancersoccur in men. For men, the risk of being diagnosed with breast cancer during lifetimeis about 1 in 1,000. Men and women all have breast tissues. The various hormones in women's bodies exhilarate the breast tissue to grow into full breasts while men's bodies normally don't form much of the breast-stimulating hormones. As a result, their breast tissue usually stays flat and small. A number of factors can increase the risk of man getting breast cancer [32]:

- High estrogen levels
- Old Age
- Klinefelter syndrome
- Radiation exposure
- A strong family history of breast cancer or genetic alterations

B. Types of Tumours

Although breast cancer is often referred to as one disease, there are actually many different types of breast cancer. All breast cancers start in the breast, so they are alike in some ways, but differ in others. Tumours are broadly categorized as:

- Non-Invasive
- Invasive

Non Invasive - DCIS (Ductal carcinoma in-situ) means the abnormal cells are contained in the milk ducts of the breast and have not dispersed to the nearby breast tissue. Although it is non-invasive but over time without treatment these abnormal cells could develop into the invasive breast cancer. Invasive – In this, the cancer spreads from either the milk ducts or the lobules into the nearby breast tissue or thelymph nodes or the other parts of the body. Due to his

reason, the invasive breast cancers have a poorer prognosis. The proportion and characteristics of each invasivebreast cancer is shown in tabular form below:

C. Imaging Modalities

To the human observer, the internal structure and functions of human body are generally not visible. So using various technologies, images are created through which the medical professionals can look into the body to diagnose abnormal conditions and guides therapeutic procedure. Identification of optimal imaging modality for surveillance imaging remains significant challenge. Various imaging modalities currently used in medical care facilities are:

- Mammography
- Ultrasonography
- Magnetic Resonance Imaging
- Positron Emission Tomography
- Computed Tomography

TABLE I

The proportion and characteristics of each invasive breast cancer is shown in tabular form below:

Type	Proportion	Characteristics
Invasive Ductal Carcinoma (IDC)	50-75%	1. Hard tumour texture 2. Irregular shape 3. Cell features vary 4. DCIS often present
Invasive Lobular Carcinoma (ILC)	5-15%	1. Normal, slightly firm on hard tumour texture 2. Cells appear in single file order
Mucinous Carcinoma (MC)	1-5%	1. Soft tumour 2. No palpable tumour 3. Cells are surrounded by excess mucous (mucin)
Tubular Carcinoma (TC)	1-5%	1. Tumours are often small in size (about 1cm or less) 2. Often no palpable tumour 3. Made up of tube like structures called “tubules”
Invasive Papillary Carcinoma	Less than 1%	1. Soft tumour 2. Cells appear as finger like branches.
Carcinoma with Medullary features	Less than 1%	1. Soft tumour 2. Cells have a sheet like appearance 3. Fleshy mass that resembles a part of the brain called the medulla 4. looks like aggressive and are highly abnormalcancer cells, but they don’t act like them 5. doesn’t grow quickly and usually doesn’t spread outside the breast to the lymph nodes

The literature shows that the only radiological technique that has significant impact on the diagnosis, staging and patient follow-up in the case of screening is low-dose mammography. Mammography is the only reliable screening test proven in breast imaging. Although it is an effective screening tool, it does have limitations, particularly in women with dense breasts. Mammography is the most widely used modality for detecting and characterizing breast cancer. It is a medical imaging that uses low-dose X-ray system to see inside the breasts. A mammography exam, called a mammogram, helps in the early detection and diagnosis of breast diseases. It has high sensitivity and specificity due to which small tumours and microcalcifications can be detected on mammograms. During mammography two views of each breast [10] are recorded as shown in Figure 1:

- Craniocaudal (CC) view: It is one of the two standard projections in a screening mammography. It is a top to bottom view that must show the medial part as well the external lateral portion of the breast as much as possible.
- Mediolateral Oblique (MLO) view: It is a side view taken at an angle. The representation of the pectoral muscle on the MLO view is a key component in assessingthe adequacy of the film. The amount of breast tissue included in the image is the
- amount of visible pectoral muscle, resulting in a good quality factor that is very important for reducing the number of false negatives and increase the sensitivity of mammography.

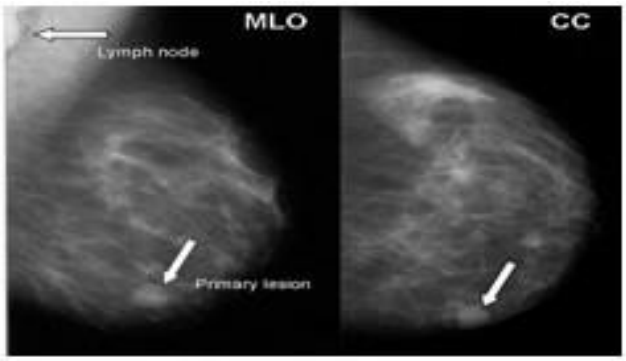


Fig..1 MLO and CC view of a mammogram

D. Types of Abnormalities

Radiologists visually search mammograms for specific abnormalities.Some of the important signs of breast abnormalities that radiologists look for are:

1) Calcifications
2) Masses
3) Architectural distortions

1) *Calcifications:* Calcifications are tiny mineral deposits (calcium) scattered throughout the mammarygland, or occur in clusters. They appear as small bright spots on the mammogram. They are characterized by their type and distribution properties. There are some characteristics of microcalcifications that help in deciding that whether it is benign or malignant. Calcifications that are very small in size (between 0.5 mm to 2 mm or even smaller than 0.5 mm) and concentrate in one place are in favor of malignancy. While the calcifications that are bigger that 2 mm are macrocalcifications and are usually benign. In context of shape, the calcifications of non- uniform, irregular and angular shape are malignant as compared to rounded, regular and uniform benign calcifications. Concentration of microcalcifications in one place is also in favor of malignancy while the scattered calcifications are usually benign [25]. When found on a mammogram, a radiologist decides whether the specks are of concern or not. Commonly, they simply indicate the presence of tiny benign cysts, but can signify the presence of early breast cancer. Two major types of calcifications found in breasttissue [17] are:

- *Macrocalcifications*: Macrocalcifications are coarse (larger) calcium deposits that are most likely due to changes in the breasts caused by aging of the breast arteries, old injuries, or inflammation. These deposits are related to non-cancerous conditions and do not require a biopsy.
- *Microcalcifications*: Microcalcifications are tiny specks of calcium in the breast as shown in Figure 2. Microcalcifications seen on a mammogram are of more concern than macrocalcifications, but they do not always mean that cancer is present. The shape and layout of microcalcifications help the radiologist judge how likely it is that cancer is present. In most cases, the presence of microcalcifications does not mean a biopsy is needed. But if the microcalcifications have a suspicious look and pattern, a biopsy will be recommended. (During a biopsy, the doctor removes a small piece of the suspicious area to be looked at under a microscope. A biopsy is the only way to tell if cancer is really present.

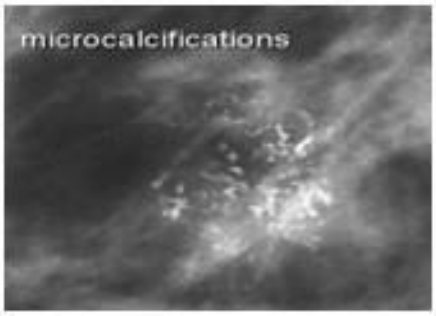


Fig..2 Microcalcifications

2) *Masses*: A mass is defined as a space occupying lesion as shown in Figure 3. A mass, with or without calcifications, is another important change seen on a mammogram. Massesare areas that look abnormal and they can be many things, including cysts (non- cancerous, fluid-filled sacs) and non-cancerous solid tumors (such as fibroadenomas) [30], but may sometimes may be a sign of cancer.

- Cysts: They can be simple fluid-filled sacs (known as simple cysts) or can be partially solid (known as complex cystic and solid masses). Simple cysts are benign (not cancerous) and don't need to be biopsied. If a mass is not a simple cyst, it is of more concern and might need to be biopsied to be sure it isn't cancer.
- Solid Tumours: If a mass is not a simple cyst (that is, if it's at least partly solid), more imaging tests may be needed. Some masses can be watched with regular mammograms or ultrasound, while others may need a biopsy. The size, shape, and margins (edges) of the mass may help the radiologist determine if cancer is likely to be present.A cyst and a tumor can feel the same on a physical exam. They can also look thesame on a mammogram. To confirm that a mass is really a cyst, a breast ultrasoundis often done. Another option is to remove (aspirate) the fluid from the cyst with athin, hollow needle.

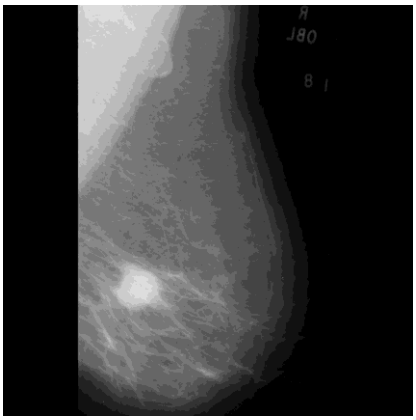


Fig..3 Masses

3) *Architectural Distortions:* The normal architecture is distorted with no definite mass visible as shown in Figure 4. This includes spiculations radiating from a point, and focal retraction or distortion of the edge of the parenchyma. It appears as a distortion in which surrounding breast tissues appear to be "pulled inward" into a focal point. While architectural distortion is a localizing sign of cancer, a surgical scar, fibrocystic change, and in some cases,the superimposition of breast tissues may give the same appearance.

Fig. 4 Architectural Distortions

E. Computer Aided Diagnosis and Detection Systems (CADx Systems)

Computer Aided Diagnosis (CADx) systems shown in Figure 5 evaluate the conspicuous structures. Computer-aided detection systems are widely used in mammography, where signs of breast cancer are often very low. It acts as a tool that enhanceshuman film-reading accuracy and classifies the signs of breast cancer as malignant or benign.

Benign: They are not cancerous. They:

- can usually be removed
- do not come back in most cases
- do not spread to other parts of the body and the cells do not invade other tissues

Malignant: They are cancerous. They: can invade and damage nearby tissues and organs metastasize (cancer cells break away from a malignant tumor and enter the blood-stream or lymphatic system to form secondary tumors in other parts of the body)

Mammograms are therapeutic pictures that are hard to decipher, in this manner a pre-preparing stage is required with the end goal to enhance the picture qualityand make the division results more precise. The initial step includes the expulsion of antique and undesirable parts out of sight of the mammogram. At that point, an up- grade process is connected to the advanced mammogram. Picture upgrade activities can be utilized to enhance the presence of pictures, to take out clamor or mistake, orto highlight certain highlights in a picture. There are reasons for the need of image pre-processing:

- improvement of image quality to meet the requirements of physician
- noise reduction
- contrast enhancement
- correction of missing or wrong pixel values
- elimination of acquisition-specific artifacts

Scanned images usually suffer from more artefacts than digital images shown in Fig-ure 6. MIAS and DDSM databases contains scanned film mammograms and so theyrequire pre-processing. The various types of artifacts [15] found are:

- Presence of duct tape which reduces intensity of breast tissue and makes segmentation by a threshold impossible.
- Unknown position and orientation of breast requires accurate registration that provides equal position and rotation to all the images.
- Orientation Tag
- Low Intensity labels
- Scanning Artifact

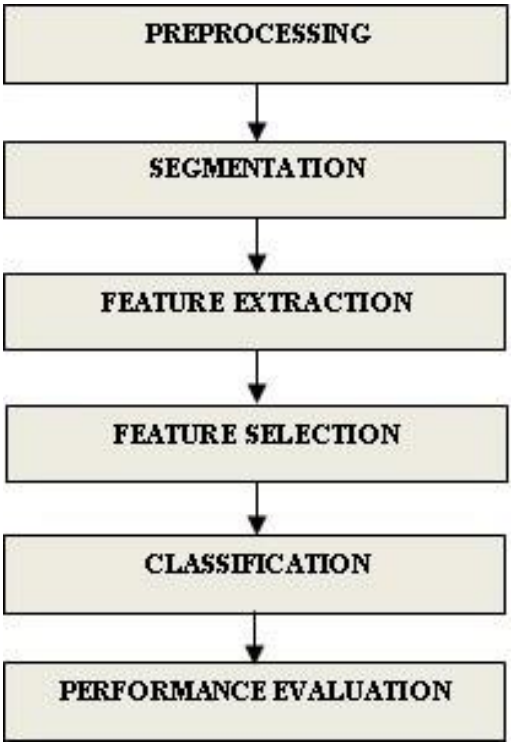


Figure 5: Block Diagram of CADx system

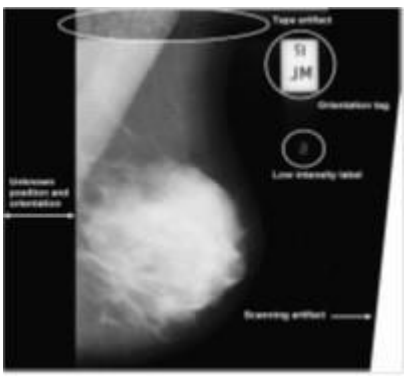


Fig. 6 Breast Image showing all types of artifacts.

Image segmentation is the parceling of a variety of estimations based on homogeneity. To be more correct, segmentation is the segmentation of a image into spatially ceaseless, disjoint and homogeneous areas. Segmentation is great and it has been recommended that image investigation prompts important questions just when the image is sectioned in 'homogenous' regions or into 'moderately homogeneous territories. The last term reflects better the 'close decomposability' of normal frameworks as spread out and we unequivocally address a specific staying inner heterogeneity. The key is that the interior heterogeneity of a parameter under thought is lower than the heterogeneity contrasted and its neighboring territories. Customary image segmentation strategies have been generally separated into three methodologies: pixel, edge and area based segmentation strategies. Pixel based techniques incorporate image thresholding and segmentation in the element space. Based on our survey and the study conducted we have classified the segmentation techniques in way which includes the primitive techniques like the object based and layer based as well as the most modern edge based or region based and ultra modern hybrid techniques like Fuzzy logic, generic algorithm etc. as seen the Figure 7. These techniques can be looked upon from different point of view in term of their performance on the basis of parameters such as spatial information, region continuity, speed, automaticity and accuracy. Several segmentation techniques are used by various researchers to carry out segmentation in mammograms.

II. PREVIOUS WORK

Many researchers have presented various techniques for computer aided detection and diagnosis of breast cancer. Few of them are discussed below:

Yang et. al. [2] told that these are the traditional methods but with time theneed and aura of segmentation has changed to whole different form with hybrid models coming into play. The hybrid model of segmentation has made it possible for the research to dive into many unexplored field that were there in the past. Satellite image analysis and medical image processing are such two fields. Now a days the classification has changed to neighbor labeling pixel based method with latest classifiers and the generative probabilistic model that composites the output of a bank of object detectors in order to define shape masks. Mammography is considered the best strategy for early recognition of breast diseases. In any case, it is troublesome for radiologists to distinguish microcalcification groups. In this way, we have built up a mechanized plan for identifying microcalcification bunches in mammograms. Mammography is the most utilized symptomatic system for breast cancer detection. Microcalcification bunches are the early indication of breast cancer and their initial location is a key to increment the survival rate of ladies. The presence of microcalcification groups in mammogram as little confined granular focuses, or, in other words recognize by radiologists as a result of its small size.

Ramani et. al. [3] showed that the initial step includes the expulsion of curio and undesirable parts out of sight of the mammogram. At that point, an improvement process is connected to the advanced mammogram. Picture upgrade tasks can be utilized to enhance the presence of pictures, to dispense with commotion or blunder, or to emphasize certain highlights in a picture.

Bandyopadhyay et. al. [4] have looked at the reproduced yield parameters, for example, picture quality, mean square mistake, Peak flag to clamor proportion, basic substance and standardized total mistake. The correlation of four sorts of channels are tried for 322 mammogram images(MIAS), from the gross function output, they come to a conclusion that image quality play a vital role in decidingthe preprocessing steps and output. The objective of this exploration is to display a calculation, which helps the radiologists recognizing breast tumors at their beginning periods.

Ibrahim et. al. [5] considers a calculation for the programmed expulsion of relics and clamor that are available in mammogram pictures utilizing morphological tasks, and afterward it upgrades complexity of mammogram pictures utilizing the Band Limited Histogram Equalization (BLHE) strategy for less demanding discovery ofinjuries or tumors. In the wake of preprocessing of mammogram pictures, this calculation sections the pictures utilizing Otsu's N thresholding strategy to identify the locale of enthusiasm for mammogram images. Charate et. al. [6] showed many filtering method have also been discussed and used in different works like mean filtering, median filtering, adaptive filtering and Wiener filtering for the purpose of preprocessing. A mammogram can recognize strange territories in the breast that resemble a growth however ends up being ordinary, this prompts false positive. Mammogram pictures are observed to be hard to translate so a CAD is turning into an inexorably critical apparatus to help radiologist in the mammographic injury elucidation. Preprocessing was considered as an essential advance in mammogram picture investigation.

George et. al. [7] told that precision of preprocessing will decide the achievement of the rest of the procedure, for example, division, characterization and so forth. In this paper, mean, middle, versatile middle, Gaussian and wiener de-noising channels are utilized to evacuate salt and pepper, spot and Gaussian commotions from a mammogram picture and these channels were analyzed dependent on the parameters, for example, PSNR, MSE and SNR to figure out which channel is better to expel these clamors in mammogram pictures.

Bezdek et. al. [8] started with an examination of different image handling procedures. The image examination exhibited here furnishing system for managing image semantics instead of pixel insights. Much of the time, data essential for the comprehension of and image isn't spoken to in single pixels however in significantimage objects and their shared relations. In a survey because of the approach ofPC innovation image preparing procedures have turned out to

be progressively vital in a wide assortment of applications. Image segmentation is an exemplary subject in the field of image handling and furthermore is a hotspot and focal point of image handling methods. A few universally useful calculations and systems have been produced for image segmentation. Since there is no broad answer for the image segmentation issue, these methods frequently must be joined with area information so as to adequately tackle a image segmentation issue for an issue area.The segmentation here is based on layer based and the block based in which further types like region and edge based are included.

Kuruvilla et. al. [9] with some other classification surveys propose the classification as thresholding based, region based, edge based, cluster based and hybrid. These were also compared on speed, accuracy etc. Even though different segmentation techniques are at hand, every method is not equally appropriate for a particular type of image. Thus the algorithm suitable for one class of image may not be suitable for another class of images. Consequently, there is no unanimously supported method for image segmentation for all categories of images and as a result it remains a challenge in image processing and computer vision. Despite the fact that distinctive segmentation procedures are close by, each technique isn't similarly proper for a specific sort of image. Accordingly the calculation appropriate for one class of image may not be reasonable for another class of images. Thusly, there is no consistently upheld strategy for image segmentation for all classifications of images and therefore it remains a test in image preparing and PC vision.

Said et. al. [10] proposed fundamental segmentation plans for compound image treatment for reducing its size while compression: object based, layer-based, also, local block based. Especially, they think about classification methods taking a shot at rough question limits, which decreases the confinement and accuracy of the segmentation, yet in return permits quicker, one-pass segmentation, low memory necessities, and a segmentation outline is better coordinated to existing strategies. They demonstrate numerical outcomes acquired on a printer application condition, where thorough standard of visual quality must be fulfilled.

Bleschke et. al. [11] previously built up a novel channel bank dependent on the idea of the Hessian lattice for grouping nodular structures and direct structures. The mammogram pictures were disintegrated into a few sub images for second distinction at scales from 1 to 4 by this channel bank. The sub images for the nodular segment (NC) and the sub images for the nodular and direct segment (NLC) were at that point acquired from investigation of the Hessian framework. Numerous locales of intrigue (ROIs) were chosen from the mammogram picture. The Bayes discriminate work was utilized for recognizing among irregular ROIs with a microcalcification bunch and two distinct sorts of typical ROIs without a microcalcification group. We assessed the recognition execution by utilizing 600 mammograms. Our mechanized plan was appeared to can possibly identify microcalcification bunches with a clinically adequate affectability and low false positives. Another productive technique to move forward symptomatic exactness in digitized mammograms is the utilization of PC Aided Diagnosis (CAD) framework.

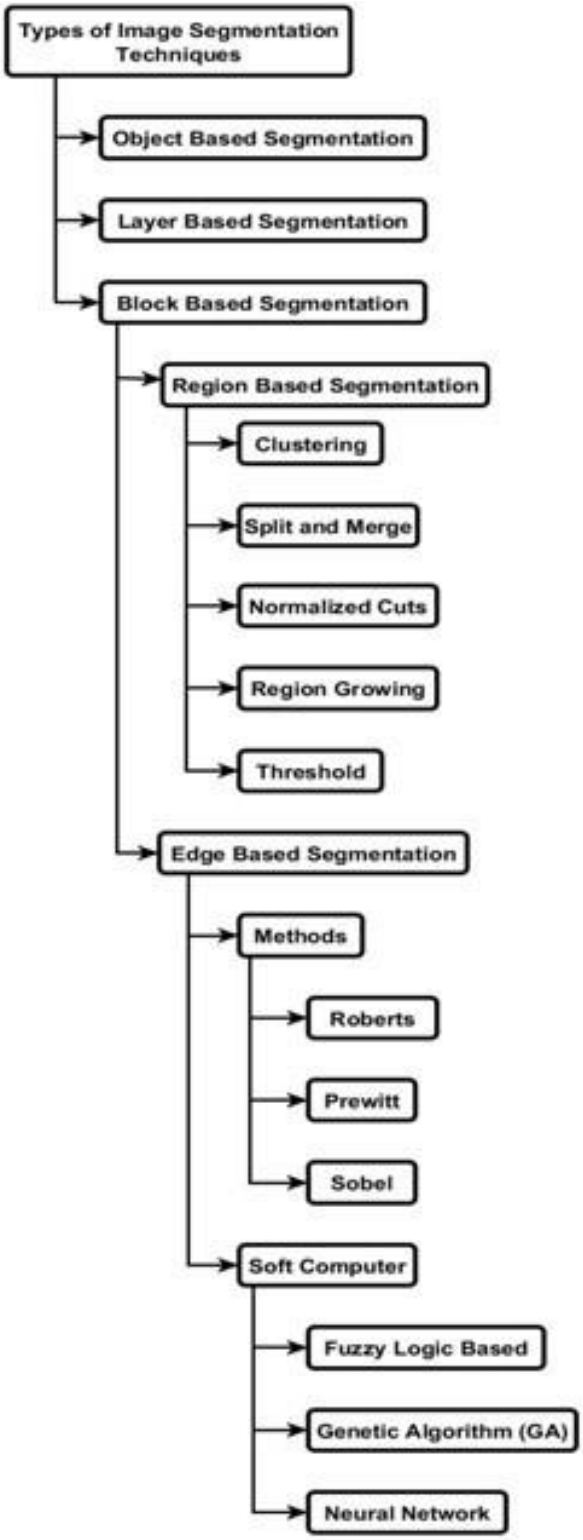


Figure 7: Classification of Segmentation methods

Zhili Chen et. al. [12] proposed a novel method for the classification of microcalcifications in mammograms. In this paper, the shape, morphological and cluster features are extracted for individual microcalcifications. Morphological and shape features like roughness, shape and size are extracted other than these features cluster features are also extracted that provides the global properties of the clusters. This work is distinct from previous approaches as here morphology of microcalcification clusters is focused such are cluster area, perimeter, eccentricity, circularity and many more. Graphs are generated at different scales to depict the topological structure of microcalcification clusters found. For classification a simple approach, K-nearest neighbors is used. This method is evaluated on two digitized datasets (MIAS and DDSM) and a full-field digital dataset. This study shows that all the features together including the topology modeling helped microcalcification analysis. Accuracy is 96% and for further improvement the author suggested to compute topological measures also.

Abdelali Elmoufidi et. al. [13] proposed a method to segment and detect the boundary of different breast tissue regions in mammograms by using dynamicK-means clustering algorithm and Seed Based Region Growing (SBRG) techniques. Firstly, the K-means clustering is applied for automatically generating the seeds points and determining the threshold values for each region. Secondly, the region growing algorithm is used with previously generated input parameters to divide mammogram into homogeneous regions according to the intensity of the pixel. Themain goal of this paper is to automatically segment and detect the boundary of different disjoint breast tissue regions in image mammography. Mammographic Image Analysis Society (MIAS) database is used for evaluation.

Akshay S. Bharadwaj et. al. [14] proposed Fuzzy C-means clustering for isolating the breast region. For eliminating the unwanted areas, the image is further segmented using Top Hat Transform. To isolate the ROI from the rest of image watershed transform is used. The Gibbs random fields are employed to analyze the pattern of pixels that matches with the devised clique patterns and detect MCs in the image.A thresholding is performed on the final image where the MCs are detected. Thedetection rate of this proposed method is 94.4%.

Anuj Kumar Singh et. al. [15] proposed a simple and easy approach for detection of cancerous tissues in mammograms. There are two phases proposed: Detection phase and Segmentation Phase. In the detection phase, ROI is obtained using an averaging filter and threshold operation is applied on original input image which outputs malignant region area. To find the malignant tissues, a rectangular window around the outputted region area is created and Max-Mean and Least-Variance technique is applied. In segmentation phase, a tumor patch is found using morphological closing operation and image gradient technique. This paper introduced a Max -Mean andLeast-Variance technique for tumor detection.

Wener Borges de Sampaio et. al. [16] proposed an adaptive algorithm capable of categorizing the image into dense and non-dense. The first stage consists of segmentation of the regions with the use of the micro-genetic algorithm. The nextstage is the reduction of false positives that were generated by previous segmentation technique. For this, two approaches were used: DBSCAN and a proximity ranking of the textures extracted from the ROIs. In the second reduction of false positives, the texture of resulting regions has been analyzed by the combination of PhylogeneticTrees, Local Binary Patterns and Support Vector Machines (SVM). The classification of masses and non-masses is performed using SVM. In this work, 1727 images fromthe DDSM database, being 1049 images of non-dense breasts and 678 images of dense breasts are used. This study is performed only on masses not on any other type of lesions.

Mellisa Pratiwi et. al. [17] proposed comparison of two classification methods: Radial Basis propagation Neural based on Gray-level Co-occurrence Matrix (GLCM) texture based features and BPNN. In this study, normal and abnormal breast image used as the standard input are taken from Mammographic Image Analysis Society (MIAS) digital mammogram database. The computational experiments show that RBFNN is better than Back-propagation Neural Network (BPNN) in performing breast cancer classification.

Chun-Chu Jen et. al. [18] proposed an efficient abnormality detection classifier (ADC) for automatic abnormality detection in mammogram images. Firstly, the input image is preprocessed for obtaining more accurate results. Preprocessing included global equalization transformation, image denoising, binarization, breast object extraction, determination of breast orientation and the pectoral muscle sup- pression. On the obtained segmented images, gray level quantization is performed. In the next step five features are extracted from the ROI and PCA is applied for determining feature weights.

J. Dheeba et. at. [19] Investigated a new classification approach for detection of abnormalities in mammograms images using Particle Swarm Optimized Wavelet Neural Network (PSOWNN). This algorithm is based on extracting Laws Texture Energy Measures (LTEM) from the mammograms and classifying the suspicious regions by PSOWNN classifier. This method is applied to real clinical database of 216 mammograms collected from different mammogram screening centers. The detection performance of the CAD system is analyzed using Receiver Operating Characteristic (ROC) curve. From the ROC curve, various parameters are derived such as sensitivity, specificity, AUC, Accuracy, Youden's Index and Misclassification Rate. The proposed classifier is compared with other classifiers such as SONN and DEOWNN and showed superior performance than these classifiers.

Xiaoming Liu et. al. [20] proposed an automatic mass detection method for breast cancer in mammographic images. Firstly, suspicious regions are found with an adaptive region growing method, named multiple concentric layers (MCL) approach. During MCL step, prior knowledge is used by tuning parameters using training dataset. Then, to improve the segmentation accuracy of masses, the initial regions are further refined with narrow band based active contour (NBAC). Texture features and geometry features are extracted from the regions of interest (ROI). The texture features are computed from gray level co-occurrence matrix (GLCM) and completed local binary pattern (CLBP). At last, the ROIs are classified by support vector machine (SVM), with supervision provided by the radiologist's diagnosis.The method was evaluated on a dataset with 429 craniocaudal(CC) view images, containing 504 masses. Among them, 219 images containing 260 masses are used to optimize the parameters during MCL step, and are used to train SVM. Theremaining 210 images (with 244 masses) are used to test the performance.

Danilo Cesar Pereira et.al. [21] Presents abnormality detection method in CC and

MLO view of mammograms. Preprocessing is performed using artifact removal algorithm followed by an image

denoising and enhancement based on wavelet trans- form and Wiener filter. Finally, for the segmentation of masses three techniques are used: multiple thresholding, wavelet transforms and genetic algorithm is employed. Database used for the work is Digital Database for Screening Mammography (DDSM). Quantitative evaluation of the work is carried out using area overlap metric (AOM). AOM achieved by the proposed method is 79%.

Shen-Chuan Tai et. al. [22] presents an automatic CADe system. In the initial step, the image is preprocessed to preserve the breast area and eliminate the structural noises in the mammograms. It is carried out using Otsu thresholding method followed by Gamma Correction. The pectoral muscle is removed using connected component labelling technique. For mending the borders of the pectoral muscles, morphological filters are used. To detect the suspicious masses on mammograms, a hierarchical matching method is used. In this study, there are three types of templates used out of which Sech Template was selected to match the suspicious area. In the next step, two feature extraction methods are used based on GLCM and ODCM. GLCM describes local texture characteristics while ODCM characterize photometric textures. Finally, Stepwise LDA is used to classify abnormal regions by selecting and rating the individual performance of each feature.

P. Shanmugavadivu et. al. [23] proposed an intuitive segmentation technique to separate the microcalcification regions from the mammogram. This mechanism first enhances the input image using an image-dependent threshold value and binarizes it to obtain an enhanced image. Then the pixels constituting the edges of microcalcification regions are grown in the enhanced image, with respect to the neighbourhood pixels. Lastly, the edge intensities of the enhanced image are remapped into the original image, using which the regions of interest are segmented. The results of the present work are compared with the ground realities of the sample images obtained from MIAS database.

Chen-Chung Liu et. al. [24] proposed an efficient algorithm for pectoral muscle extraction on MLO view of mammograms. The detection of pectoral muscle is very important for carrying out the next step i.e. Segmentation. So it is very important to have an efficient algorithm for preprocessing. For the pectoral muscle detection, various steps are performed. In the first step, Iterative Otsu Thresholding is applied. Then in the second step, mathematical morphological operations are applied to find a rough border of pectoral muscle. In the last step, multiple regression analysis (MRA) is employed at different degrees to obtain an accurate segmentation of pectoral muscle. This algorithm is tested on MIAS database.

Aya F. Khalaf et. al. [25] proposed a CAD system for detection of microcalcifications in mammograms based on novel feature set. In the preprocessing stage, the ROI's of the original image are cropped and are subjected to DWT using Db4 wavelet. The corresponding approximation, horizontal, vertical and diagonal coefficients are then calculated. Then in the feature extraction stage, new features are computed from several statistical observations such as higher order statistical (HOS) features, Discrete Wavelet Transform (DWT) and Wavelet Packet Decomposition (WPD). For feature selection, Student's t-test is used. Support vector machines (SVM) with two defined kernels, linear and RBF kernels are used.

Mario Mustra et. al. [26] proposed adaptive contrast enhancement method for breast skin–air interface detection which combines usage of adaptive histogram equalization method on small region of interest which contains actual edge and edge detection operators. Pectoral muscle detection method uses combination of contrast enhancement using adaptive histogram equalization and polynomial curvature estimation on selected region of interest. This method makes segmentation of very low contrast pectoral muscle areas possible because of estimation used to segment areas which have lower contrast difference than detection threshold.

Rangaraj M. Rangayyan et. al. [27] presents an overview of various digital image processing and pattern analysis techniques for problem solution in several areas of breast cancer disgnosis. This includes: contrast enhancement, detection and analysis of calcifications, masses and tumors, asymmetric shapes analysis and detection of architectural distortion.

Arnau Olivera et. al. [28] proposed an automatic detection approach for micro- calcifications and clusters in mammographic images. In this work, the morphology of the mc's is identified using the local features that are extracted from a bank of filters. In the initial training step, the system learns automatically and selects the most relevant features that are used by Boosted classifier for detection of individual microcalcifications. Further, microcalcification detection method is extended to cluster detection. The work is performed on MIAS database and a non-public database of 280 images.

Siti Salmah Yasiran et. al. [29] proposed comparison of Sobel, Prewitt and Laplacian of Gaussian (LoG) edge detection techniques in segmenting the boundary of micro- calcifications. After satisfying the breast phantom scoring criteria, segmentation phase is carried out. Then, all of the edge detection techniques were implemented in the Enhanced Distance Active Contour (EDAC) model for the segmentation process. Results obtained from Area under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve shows that the Prewitt edge detection has the highest value of AUC, followed by the Sobel and LOG which are 0.79, 0.72 and 0.71 respectively.

Indra Kanta Maitra et. al [30] proposed an efficient algorithm for contrast enhancement and suppression of pectoral muscle. For contrast enhancement, CLAHE technique is used. And for identifying the pectoral muscle, a rectangle is defined that isolates the pectoral muscle from the ROI and finally pectoral muscle is suppressed using

modified seeded region growing algorithm (SRG). The evaluation of algorithm is carried on 322 images of MIAS database.

Min Chen et. al. in [31] proposed a clustering approach based on combination of fuzzy C-mean clustering (FCM) with PSO. As fuzzy C-mean clustering has some drawbacks such as the number of clusters needs to be specified in advance and also we need to have knowledge of the ground truth. The data points in overlapping areas cannot be correctly categorized so to overcome all these drawbacks PSO is used in combination with FCM. The result of this algorithm shows that it can automatically find the optimal number of clusters.

Harry Strange et.al. [32] proposed a novel method for classification of microcalcification clusters as benign or malignant. This paper worked on discrete mereotopological relations between the individual microcalcifications over a range of scales. This range of scales is represented in the form of barcode. This barcode based representation helps representing complex relations that exist between multiple regions. This method gave classification accuracy of 95% for MIAS database and 80% for DDSM database.

Alan Joseph Bekker et. al. [33] proposed a two-step classification method for classifying the clusters of microcalcifications as benign or malignant. In this method, both the views of mammograms MLO and CC view are taken into consideration. Feature vectors of both the views are calculated and later on a linear regression classifier is used separately for both views to classify the ROI as benign or malignant. Then both views are combined to form a multi-view and a joint decision is taken. In this work, Expectation-Maximization (EM) algorithm is also combined with linear regression classifiers to find maximum likelihood parameters for a model. This algorithm is tested on a large database of 1410 images consisting of 750 pairs of CC and MLO views. The classification results are shown separately for dense and fatty breast tissues. Accuracy achieved for fatty breast tissues is 73.19% and for dense breast tissues it is 66.5%.

Gwenole Quellec et. al. [34] describes a CADx system for breast cancer diagnosis. In this work, the input image is first preprocessed and then the breast area is segmented. This output image is then partitioned adaptively into regions with respect to their distance from the breast edges. Then from each region features are derived along with the texture features for classification of normal or abnormal regions. Whenever an abnormality is found, the region from where it is induced is highlighted. To define the anomaly detector, two strategies are used. The first one is in which manual segmentation of lesions are used to train an SVM that assigns an anomaly index to each region and then these local anomalies are combined into global anomaly index. In the second way, the local and global anomaly detectors are trained simultaneously without manual segmentation using MIL algorithms such as Diverse Diversity (DD), Axis-Parallel Rectangles (APR), Multiple-Instance Support Vector Machines, MI-SVM, MILBoost. The work is carried on DDSM database.

Washington W. Azevedo et. al. [35] performed a method on IRMA database of mammograms that contains four types of tissues: fat, fibroid, dense and extremely dense. In this work, Morphological Extreme Learning machine is proposed with a hidden layer kernel based on morphological operators: dilation and erosion for classifying the masses as benign or malignant. At the feature extraction stage, Zernike moments and Haralick features are used. Comparison of sigmoid and morphological kernels used in ELM is evaluated through two parameters: Classification Rate and Kappa Index.

Fatemeh Pak et. al. [36] proposed a new algorithm for breast cancer detection and classification in digital mammography. The presented algorithm evolves three major steps: preprocessing, feature extraction and classification. For preprocessing, firstly the additional margins are removed which makes image smaller that reduces computational time. Then further, pectoral muscle is removed, for this all the images were aligned to left side first and then thresholding method and erosion morphological operator are used. Then for further improvement, the abnormalities are highlighted using Non subsampled Contourlet Transform (NSCT) followed by Super resolution (SR) algorithm. Before applying NSCT, the image dimensions are decreased by half to make computation fast. Then three levels NSCT decomposition is applied to divide the image in three levels of sub bands. In the first level, there were two sub bands. In the second level, there were four sub bands and in the third level there were eight sub bands. From all these sub bands, the sub bands with high frequency components are used only and then prewitt edge detector technique is used to obtain stronger edges of each selected sub band. After this the image is reconstructed then SR algorithm based on fuzzy learning algorithm is applied and finally a high pass filter is used for sharpening and highlighting the desired regions. At the feature extraction stage, several features including shape features and texture features based on GLCM are computed. At the end AdaBoost algorithm is used to perform classification to determine the probability of benign and malignant cases. The whole algorithm is applied on MIAS database and obtained accuracy is 91.43%. Jinchang Ren [37] proposed a new strategy for classification of imbalanced samples using SVM and ANN classifiers. Although the literature says SVM performs better than ANN but performance may become worse for imbalance training samples. When balanced learning and optimized decision making is applied then the performance of two classifiers becomes vary comparable. A balanced learning strategy is applied to individual classifier in this paper and then their performance is evaluated for successful classification of clustered microcalcifications.

R.V. Rao et. al. [38] proposed a new optimization method known as Teaching Learning- based optimization (TLBO). TLBO method is based on the influence of a teacher on the learners. TLBO is a population-based method. There are

many other nature- inspired optimization techniques like Artificial Bee Colony (ABC), Ant Colony Optimization (ACO), Genetic Algorithm (GA), and Particle Swarm Optimization (PSO). TLBO is also a population based method that usesa population of solution to reach to a global solution. For TLBO, population is the group of learners. As every optimization algorithm consists of different design variables, so in TLBO also we have different design variables that will be the different subjects taught to the learners and the result of the learners will correspond to the 'fitness' value. The best solution considered is the teacher. This process consists two phases: Teacher Phase and Learner Phase. The teacher phase means learning from the teacher and learner's phase means learning through the learners through interaction, discussion etc.

Arianna Mencattini et. al. [39] proposed a CADx system for classification of breast masses in mammograms. This system consists of five main steps: preprocessing, segmentation, feature extraction, feature selection and classification. In the initial step, preprocessing is carried out for removing the artifacts and contrast enhancement. Then segmentation is done using region growing technique by manually selecting a starting point and setting a stopping criteria based on selected threshold value. Two types of features are extracted from the extracted ROI, textural features based on GLCM and geometrical features. Then from these entire features, an optimal set of features are selected using ROC curves. The area under the ROC curve is taken as the criteria for selecting an optimal set of features. And finally classification is carried out using Na¨ıve Bayes classifier. This work is carried out on MIAS database.

Yiming Ma et. al.[40] proposed a shape analysis method for classification of micro- calcifications that are difficult to diagnose. For detection of each microcalcification, a region growing method is employed. After this close contour points of each MC within the ROI are obtained using Gradient Vector Flow (GVF) Active contour method. After the closed contour determination, distance signature is defined by calculating the Euclidean distance of each contour point to the centroid. Then frequency content of normalized distance signature is exploited to determine a metric to quantify the roughness of a contour. For band pass approximation of normalized distance signature, three level wavelet transform is used.

H. D. Cheng et. al. [41] for mass detection and classification and also compares their advantages and disadvantages. As for every CAD system, we need to follow few defined steps. Step one is image preprocessing, various methods are discussed in this paper such as global histogram modification approach, local processing approach, multiscale processing approach. Then for step two, image segmentation the techniques are classified in four groups: classical techniques, fuzzy techniques, multiscale technique and bilateral image subtraction. Then for the third step,various features are discussed such as intensity features, shape features and texturalfeatures. Step four is feature selection; under this two major methods are discussed:Stepwise feature selection and Genetic Algorithm. At the end, for the classification step few classifiers are discussed such as Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANN) and Binary Decision Tree.

Ozden et. al. [42] presented Multi resolution based calculation for microcalcification location in mammograms. The discovery of microcalcifications is accomplished by decaying the mammogram by wavelet change without testing administrator into various sub-groups, stifling the coarsest estimate sub band, lastly reproducing the mammogram from the sub bands containing just noteworthy detail data. Thenoteworthy points of interest are gotten by favorable ideas. Exploratory outcomes demonstrate that the proposed strategy is better in distinguishing the microcalcification groups than other wavelet decay strategies.

Haralick et. al. presented a Computer Aided Detection (CAD) strategy, or, in other words recognize knobs (microcalcification) in mammograms. Haralick et. al. [43] has planned a multi-scale channel bank dependent on the idea of second-arrange halfway subsidiaries (Hessian lattice). Districts of Interest (ROI) are distinguished by a multi resolution based histogram procedure. This ROI of mammogram is deteriorated into sub-groups, the low-recurrence sub band is stifled and afterward the high-recurrence sub bands which contain just knob like structures are remade. This structure is resolved by the eigen values of the Hessian lattice. The identification execution of the proposed strategy is assessed by contrastingour outcomes and two customary wavelet based strategies. Exploratory outcomes demonstrate that the microcalcifications can be productively identified by proposedtechnique and it has high obvious positive proportion in contrast with different strategies.

Ting et. al. [44], in an auto-testing breast malignancy mass division (ABC-MS)is proposed to help restorative specialists in breast disease finding. This techniqueis based on the single point region growing method. Manual division is actualizedas standard conclusion strategy for restorative specialists. This calculation candistinguish and portion the breast disease variation from the norm without earlier learning in regards to its quality. Mechanized single seed point locale developing is used in this calculation to play out the mass discovery and division naturally. The Mean Median Intersection point (MMIP) calculation is connected to process the edge an incentive to fragment the breast ROI. Single seed point locale developing is performed sequentially until the principal worldwide least of sleekness descriptoris come to. Emphasis will stop when the force contrast affected area mean and chose new pixel ends up bigger than first worldwide least of sleekness descriptor. The method was tested on MIAS dataset. The investigations are performed on computerized mammograms datasets given by Mammographic Image Analysis Society (MIAS). The dataset comprises of the first advanced mammograms at 50 micron goals in "Versatile Gray Map" (PGM) picture arrange and connected with separate analyzed ground truth information. Add up to 35 generous

cases, 32 threatening cases and 33 ordinary cases are included for the trial contemplates. The analyzed ground truth information is comprised of 7 segments information. The proposed technique is contrasted and business programming named as 3D-Doctor which is utilized as a semi-robotized division calculation. The presented work is designed to assist medical doctors in breast cancer diagnosis with.

Senthikumar et. al. [45], told that they have made enhancements in locale developing picture division for mammogram pictures to recognize the breast malignancy. Specific middle channel is utilized for preprocessing, CLAHE (Contrast Limited Adaptive Histogram Balance) strategy is utilized for the improvement, Harris corner recognize hypothesis is utilized to auto discover developing seeds and the seeded district developing tenet for the advancement of areas. This work likewise incorporates another vulnerability theory Cloud Model to acknowledge programmed and versatile division edge choosing, which thinks about the vulnerability of picture furthermore, separates ideas from qualities of the area to be divided as being human. They discovered this technique works solid on homogeneity and area qualities. Moreover, the technique has been tried for more than 40 test pictures and the outcomes found theyre great. Initially, the seed pixel must have high likeness to its neighbors. Second, for an expected locale, no less than one seed must be created in request to create this area. Third, seeds for various districts must be separated. In view of the programmed seed choice criteria, they utilize Harris corner recognize hypothesis to acknowledge programmed seed choice. The Harris corner indicator is a well known intrigue point identifier because of its solid invariance to turn, scale, brightening variety what's more, picture commotion. Harris corner indicator depends on the nearby auto-relationship capacity of a flag; where the neighborhood autocorrelation work estimates the nearby changes of the motion with patches moved by a little sum in various headings [9]. This calculation can discover the snack point and the tumor territory (Calcification) as appeared in the result. It additionally joins with particular middle sifting, CLAHE and performs well in breast disease discovery.

Swetha et. al. [46], told that the handled mammography pictures are helpful for conclusion of tumor. The antiquities and individual detail data's of pictures are killed with pre-handling phases of division. Further, the tumor's edge points of interest can be registered by utilizing different division strategies. These sectioned outcomes are further analyzed by the specialists. In those division strategies this paper works with Hybrid picture division and Otsu's thresholding strategies. In these strategies, the Hybrid picture division depends on quick clearing calculation and double front advancement with laplacian or slope and Otsu's thresholding with 10 limit levels are utilized. The mix of these two will give better execution. With this procedure, the antiques are to be wiped out and the malignant breast tumor is distinguished. These points of interest give the measure of the tumor and phase of the disease. The principle job of the proposed calculation in this paper is division.In this dynamic shape quick walking technique we can do division without much ofa stretch handle convexities, concavities, and topological changes in a picture. By utilizing this strategy we can get hearty and adaptable restorative picture division. Also, it contains two-organize approach, that is in light of a crossover edge and localebased strategy with quick clearing advancement and a double front development display arranged also. This work proposed antiquities disposal in preprocessing stageand division with Hybrid picture division and Otsu's thresholding. By consolidating these two sectioned strategies the exactness of tumor edges are identified. These techniques are precise and basic computational intricacy.

Thawkar et. al. [47] introduced an approach for division of masses in mammogram images. The proposed calculation utilizes Median filtering, Optimal Global thresh- olding utilizing Otsu's strategy and morphological operations with the end goal to improve nature of mammography picture, Segmentation of masses dependent on edge around region of interest separates sectioned masses from image. Division technique are enhanced extensively if the histogram crests are tall, limited, and symmetric and isolated by profound valleys. One approach to enhance state of histogram is to consider just those pixels that lie on or close to the edges amongprotest and foundation. Following algorithm is utilized to perform thresholding with edge location:

1. Obtain the edge data from input picture f(x,y) utilizing total estimation of Laplacian.
2. Specify a limit value i.e. a threshold to create a binary picture. This picture actsas a marker image.
3. Histogram is computed utilizing just the pixels in f(x, y) that relates the pixelsin twofold (binary) image and use this histogram to section input picture utilizing Otsu technique.

Mass division is accomplished through ideal worldwide thresholding technique i.e. utilizing Otsu's strategy. The utilization of middle filtering, picture editing and fringe expulsion as a preprocessing step is extremely valuable for improving the picture quality and evacuation of marks and non-mass locales appended to picture outskirt. The proposed strategies are executed and tried in Matlab on 50 mammography pictures (ordinary and tumor) to get ROI.

Nakayama et. al. [48] initially built up a novel channel bank dependent on the idea of the Hessian network for characterizing nodular structures and straight structures. The mammogram pictures were decayed into a few sub images for second contrast at scales from 1 to 4 by this channel bank. The sub images for the nodular segment (NC) and the sub images for the nodular and direct part (NLC) were at that point acquired from examination of the Hessian lattice. Numerous districts of intrigue (ROIs) were chosen from the mammogram picture. In every ROI, eight highlights were resolved from the sub images for NC at scales from 1 to 4 and the sub images for NLC at scales from 1 to 4. The Bayes discriminant work was utilized for recognizing among unusual ROIs with a microcalcification group and two unique sorts of ordinary ROIs without a microcalcification.

Balakumaran et. al. [49] presented Multi resolution based foveal calculation for microcalcification recognition in mammograms. The recognition of microcalcifications is accomplished by decaying the mammogram by wavelet change without testing administrator into various sub-groups, stifling the coarsest estimate sub band, lastly remaking the mammogram from the sub bands containing just noteworthy detail data. The critical points of interest are gotten by foveal ideas. Exploratory outcomes demonstrate that the proposed strategy is better in identifying the microcalcificationgroups than other wavelet decay techniques.

Balakumaran et. al. [50] have structured a multi-scale channel bank dependent on the idea of second-arrange halfway subordinates (Hessian network). Districts Of Interest (ROI) are recognized by a multiresolution based histogram procedure. This ROI of mammogram is disintegrated into sub-groups, the low-recurrence sub band is smothered and after that the high-recurrence sub bands which contain just knob like structures are remade. This structure is controlled by the eigen values of the Hessian lattice. The location execution of the proposed technique is assessed by contrasting our outcomes and two customary wavelet based strategies. Exploratory outcomes demonstrate that the microcalcifications can be proficiently distinguished by proposed technique and it has high evident positive proportion in contrast with different strategies.

III. CONCLUSIONS

Scanned images usually suffer from many artifacts than digital images so they require pre-processing. The purpose of pre-processing is to improve the quality ofthe image being processed. There are reasons for the need of image pre-processing:

- improvement of image quality to meet the requirements of physician
- noise reduction contrast enhancement
- correction of missing or wrong pixel values
- eliminationof acquisition-specific artifacts

Various preprocessing techniques have been discussed that includes different types of filters used for smoothing the input image. Along with that many techniques are discussed for pectoral muscle removal as it plays a vital role in incorrect prediction of breast cancer because the intensity levels of abnormalities and the pectoral are almost same in the mammogram. So if we don't remove the pectoral muscle part in the preprocessing step it may give fatal results in later stages. For Segmentation a number of techniques and methods have been studied. The techniques studied range from primitive to the ultra modern segmentation techniques. On the basis of survey, segmentation methods are disintegrated as thresholding based, region based, edge based, cluster based and hybrid. These were also compared on the basis of speed, accuracy etc. Even though different segmentation techniques are at hand, every method is not equally appropriate for a particular type of image. Thus the algorithm suitable for one class of image may not be suitable for another class of images. Consequently, there is no unanimously supported method for image segmentation for all categories of images and as a result it remains a challenge in image processing and computer vision. Despite the fact that distinctive segmentationprocedures are close by, each technique isn't similarly proper for a specific sort of image. Accordingly the calculation appropriate for one class of image may not be reasonable for another class of images. Thus, there is no consistently upheld strategy for image segmentation for all classifications of images and therefore it remains a testin image preparing and PC vision.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] Y. Yang, S. Hallman, D. Ramanan, and C. C. Fowlkes, "Layered object models for image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 9, pp. 1731–1743, 2012.

[2] R. Ramani, N. S. Vanitha, and S. Valarmathy, "The Pre-Processing Techniques for Breast Cancer Detection in Mammography Images," Int. J. Image, Graph. Signal Process., vol. 5, no. 5, pp. 47–54, 2013.

[3] S. K. Bandyopadhyay, "Pre-processing of Mammogram Images," Int. J. Eng. Sci. Technol., vol. 2, no. 11, pp. 6753–6758, 2010.

[4] F. E. A. I. N. S. A. S. N. F. A. M. and El-Samie, "An algorithm for pre- processing and segmentation of mammogram images," in Proceedings of 2016 11th International Conference on Computer Engineering and Systems, ICCES 2016, 2017, pp. 187–190.

[5] S. B. C. A.P. and Jamge, "The Preprocessing Methods of Mammogram Images for Breast Cancer Detection," Int. J. Recent Innov. Trends Comput. Commun., vol. 5, no. 1, pp. 261–264, 2017.

[6] Kshema, M. J. George, and D. A. S. Dhas, "Preprocessing filters for mammogram images: A review," in 2017 Conference on Emerging Devices and Smart Systems, ICEDSS 2017, 2017, pp. 1–7.

[7] J. C. Bezdek, L. O. Hall, and L. P. Clarke, "Review of MR image segmentation techniques using pattern recognition," Med. Phys., vol. 20, no. 4, pp. 1033–1048, 1993.

[8] J. Kuruvilla, D. Sukumaran, A. Sankar, and S. P. Joy, "A review on image processing and image segmentation," in Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016, 2016, pp. 198–203.

[9] A. Said and A. Drukarev, "Simplified segmentation for compound image compression," in IEEE International Conference on Image Processing, 1999, vol. 1, pp. 229–233.

[10] T. Blaschke, C. Burnett, and A. Pekkarinen, Image Segmentation Methods for Object-based Analysis and Classification. 2004.

[11] Z. Chen, H. Strange, A. Oliver, E. R. E. Denton, C. Boggis, and R. Zwiggelaar, "Topological Modeling and Classification of Mammographic Microcalcification Clusters," IEEE Trans. Biomed. Eng., vol. 62, no. 4, pp. 1203–1214, 2015.

[12] A. Elmoufidi, K. El Fahssi, S. Jai-Andaloussi, and A. Sekkaki, "Automatically density based breast segmentation for mammograms by using dynamic K-means algorithm and Seed Based Region Growing," in Conference Record - IEEE Instrumentation and Measurement Technology Conference, 2015, vol. 2015-July, pp. 533–538.

[13] A. S. Bharadwaj and M. Celenk, "Detection of microcalcification with top-hat transform and the Gibbs random fields," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2015, vol. 2015-Novem, pp. 6382–6385.

[14] A. K. Singh and B. Gupta, "A Novel Approach for Breast Cancer Detection and Segmentation in a Mammogram," in Procedia Computer Science, 2015, vol. 54, pp. 676–682.

[15] W. B. De Sampaio, A. C. Silva, A. C. De Paiva, and M. Gattass, "Detection of masses in mammograms with adaption to breast density using genetic algorithm, phylogenetic trees, LBP and SVM," Expert Syst. Appl., vol. 42, no. 22, pp. 8911–8928, 2015.

[16] M. Pratiwi, Alexander, J. Harefa, and S. Nanda, "Mammograms Classification Using Gray-level Co-occurrence Matrix and Radial Basis Function Neural Network," in Procedia Computer Science, 2015, vol. 59, pp. 83–91.

[17] C. C. Jen and S. S. Yu, "Automatic detection of abnormal mammograms in mammographic images," Expert Syst. Appl., vol. 42, no. 6, pp. 3048–3055, 2015.

[18] J. Dheeba, N. Albert Singh, and S. Tamil Selvi, "Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach," J. Biomed. Inform., vol. 49, pp. 45–52, 2014.

[19] X. Liu and Z. Zeng, "A new automatic mass detection method for breast cancer with false positive reduction," Neurocomputing, vol. 152, no. C, pp. 388–402, 2015.

[20] D. C. Pereira, R. P. Ramos, and M. Z. do Nascimento, "Segmentation and detection of breast cancer in mammograms combining wavelet analysis and genetic algorithm," Comput. Methods Programs Biomed., vol. 114, no. 1, pp. 88–101, 2014.

[21] S. C. Tai, Z. S. Chen, and W. T. Tsai, "An automatic mass detection sys- tem in mammograms based on complex texture features," IEEE J. Biomed. Heal. Informatics, vol. 18, no. 2, pp. 618–627, 2014.

[22] P. Shanmugavadivu and S. G. L. Narayanan, "Segmentation of microcalcifications in mammogram images using intensity-directed region growing," in 2013 International Conference on Computer Communication and Informatics, ICCCI 2013, 2013, pp. 1–6.

[23] C. C. Liu, C. Y. Tsai, J. Liu, C. Y. Yu, and S. S. Yu, "A pectoral muscle segmentation algorithm for digital mammograms using Otsu thresholding and multi- ple regression analysis," Comput. Math. with Appl., vol. 64, no. 5, pp. 1100–1107, 2012.

[24] A. F. Khalaf and I. A. Yassine, "Novel features for microcalcification detection in digital mammogram images based on wavelet and statistical analysis," in Proceedings - International Conference on Image Processing, ICIP, 2015, vol. 2015-Decem, pp. 1825–1829.

[25] M. Mustra and M. Grgic, "Robust automatic breast and pectoral muscle segmentation from scanned mammograms," Signal Processing, vol. 93, no. 10, pp. 2817–2827, 2013.

[26] R. M. Rangayyan, F. J. Ayres, and J. E. Leo Desautels, "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs," J. Franklin Inst., vol. 344, no. 3–4, pp. 312–348, 2007.

[27] A. Oliver et al., "Automatic microcalcification and cluster detection for digital and digitised mammograms," Knowledge-Based Syst., vol. 28, pp. 68–75, 2012.

[28] S. S. Yasiran et al., "Microcalcifications segmentation using three edge detection techniques," in International Conference on Electronic Devices, Systems, and Applications, 2012, pp. 207–211.

[29] I. K. Maitra, S. Nag, and S. K. Bandyopadhyay, "Technique for preprocessing of digital mammogram," Comput. Methods Programs Biomed., vol. 107, no. 2, pp. 175–188, 2012.

[30] M. Chen and S. A. Ludwig, "Fuzzy clustering using automatic particle swarm optimization," in IEEE International Conference on Fuzzy Systems, 2014, pp. 1545–1552.

[31] H. Strange, Z. Chen, E. R. E. Denton, and R. Zwiggelaar, "Modelling mammographic microcalcification clusters using persistent mereotopology," Pattern Recognit. Lett., vol. 47, pp. 157–163, 2014.

[32] A. J. Bekker, M. Shalhon, H. Greenspan, and J. Goldberger, "Multi-view probabilistic classification of breast microcalcifications," IEEE Trans. Med. Imaging, vol. 35, no. 2, pp. 645–6536, 2016.

[33] G. Quellec, M. Lamard, M. Cozic, G. Coatrieux, and G. Cazuguel, "Multiple- Instance Learning for Anomaly Detection in Digital Mammography," IEEE Trans. Med. Imaging, vol. 35, no. 7, pp. 1604–1614, 2016.

[34] and W. P. dos S. Washington W. Azevedo, Sidney M. L. Lima, Isabella
M. M. Femandes, Arthur D. D. Rocha, Filipe R. Cordeiro, Abel G. da Silva-Filho, "Morphological Extreme Learning Machines applied to detect and classify masses in mammograms," in International Joint Conference on Neural Networks (IJCNN), 2015, pp. 1–8.

[35] F. Pak, H. R. Kanan, and A. Alikhassi, "Breast cancer detection and classification in digital mammography based on Non-Subsampled Contourlet Transform (NSCT) and Super Resolution," Comput. Methods Programs Biomed., vol. 122, no. 2, pp. 89–107, 2015.

[36] J. Ren, "ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging," Knowledge-Based Syst., vol. 26, pp. 144–153, 2012.

[37] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-learning-based opti mization: A novel method for constrained mechanical design optimization problems," CAD Comput. Aided Des., vol. 43, no. 3, pp. 303–315, 2011.

[38] A. Mencattini, M. Salmeri, G. Rabottino, and S. Salicone, "Metrological characterization of a CADx system for the classification of breast masses in mammograms," IEEE Trans. Instrum. Meas., vol. 59, no. 11, pp. 2792–2799, 2010.

[39] Y. Ma, P. C. Tay, R. D. Adams, and J. Z. Zhang, "A novel shape feature to classify microcalcifications," in Proceedings - International Conference on Image Processing, ICIP, 2010, pp. 2265–2268.

[40] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du, "Approaches for automated detection and classification of masses in mammograms," Pattern Recognit., vol. 39, no. 4, pp. 646–668, 2006.

[41] M. O¨ zden and E. Polat, "Image segmentation using color and texture features," in 13th European Signal Processing Conference, EUSIPCO 2005, 2005, pp. 2226–2229.

[42] L. G. Haralick, R.M. and Shapiro, "Image segmentation techniques. In Applications of Artificial Intelligence," Int. Soc. Opt. Photonics, vol. 548, pp. 2–10, 1985.

[43] F. F. Ting, K. S. Sim, and S. S. Chong, "Auto-probing breast cancer mass segmentation for early detection," in Proceeding of 2017 International Conference on Robotics, Automation and Sciences, ICORAS 2017, 2018, vol. 2018-March, pp. 1–5.

[44] B. Senthilkumar and G. Umamaheswari, "A novel edge detection algorithm for the detection of breast cancer," in European Journal of Scientific Research, 2011, vol. 53, no. 1, pp. 51–55.

[45] T. L. V. N. Swetha and C. H. H. Bindu, "Detection of Breast cancer with Hybrid image segmentation and Otsu's thresholding," in 2015 International Conference on Computing and Network Communications, CoCoNet 2015, 2016, pp. 565–570.

[46] R. Thawkar, Shankar &Ingolikar, "Segmentation of Masses in Digital Mammograms using Optimal Global Thresholding with Otsu's method," Int. J. Comput. Sci. Technol., vol. 5, no. 3, pp. 129–132, 2014.

[47] R. Nakayama, Y. Uchiyama, K. Yamamoto, R. Watanabe, and K. Namba, "Computer-aided diagnosis scheme using a filter bank for detection of microcalcification clusters in mammograms," IEEE Trans. Biomed. Eng., vol. 53, no. 2, pp. 273–283, 2006.

[48] T. Balakumaran and I. Vennila, "Detection of microcalcification clusters in digital mammograms using Multiresolution based foveal algorithm," in Proceedings of the 2011 World Congress on Information and Communication Technologies, WICT 2011, 2011, pp. 657–660.

[49] T. Balakumaran, I. L. A. Vennila, and C. G. Shankar, "Microcalcification detection in digital mammograms using novel filter bank," Procedia Comput. Sci., vol. 2, pp. 272–282, 2010.

ENSEMBLE BASED VOTING CLASSIFIER FOR PREDICTION OF DDOS ATTACK

Taqdir, Amit Dogra
Department of CSE,GNDU, BGSBU
taqdir_8@rediffmail.com,
amitdogra004@gmail.com

ABSTRACT- Wireless sensor network provides resources as per requirement of the user. WSN consists of sensors arranged in sequence for sending and receiving signals. WSN is hampered with the attacks such as distributed denial of service, WORM hole attack etc. This work present ensemble-based approach for detecting DDOS attack caused by malicious users. The impact of DDOS attack on the WSN along with need to tackle DDOS attack is discussed. For accomplishing the detection process, ensembles based voting classifiers is designed. Ensemble based algorithms used for demonstrating the DDOS attack includes logistic regression, support vector machine, random forest, naïve bayes and KNN. The classification accuracy of combined classifier is better as demonstrated within the result section.

KEYWORDS- WSN, ensemble of algorithms, voting classifiers, classification accuracy

I. INTRODUCTION

Wireless sensor network (WSN) indicates the networks of spatially distributed sensors. These sensors are generally dedicated and meant to provide specific service only. Performance of WSN depends upon many distinct factors including environmental conditions (Humidity, pollution, sound etc) (Muhammad, Hussain and Yousaf, 2015). As the authenticated users becomes part of WSN so does unauthorized users. Thus, performance of the WSN also impacted by unauthorized access. Unauthorized users may cause multiple attacks within the network and thereby hampering the performance of the system (Singh, Singh and Kumar, 2017). The most common type of attack includes distributed denial of service attack. This attack is caused through distinct mechanisms. Some of these mechanisms includes

- **Flooding**

With flooding, thousands of packets are dispersed by the attackers over the network. This will cause other users to be in deadlock situations. This means they will not able access the resources and entire system will be in unstable state.

- **Protocol attacks**

These type of attacks eats the communication channel along with server resources. Thus, server resources will always be in deadlocked state.

- **Application layers**

Application layer attacks generally caused through cookies, capturing slots and bad bots. This type of attack could generate multiple identity attacks (AAMIR and ZAIDI, 2013). Source may not able to check the correct destination for transmission pf packets.

These type of attacks causes the distortion as well disturbance within the network. Extra energy loss could also be caused through this distributed denial of service attack. The next section presents the literature discussing the DDOS attack impacts on the networks along with the mitigation strategies. Section 3 gives the proposed methodology followed to accurately predict DDOS attack. The performance analysis and result are discussed in section 4, the conclusion and future scope is presented in the section 5, the last section gives the references.

II. LITERATURE SURVEY

The literature gives the tabular comparative analysis of techniques used for the detection of DDOS attack along with impact of attack on WSN.

TABLE I
Techniques for DDOS attack prediction along with impacts and issues

References	Techniques	Impact	Issues
(Lara and Ramamurthy, 2016)	Network policy-based mechanism for attack detection	Once attack occurs OpenSec system will fail, and resources will be consumed as a result	Energy consumption of sensor is not considered.
(Ganapathy *et al.*, 2013)	Discussed tools and techniques used for intelligent feature selection and classification of intrusion	Intrusion detection in case of multiple identity attacks could be detected.	The packet drops ratio and the energy consumption is not taken into account.
(Kumar and Santhi Tilagam, 2011)	Flooding attack detection	Only low-rate attacks could be detected but high rated attacks may cause traffic within the network	In this case, it is found that the Packet drop ratio is high.
(Bukac and Matyas, 2015)	Traffic pattern analysis in case of DOS standalone attack	Distributed attacks could hamper performance of the network and denial of service requests	Lifetime of the network will be reduced but not considered in this literature

(Behal, Kumar and Sachdeva, 2018)	D-Face based approach for early detection of DDOS attack	Internet domain will be impacted with this type of attack	Internet domain-based attack could reduce the packet to base station and should be a part of DDOS attack metric consideration
(Meenakshi, Kumar and Behal, 2021)	Deep learning based approach for DDOS attack detection	LSTM based mechanism applied detect the impact of resource wastage within WSN	In the detection process, the classification accuracy is low and that is an issue.
(Nguyen *et al.*, 2021)	Detection of DDOS with the Deep learning and gaussian model	This model evaluates the impact on network resources and also determine the percentage resource consumption	The classification accuracy of the detection method is low.

III. PROPOSED METHODOLOGY

The methodology followed is based upon different classifiers including KNN, random forest, SVM, naïve bayes, logistic regression and ensemble based coting classifier. The ensemble-based approach produced better classification accuracy as compared to individual approaches. Fine tuning of classification accuracy also resulted in better learning rate and low error rates. Dataset for demonstration is fetched from Kaggle. The pseudo codes for the approaches are presented in this section

TABLE II
Pseudo code corresponding to different classifiers

KNN

knn = set_nearest_neighbourcount(7)
Evaluate distance between neighbours
Check prediction with KNN classifers

Outcome-Prediction

```
array([1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1,
       1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0,
       1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1,
       1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0,
       0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0,
       1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1,
       0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0],
      dtype=int64)
```

Logistic Regression

Build LR model
LR.train(X_train,Y_train)
Perform Prediction using Linear regression

```
array([1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,
       0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0,
       0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0,
       1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,
       0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0,
       1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1,
       0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0],
      dtype=int64)
```

SVM

Build support vector machine with predefined kernel function
Initialize hyperplanes within SVM classifier
Perform classification using test data

```
array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
      dtype=int64)
```

Random Forest
Initialize random forest classifier Initialize number of tree blocks Train rf(test,train) Perform prediction
array([1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0], dtype=int64)
Initialize naïve bayes algorithm Initialize kernel(Linearity=5,Learning Rate=0.1) Train NB(train,test) Prediction using Naïve bayes
array([0, 0], dtype=int64)

Almost all the classifiers given the mix result where '1' indicates that attack has been detected and '0' means no attack is detected. The voting classifiers with fine tuning gives the result by accommodating good features of all the classifiers and present the result as per majority. This means that if 3 out of five classifiers give '1' as prediction and 2 classifiers yields '0' for the same data, then voting classifier will give '1' as prediction. The pseudocode for the same is given as under

TABLE III
Voting classifier demonstration

Voting Classifier
Ensemmble_Algorithm(logistic_regression,random_forest,naïve_bayes,support_vector,knn) Fit the data ensemble.fit(train,test) Perform prediction using ensemble
VotingClassifier(estimators=[('lr', LogisticRegression()), ('rf', RandomForestClassifier()), ('nb', GaussianNB()), ('svc', SVC(kernel='linear', probability=True)), ('knn', KNeighborsClassifier(n_neighbors=7))], voting='soft', weights=[1, 1, 2, 2, 1])
Ensemble_prediction(test) Calculate parameter confusion_matrix(test,train) Perform prediction_comparison(validation)
array([1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 1, 0], dtype=int64)

I. PERFORMANCE ANALYSIS AND RESULT

Since the majority of the classifiers gives a similar outcome only the SVM classifier gives the value '0' for all the predictions hence voting classifier generated the above-listed outcome. In the form of metrics, the result is shown below.

TABLE IV
Performance Analysis and result in terms of different metrics

Print roc curve from confusion_matrix
From specificity, sensitivity , f-score and other metrics from the confusion matrix
Result1=LR()
Result2=RF()
Result3=KNN()
Result4=SVM()
Result5=NB()
Result5=Voting_Classifer()

	Model	Accuracy	Precision	Recall	F1 Score	ROC
0	Logistic Regression	0.720779	0.590164	0.666667	0.626087	0.708333
1	Random Forest	0.694805	0.550725	0.703704	0.617886	0.696852
2	KNN	0.766234	0.655172	0.703704	0.678571	0.751852
3	SVC Linear	0.649351	0.000000	0.000000	0.000000	0.500000
4	NB	0.649351	0.000000	0.000000	0.000000	0.500000
5	Voting Classifier	0.727273	0.590909	0.722222	0.650000	0.726111

V.CONCLUSION AND FUTURE SCOPE

This work presented the detection of DDOS attack using distinct classifiers including logistic regression, random forest, KNN, SVM, naïve bayes and voting based classifiers. The result section demonstrates that result of logistic regression classifier in the detection process is highest. However still, classification accuracy is below desired levels. To accomplish better result, voting based classifier with fine tuning mechanism is applied. Voting classifier yield better accuracy but improvement is limited. In future, outlier detection mechanism along with missing values handling could be used for better results.

REFERENCES

[1] AAMIR, M. and ZAIDI, M. A. (2013) 'A Survey on DDoS Attack and Defense Strategies: From Traditional Schemes to Current Techniques', *Interdisciplinary Information Sciences*, 19(2), pp. 173–200. doi: 10.4036/iis.2013.173.

[2] Behal, S., Kumar, K. and Sachdeva, M. (2018) 'D-FACE: An anomaly based distributed approach for early detection of DDoS attacks and flash events', *Journal of Network and Computer Applications*, 111, pp. 49–63. doi: 10.1016/j.jnca.2018.03.024.

[3]Bukac, V. and Matyas, V. (2015) 'Analyzing traffic features of common standalone DoS attack tools', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9354, pp. 21–40. doi: 10.1007/978-3-319-24126- 5_2.

[4] Ganapathy, S. *et al.* (2013) 'Intelligent feature selection and classification techniques for intrusion detection in networks: a survey', *EURASIP Journal on Wireless Communications and Networking*, 2013(1), p. 271. doi: 10.1186/1687-1499-2013-271.

[5] Kumar, A. and Santhi Tilagam, P. (2011) 'A Novel Approach for Evaluating and Detecting Low Rate SIP Flooding Attack', *International Journal of Computer Applications*, 26(1), pp. 31–36. doi: 10.5120/3067-4192.

[6] Lara, A. and Ramamurthy, B. (2016) 'OpenSec: Policy-Based Security Using Software-Defined Networking', *IEEE Transactions on Network and Service Management*, 13(1), pp. 30–42. doi: 10.1109/TNSM.2016.2517407.

[7] Meenakshi, Kumar, K. and Behal, S. (2021) 'Distributed denial of service attack detection using deep learning approaches', *Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development, INDIACom 2021*, pp. 491–495. doi: 10.1109/INDIACOM51348.2021.00087.

[8] Muhammad, S., Hussain, S. and Yousaf, M. (2015) 'Neighbor Node Trust Based Intrusion Detection System for WSN', *Procedia - Procedia Computer Science*, 63, pp. 183–188. doi: 10.1016/j.procs.2015.08.331.

[9] Nguyen, T. T. *et al.* (2021) 'Detection of unknown DDoS attacks with deep learning and Gaussian mixture model', *Proceedings - 2021 4th International Conference on Information and Computer Technologies, ICICT 2021*, pp. 27–32. doi: 10.1109/ICICT52872.2021.00012.

[10] Singh, K., Singh, P. and Kumar, K. (2017) 'Application layer HTTP-GET flood DDoS attacks: Research landscape and challenges', *Computers & Security*, 65, pp. 344–372. doi: 10.1016/j.cose.2016.10.005.

`REVIEW ON APPLICATION AREAS OF IMAGE PROCESSING

*Ravi Kumar Verma[*1], Dr Lakhwinder Kaur[#2],ER Navneet Kaur[#3]*
[#]Department of Computer Science and Engineering Punjabi University, Patiala, Punjab,India
[1]*ravisadhak@gmail.com*
[2]*mahal2k8@yahoomail.com*
[3]*navneetmavi88@gmail.com*

ABSTRACT:— Researchers are always interested in image processing due to its importance in enhancement of visual information for benefit of society, it can also process images for machine vision as according to HVS (Human Visual System). Images which are digital in nature can be processed by image processing and can provide benefits to users of image data as well as scientists, engineers in their research work. If we talk about methodologies in image processing then we can do various operations like converting images to greyscale, segmentation of images, detecting edges, extracting some features from images and most important classification of images. In this paper we are going to discuss different methods of image processing used by researchers. In order to make the images noise free and clear, edge detection tools are used. If we talk about pattern recognition in images and image division then the first step is image segmentation. Feature extraction is also discussed which mainly involve shape and colour features. We also discuss how machine learning algorithms are used for classification of images. This paper is beneficial for researchers who want to acquire knowledge regarding identification and processing of images. There are number of applications of image processing like computer vision, sensing object remotely, extracting features from images, optical character recognition, detection of finger prints etc.

KEYWORDS Image Processing, Segmentation, Feature Extraction, Machine Learning, Gray level co-occurrence matrix (GLCM)

I. INTRODUCTION

Using image processing we can convert image from analog to binary representation and then perform operations on individual bits so that image quality will be increased and useful information can be extracted from the image. First of all, image is fetched by processing software, it can be acquired online or read operation is performed on hard disk where the image is stored and then image is processed according to instructions written inside the sub process of the processing software and output will be produced in the form of new image, information from the image or some features which are extracted from the image [1]. Segmentation of image is initial step of image processing. If we talk about good image segmentation algorithm it must be fast, shaping of objects must be good, so that picture can be highlighted easily. Incomplete shape lead to extra work regarding over- sections output [2]. When image processing is applied in computer vision, image is divided into different sectors for analysis so that important features of the image can be identified and time for analysis becomes less. Division is performed so that objects can be easily identified in the picture. Using division each pixel of the image is assigned a mark, so that pixels having same properties can be identified [3]. Without division we cannot perform image analyzation and interpretation [4]. Using division, we can resolve issues in machine level initialization and controlling irregular pictures. Division is required in almost every application of image processing like finding and identifying objects in images. There are many steps of image processing (Figure 1) some are discussed below.

II.FUNDAMENTAL STEPS IN IMAGE PROCESSING

A. Image Pre-processing

Before images are actually used for processing there are some steps which must be taken to format images. For e.g.: image resizing, correcting color and image orientation but it is not limited to only these steps as steps can be more advanced depending upon the type of project. Ganzalez et. al. [5] discussed a contrast enhancement method which is based on histogram in which brightness is flattened which is across the picture, in this technique darker areas will be enhanced and brighter areas will not be over exposed. Grisan et al. [6] have discussed a mathematical model related to illumination of background and conclude that normalization of contrast is effecting lesion segmentation algorithms negatively. By smoothing of green image using median filter background image was estimated. Lee et al.[7] deployed 56x56 median filter to acquire the image which is shade corrected.

B. Image Segmentation

Process of image segmentation is applied so that image can be partitioned into various segments. Image segmentation alters image representation, converting it to more systematic and meaningful form which is easier to analyze. Using segmentation object can be located in the image and boundaries can be created. Saleh et. al.[8] discussed about edge based segmentation in which there is application of edge filter to the image and according to the filter output pixels are classified into two categories edge and non-edge filter. There is another method named Region Based segmentation in which similarities are identified in pixels which are adjacent Unique regions are allocated for similar pixels [9].One of the simple methods of image segmentation is Thresholding, binary images are created from grey scale image Thresholding[10].Using segmentation binary images are developed from color images[11].

C. Feature Extraction

Relevant shape information of image is retrieved by feature extraction, using it classification problem of images can be easily solved using a procedure which is formal. In image processing feature extraction can be used for dimensionality reduction. There is another technique which is based on color vector [12,13] in this standard deviation and mean of the image pixel intensity are extracted to extract feature. Shape Features extraction technique [14] is a continuous approach in which the shape is not divided into subpart. In this approach boundary is used which is integral to the feature vector. Heurtier et. al.[15] discussed about texture feature approach in which it is observed that there is requirement of square region with sufficient size to extract texture.

D. Feature Selection

Using feature select a large set of features is reduced from the image and only effective features are selected and redundant features will be discarded from the image. Subset of features can be selected using ant features selection algorithm, Measure of importance related to local feature and overall performance regarding subsets [16] so that feature space can be searched for ideal solutions. Randomized algorithm are used detecting objects quickly in image which are noisy [17]. Complex optimization problems can be solved by a statistical mechanics method known as simulated healing, also it can be used to solve problems which occur in image processing [18]

E. Image Classification

In order to extract information, pixels and labels from image classification is performed. Multiple images of same object are needed to perform classification. In order to perform classification effectively suitable number of training samples as well as right classification technique is required. Kumar et. al.[19] discussed Artificial Neural Networks in image classification in this technique in which value of the number of information classes is same as output layer nodes number and dimension of each pixel is equal to the node present in the input layer. Moustakidis et. al.[20] discussed Fuzzy Measure in which accuracy and performance of classification is totally dependent on fuzzy integral as well as threshold selection. Taherkhani et. al[21] discussed about Naïve Bayes classifier in image classification ,concept of probability representation is used after that a class is assigned and the probability of assigned class to the feature vector which is acquired from ROI is greatest.

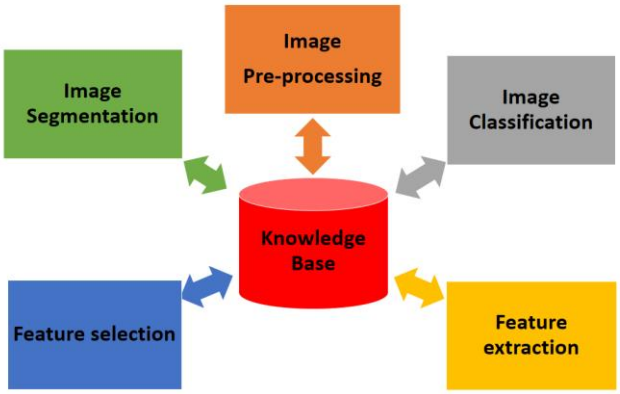


Figure 1 Image processing Fundamental steps.

III. APPLICATION AREAS

There are so many fields of image processing like recognition of images, segmentation of images and classification of images. It is the base for other applications like recognizing patterns, identifying objects. Mainly image processing deals with binary image processing although optical processing is also possible with additional efforts. In this paper we will discuss about commonly used mechanisms of image processing. In Image processing discrete structure of images are converted to signal format and then processed and after processing signals are converted back into digital image. Here we are going to discuss some of the research papers regarding image processing mechanisms.

A. Image Processing in Crack Detection.

In Image restoration a clean and original image is produced from an image which is corrupt or noisy. Motion blur, camera miss-focus and noise are some forms of image corruption. In order to recover the object advanced image processing techniques are applied. An algorithm is proposed which is used to detect cracks using image processing [22]. Firstly, the input image is smoothened and then threshold method is applied for segmentation. In order to analyse the image the computations which needs to be performed are related to area and perimeter of roundness. After analysis, the presence of crack is determined. Adhikari et.al.[23] developed a mechanism in which cracks are identified numerically. In this algorithm crack quantification is implemented using neural network and 3d visualisations. An Algorithm is proposed which is based on image stitching in which registration is feature based [24]. Mechanism of skeletonization procedure is

used so that crack segments can be retrieved. Cracks are detected on the basis of size is dependent on quantification of cracks. Crack length can be detected and changed with the help of neural network using which depth of the crack can be predicted and patterns of crack can be visualized using 3d graphs.

A mechanism is developed in which measurements of cracks is automated using computer vision [25]. Proper dimension estimation regarding cracks is evaluated by processing images, using a single camera. Mainly used algorithm are HSB and RSV with which images are processed to detect cracks. Images are applied as input to the algorithm and images with red particles will be produced which shows the detected cracks. Pixel data will be stored in a data structure like vector and transferred to the algorithm which measure cracks. Cross section is analysed by the algorithm with pixel positions so that number of pixels in cross section can be obtained and crack dimension will be produced.

A new approach is developed for crack detection, in which dark colour methodology and low contrast are used by applying curvelet waveform and analysis of text [26]. In this, image decomposition and reconstruction is performed using FDCT algorithm. Texture features will be measured to find the thresholds and surface texture will be removed. Contours will be extracted from the final images, in the final output image there will be no texture, but crack defects will be detected.

B. Image Processing in Production Industry

Inspection of products using Image processing is very important so that product quality and productivity can be enhanced in production industries [27]. If we talk about bulb production industry then filaments can be inspected, the process of manufacturing of bulbs can be monitored with the help of visual inspection. The reason behind filament get fused is error in shape and size of filament. For e.g. Pitch of wiring is not uniform. Manually it is not possible to detect the problem accurately. In Inspection system which are vision based, filament image is produced and it is binary in nature. Silhouette of the filament is produced from that image. Then from that silhouette information we can get the error in the pitch of filament shape and size. General Electric corporation developed an automated vision inspection system regarding bulb filament error detection. Faulty components in electronic systems like faulty resistance, capacitor can also be detected using automated visual inspection

systems [28]. Thermal energy generation from faulty components is more as compared to non-faulty components. From the distribution of thermal energies, infrared images are produced. Using these Infra-red rays one can detect faulty component easily. Identification of flaws in industries is very important in metal related industries [29] specially related to surface. If we talk about rolling mills[30], errors on the metal surface are very important to detect. Detection of edges, identification of texture and fractal analysis are the image processing techniques which can be used to perform automatic surface inspection.

C. Tracking Moving Objects

Tracking of moving objects[31] is also very important area of image processing, using it we can measure parameters of motion of object and a record is maintained regarding visual elements of object. There are two approaches in tracking moving objects one is based on recognition and second one is based on motion. There are motion based[32] predictive techniques which are used to detect missile, aircraft etc. For e.g. Kalman Filtering[33], particle filtering etc. In object tracking systems where automated image processing is applied, when the target object enters the boundary of the sensor, it is detected automatically so no human intervention is required. Each and every image frame is recognised in recognition based tracking and positional information of the object is gathered and then object is recognised.

D. Vehicle Detection

Licence plates in vehicle image can be detected[34] as they are rectangular in shape and their edge information is easier to detect using light and heuristic energy with the help of histogram approach Line and Clip functions are used to detect plates in morning as well as night with which Gaussian function is analysed Otsu's algorithm is used to recognise vehicle[35] number plates which uses templates, in this threshold partitioning[36] is used.

E. Image Retrieval from Large Databases

Extraction of images for a large dataset is also very important application area of image processing. In today's world databases are very large in size in which multimedia content is stored, especially due to rise in social media, so in order to extract accurate information from such large dataset; image processing can be used to make the process faster. If we talk about today's search engines then extracting text information is fast and easy, but when it comes to color images process is slow if image processing is not deployed. Approaches which are traditional are not fast and requires high cost. There is need to develop a proper image processing based search system, which can integrate image processing with search techniques to achieve higher accuracy and speed. Shape, texture, color which are feature of an image can be used to search information form large databases. Content-based retrieval of images is discussed in [37,38]. Various other approaches had also been reported for this technology such as color correlogram [39],texture similarity based technique[40],technique based on distance metrics performance analysis[41].

F. Medical Image Processing

In human body many structures are hidden which are not possible to identify with naked eyes. So, medical image processing plays a vital role in processing the images to extract those hidden structures. Medical image processing is also used for diagnosis, recognition and treatment of diseases. If a doctor/radiologist want to view the internal parts of the body

without opening the human body, for e.g. in laparoscopic surgeries then he/she needs a imaging modality for that. Various imaging modalities are X-ray, MRI(Magnetic Resonance Imaging), Ultrasound etc. Using this doctor/ radiologist can look inside the body by viewing its hidden dimensions. We are going to discuss here about Radiography and MRI although there are many more medical imaging modalities available.

G. Radiography

Electromagnetic radiation is used in radiography to view the object which is not transparent and its composition is changing, for e.g. body of human beings. There is a machine called X-Ray Generator [42] which is used to generate a beam of X-rays which is heterogeneous in nature and these X-rays are applied on the object whose internal detail we want to extract. When the X-Rays penetrate through the target object, there is detector which captures the X-rays [43], and 2d representation of the internal structure of the target organ is provided. In Fig 1 there is an X-Ray image which provides information about human teeth. Radiography can be used in dental treatments [44] related to orthopaedic surgeries. There is two types of radiography: medical radiography and industrial radiography. If the object examined by radiography is living, then it is referred to as medical where these images are further used for some disease detection otherwise it is termed as industrial.

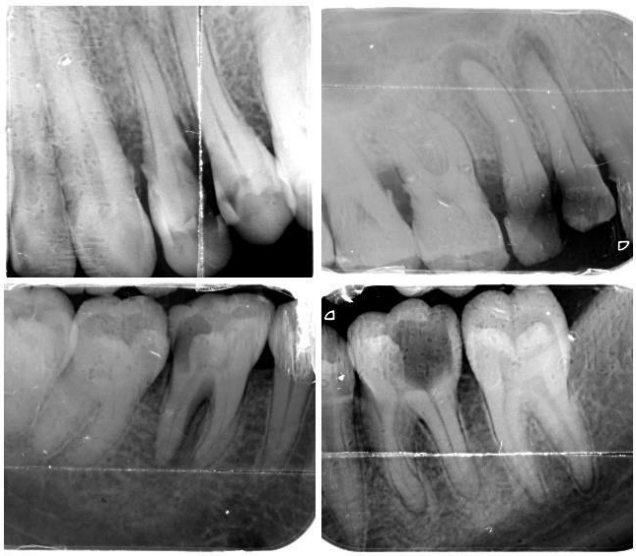


Fig 1 Human teeth X-ray[45]

H. Magnetic Resonance Imaging (MRI):

In MRI [46] physiological processes and anatomy of the human body is converted into image. Magnetic fields and radio waves are used to extract images. In Fig 2 MRI scan of human brain is shown. In MRI scanner[47] effects of Radio Waves and Magnetic field is used, so that images of human body organ can be obtained. It can be used in medical diagnosis as well as research related to bio-medical images. Apart from digital images other clinical data is also produced by MRI. Data produced by MRI is voluminous in nature and there is no need of manual segmentation. Algorithms like image denoising[48],skull stripping, image reconstruction[49]are applied to so that use of brain images can be simplified and the information which is obtained can be enhanced.

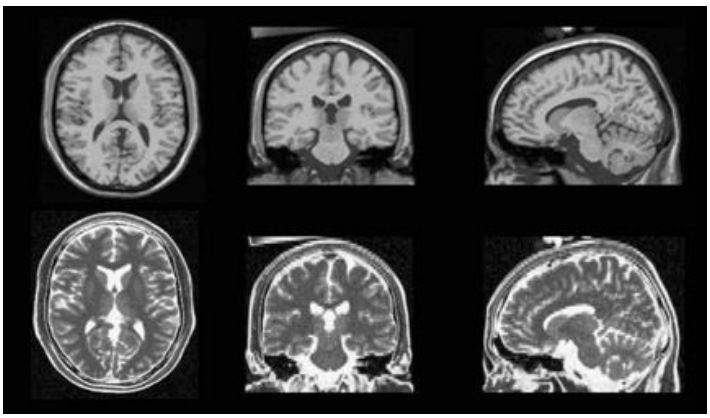


Fig 2Human Brian MRI [50]

ANN and SVM can also be used in medical image processing. Thai et.al.[51] have discussed regarding classifying images using Artificial Neural Networks(ANN) and Support Vector machines(SVM). First the receptive class is created using Artificial neural networks and using support vector machine all the categories will be interpreted. After

processing images, segmentation of images and extraction of features is done in the form of vector. In the output there is large portrayal pace as well as sub regions. Then the picture removal process occurred from sub regions as a vector. Then the extracted vector will be provided as an input to ANN [52]. If we talk about processing of ANN then there are three types input, hidden and output. Feature vector elements are always equals to input layer nodes. Class of Artificial neural network is equal to number of nodes in output layer. Support vector machine is used to find the appropriate weight in this system. First of all, support vector machine is trained, SVM is synchronized [53] with the particular image problem by manipulating its parameters. All the classifications of ANN are joined by SVM. Whole process of classification is proposed using ANN and SVM which is less time consuming. Main problem with this approach is training time for SVM with large datasets. Hussain et. el[54] proposed a method in which skull stripping and median filtering is used for pre-processing of image as well as GLCM technique is used for extracting features and Classifier which is used is SVM. Kadam et el.[55] proposed an algorithm in which malign brain cells are darkened to detect brain tumor as well as it helps the doctor by enhancing MRI images so that other diseases can also be detected inside human brain. Useful information is extracted by GLCM features ANN classifier is used to show reasonable results.

IV.CONCLUSION

Applications of image processing are very large in number, so it depends on the person who wants to research that which stream of image processing to choose. So many techniques are developed after research, but there is always more to happen. As we have high speed CPU's, graphic processers and signal processers, digital image processing can be easily implemented and is used in almost every trade because of it versatile nature and also it is less expensive. Researchers use image processing techniques in their research work on different types of images to achieve accuracy in their research area. Image quality is mainly based on clarity of the image and the technique which is applied so researchers are always in the search of best techniques.

REFERENCES

[1] K.Sumithra,S.Buvana,R.Somasundaram."ASurvey on Various Types of Image Processing Technique" *International Journal of Engineering Research & Technology(IJERT),* ISSN: 2278-0181,Vol. 4, March-2015

[2] Kumar, V., Lal, T., Dhuliya, P., & Pant, D. (2016). A study and comparison of different image segmentation algorithms. *2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall)*. https://doi.org/10.1109/icaccaf.2016.7749007

[3] Petrellis, N. (2017). A smart phone image processing application for plant disease diagnosis. *2017 6th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. https://doi.org/10.1109/mocast.2017.7937683

[4] Radha, R., & Jeyalakshmi, S. (2014). An effective algorithm for edges and veins detection in leaf images. *2014 World Congress on Computing and Communication Technologies*. https://doi.org/10.1109/wccct.2014.1

[5] 1)R. C. Ganzalez and R. E. Woods, Digital image processing, Second edition. Prentice Hall: New Jersey, 2001.

[6] Grisan, E., Giani, A., Ceseracciu, E., & Ruggeri, A. (n.d.). Model-based illumination correction in retinal images. *3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano, 2006*. https://doi.org/10.1109/isbi.2006.1625085

[7] Lee, S. C., Wang, Y., & Lee, E. T. (1999). . *Medical Imaging 1999: Image Perception and Performance*. https://doi.org/10.1117/12.349664

[8] S. Saleh, N. V. Kalyankar and S. Khamitkar, "Image segmentation by using edge detection", (IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010.

[9] Angelina., S., Suresh, L. P., & Veni, S. K. (2012). Image segmentation based on genetic algorithm for region growth and region merging. *2012 International Conference on Computing, Electronics and Electrical Technologies (ICCEET)*. https://doi.org/10.1109/icceet.2012.6203833

[10] Zhang, Y. (n.d.). An overview of image and video segmentation in the last 40 years. *Advances in Image and Video Segmentation*. https://doi.org/10.4018/9781591407539.ch001

[11] Lindeberg, T., & Li, M. (1997). Segmentation and classification of edges using minimum description length approximation and complementary Junction cues. *Computer Vision and Image Understanding*, *67*(1), 88-98. https://doi.org/10.1006/cviu.1996.0510

[12] Pachouri, K. K. (2015). A Comparative Analysis & Survey of various Feature Extraction Techniques, 6(1), 377–379.

[13] Liu Jinxia, & Qiu Yuehong. (2011). Application of SIFT feature extraction algorithm on the image registration. *IEEE 2011 10th International Conference on Electronic Measurement & Instruments*. https://doi.org/10.1109/icemi.2011.6037882

[14] Zhang, D., & Lu, G. (2004). Review of shape representation and description techniques. *Pattern Recognition*, *37*(1), 1-19. https://doi.org/10.1016/j.patcog.2003.07.008

[15] A. Humeau-Heurtier, "Texture Feature Extraction Methods: A Survey," in IEEE Access, vol. 7, pp. 8975-9000, 2019, doi: 10.1109/ACCESS.2018.2890743.

[16] Dewi Nasien, Habibollah Haron and Siti S. Yuhaniz. (2010). Metaheuristics Methods (GA & ACO) For Minimizing the Length of Freeman Chain Code from Handwritten Isolated Characters, World Academy of Science Engineering and Technology, Vol. 62, February 2010, ISSN: 2070-3274, Article 41, pp. 230-235Xyz

[17] Dewi Nasien, Habibollah Haron, Siti Sophiayati Yuhaniz. (2011). The Heuristic Extraction Algorithm for Freeman Chain Code of Handwritten Character. International Journal of Experimental Algorithms (IJEA). Publisher: CSC Press, Computer Science Journals, Volume: 1, Issue: 1, pp. 1-20, ISSN: 2180-1282.

[18] Azmi, R., Pishgoo, B., Norozi, N., Koohzadi, M., & Baesi, F. (2010). A hybrid GA and SA algorithms for feature selection in recognition of hand-printed Farsi characters. *2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*. https://doi.org/10.1109/icicisys.2010.5658728

[19] Vijay Kumar, Priyanka Gupta "Importance of Statistical Measures in Digital Image Processing" International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 2,Issue 8, 2012

[20] Moustakidis, S., Mallinis, G., Koutsias, N., Theocharis, J. B., & Petridis, V. (2012). SVM-based fuzzy decision trees for classification of high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, *50*(1), 149-169. https://doi.org/10.1109/tgrs.2011.2159726

[21] Taherkhani, A. (2010). Recognizing sorting algorithms with the C4.5 decision tree classifier. *2010 IEEE 18th International Conference on Program Comprehension*. https://doi.org/10.1109/icpc.2010.11

[22] Yiyang, Z. (2014). The design of glass crack detection system based on image preprocessing technology. *2014 IEEE 7th Joint International Information Technology and Artificial Intelligence Conference*. https://doi.org/10.1109/itaic.2014.7065001

[23] R.S. Adhikari, O. Moselhi1, A. Bagchi, Image-based retrieval of concrete crack properties for bridge inspection, Autom. Constr 39 (2014) 180–194.

[24] Online video Lectures by Prof. P.K. Biswas, IIT Kharagpur,Department of Electronics and communication engineering, <http:/nptel.iitm.ac.in/syllabus/syllabus.php?subjectid=117105079>.

[25] Lins, R. G., & Givigi, S. N. (2016). Automatic crack detection and measurement based on image analysis. *IEEE Transactions on Instrumentation and Measurement*, *65*(3), 583-590. https://doi.org/10.1109/tim.2015.2509278

[26] Li, X., Jiang, H., & Yin, G. (2014). Detection of surface crack defects on ferrite magnetic tile. *NDT & E International*, *62*, 6-13. https://doi.org/10.1016/j.ndteint.2013.10.006

[27] D. T. Pham and R. Alcock, Smart Inspection Systems: Techniques and Applications of Intelligent Vision, Academic Press, Oxford, 2003.

[28] Moganti, M., Ercal, F., Dagli, C. H., & Tsunekawa, S. (1996). Automatic PCB inspection algorithms: A survey. *Computer Vision and Image Understanding*, *63*(2), 287-313. https://doi.org/10.1006/cviu.1996.0020

[29] Kah W. Ng and Kee S. Moon, Measurement of 3-D Tool Wear Based on Focus Error and Micro-Coordinate Measuring System, Decision and Control, 1998. Proceedings of the 37th IEEE Conference, Volume: 3

[30] O'Leary, P. (2005). Machine vision for feedback control in a steel rolling mill. *Computers in Industry*, *56*(8-9), 997-1004. https://doi.org/10.1016/j.compind.2005.05.023

[31] Ross, M. (2006). Model-free, statistical detection and tracking of moving objects. *2006 International Conference on Image Processing*. https://doi.org/10.1109/icip.2006.312486

[32] Keivani, A., Tapamo, J., & Ghayoor, F. (2017). Motion-based moving object detection and tracking using automatic K-means. *2017 IEEE AFRICON*. https://doi.org/10.1109/afrcon.2017.8095451

[33] Li, Q., Li, R., Ji, K., & Dai, W. (2015). Kalman filter and its application. *2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS)*. https://doi.org/10.1109/icinis.2015.35

[34] Al-Smadi, M., Abdulrahim, K., Salam, R.A. (2016). Traffic surveillance: A review of vision based vehicle detection, recognition and tracking.

[35] Zhang, J., Song, B., & Sun, G. (2008). An advanced control method for ABS fuzzy control system. *2008 International Conference on Intelligent Computation Technology and Automation (ICICTA)*. https://doi.org/10.1109/icicta.2008.300

[36] Sari, Y., & Prakoso, P. B. (2018). Detection of moving vehicle using adaptive threshold algorithm in varied lighting. *2018 5th International Conference on Electric Vehicular Technology (ICEVT)*. https://doi.org/10.1109/icevt.2018.8628398

[37] A. K. Ray and T. Acharya. Information Technology: Principles and Applications, Prentice Hall of India, New Delhi, India, 2004.

[38] Naseera, S. (2016). Client server architecture for embedding patient information on X-ray images. *Research Journal of Pharmacy and Technology*, *9*(9), 1337. https://doi.org/10.5958/0974-360x.2016.00255.9

[39] Jing Huang, Kumar, S., Mitra, M., Wei-Jing Zhu, & Zabih, R. (n.d.). Image indexing using color correlograms. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.1997.609412

[40] Deepak S. Shete1, Dr. M.S. Chavan (2012) "Content Based Image Retrieval: Review" International Journal of Emerging Technology and Advanced Engineering ISSN, Volume 2, pp2250-2459.

[41] Malik, F., & Baharudin, B. (2013). Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain. *Journal of King Saud University - Computer and Information Sciences*, *25*(2), 207-218. https://doi.org/10.1016/j.jksuci.2012.11.004

[42] Kusano, H., Oyama, Y., Naito, M., Nagaoka, H., Kuno, H., Shibamura, E., Hasebe, N., Amano, Y., Kim, K. J., & Matias Lopes, J. A. (2014). Development of an X-ray generator using a pyroelectric crystal for X-ray fluorescence analysis on planetary landing missions. *Hard X-Ray, Gamma-Ray, and Neutron Detector Physics XVI*. https://doi.org/10.1117/12.2061547

[43] Evans, S. (1988). Quality assurance and image improvement in diagnostic radiology with X-rays. *The Physics of Medical Imaging*. https://doi.org/10.1201/9781439822081.ch3

[44] Integration of image processing and 3D techniques to simulate aesthetic dental treatments. (2014). *Biodental Engineering III*, 129-134. https://doi.org/10.1201/b17071-26

[45] Misra, Diganta & Arora, Vanshika. (2018). Image Processing on IOPA Radiographs: A comprehensive case study on Apical Periodontitis.

[46] Revett, K. (2011). An introduction to magnetic resonance imaging: From image acquisition to clinical diagnosis. *Innovations in Intelligent Image Analysis*, 127-161. https://doi.org/10.1007/978-3-642-17934-1_7

[47] Oldendorf, W., & Oldendorf, W. (1988). The MRI scanner. *Basics of Magnetic Resonance Imaging*, 89-114. https://doi.org/10.1007/978-1-4613-2081-4_7

[48] Yuan, J., & Wang, J. (2018). Compressive sensing based on L1 and hessian regularizations for MRI denoising. *Magnetic Resonance Imaging*, *51*, 79-86. https://doi.org/10.1016/j.mri.2018.04.015

[49] Kayvanrad, M. H., McLeod, A. J., Baxter, J. S., McKenzie, C. A., & Peters, T. M. (2014). Stationary wavelet transform for under-sampled MRI reconstruction. *Magnetic Resonance Imaging*, *32*(10), 1353-1364. https://doi.org/10.1016/j.mri.2014.08.004

[50] Kathiravan S and Kanakaraj J,"A Review of Magnetic Resonance Imaging Techniques",Smart Computing Review, vol. 3, no. 5, October 2013,[358-366]

[51] L.H.Thai,T.S.Hai,NguyenThanhThuy."ImageClassificationusingSupportVectorMachineandArtificialNeuralNetwork" *InternationalJournalonInformationTechnologyandComputerScience*,2012,5,32-38.

[52] Li, C., & Cheng, C. (n.d.). Imperfect tactile image classification using artificial neural network. *1991., IEEE International Sympoisum on Circuits and Systems*. https://doi.org/10.1109/iscas.1991.176041

[53] Misumi, M., Orii, H., Sharmin, T., Mishima, K., & Tsuruoka, T. (2016). Image classification for the painting style with SVM. *The Proceedings of the 4th International Conference on Industrial Application Engineering 2016*. https://doi.org/10.12792/iciae2016.046

[54] Hussain, A., & Khunteta, A. (2020). Semantic segmentation of brain tumor from MRI images and SVM classification using GLCM features. *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. https://doi.org/10.1109/icirca48905.2020.9183385

[55] Kadam, DB & Gade, Sachin & Uplane, Mahadev & Prasad, RK. (2013). An Artificial Neural Network Approach for Brain Tumor Detection Based on Characteristics of GLCM Texture Features. International Journal of Innovations in Engineering and Technology. 2. 193-199.

A SYSTEMATIC LITERATURE REVIEW ON SPEECH TO TEXT TRANSLATION

Satwinder Singh[#1], Aswin P[#2], Dilshad Kaur[#3]
[#]*Department of Computer Science and Technology, Central University of Punjab, Bathinda*
[1] satwindercse@gmail.com
[2] aswinp2610@gmail.com
[3]dilshadkaur@outlook.com

ABSTRACT— Speech is one of the most nominal way nowadays to present your ideas to the public. Speech is mostly given in one's native language. With globalization, everyone wants to listen or read the ideas or views of the presenter. So, there is huge requirement of such systems that can act as a boon in the field of speech to text translation. Different techniques and methodologies came into existence in the last decade that worked in this field of translation. This work presents a survey report on such speech-to-text translation techniques. In the field of Speech to text translation, a speech is given in a particular language and then it is converted into a required textual language. It incorporates four stages that is speech database, preprocessing, extracting features and recognition. This paper incorporates different speech interpretation projects using various methodologies for speech translation and recognition. It also highlights the advantages and disadvantages of different speech to text translation systems. The technologies related to languages can solve many problems. The speech recognition technique helps machine to track the human voice and accordingly understand languages spoken by human beings. This system makes it fruitful for the uneducated rural groups or the academically poor people to understand the speech in native languages.

KEYWORDS— Automatic Speech Recognition (ASR), speech-to-text translation system (STT), multilingual, bilingual, speech-to-text translation

I. INTRODUCTION

Speech is considered as the most usual mode of interaction between the humans. The processing of speech plays an important part in field of research on the signal processing [14]. Nowadays speech is playing a major role in the field of speech recognition. Speech recognition seen as a method which is used for extracting important information from the speech signals. Speech recognition includes different kinds of information, like the speaker's information, the linguistic data etc. This kind of information has motivated the developers for the development of the technologies that automatically process the speech in the form of enhancement of speech, the synthesis of speech, the compression of speech, the recognition of speaker, the speech recognition and cross checking. Speech recognition is of two types: dependent of speaker and independent of person who speaks. Also, Speech recognition is mostly developed for different foreign languages. The technologies related to languages can solve many problems like it can encourage the people who speaks different kind of languages for communication and data transfer [32]. The speech recognition technique helps machine to track the human voice and accordingly understand languages spoken by human beings. This system makes it fruitful for the uneducated rural groups or the academically poor people to understand the speech in native languages. The kinds of speech are made understandable by the utterances in the signal of speech which were defined in various classes as follows. The Fig.1 illustrates each class of speech.

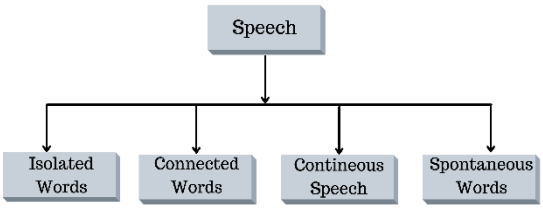


Fig. 1. Different Types of Speech

1) *Isolated Words:* The words or the utterances can be taken as single unit. It accepts at a time either one pronunciation or word [19].
2) *Connected Words:* This kind of speech is normally of two or more words. These will be included a connective in between each word [48].
3) *Continuous/Non-stop speaking words:* As the name specifies, the words will be continuous. The humans speak normally in this type of speech [51].
   Unordered words /Spontaneous: This is the kind of speech which is spontaneous and is used for usual communication [10].

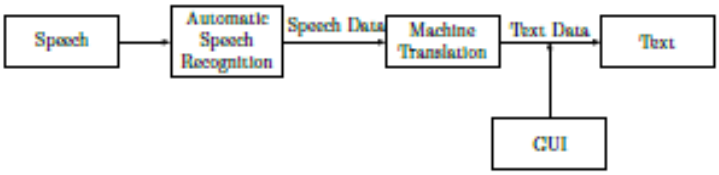


Fig. 2. General Overview of Speech to Text Translation

This survey report is divided into different sections. First section will give us a brief introduction to the speech and the speech recognition. The Speech to Text translation and Text Analysis techniques used by various researchers are mentioned in section two of Literature survey. The third section explains the research process followed in making of this literature survey. Then next the paper proceeds towards the results and discussion area which gives the facts and figures involved in making of this review paper. Finally, the fifth section provides us with the conclusion.

II. RELATED WORKS

Many research works have been conducted in the domain of speech-to-text translation and automatic speech recognition. In this survey, we focus and compiles all the recent works done in the field of speech-to-text translation. Different methodologies/models, techniques and frameworks used by different researchers are mentioned below.

A. Mel-Frequency-Cepstral-Coefficients (MFCCs)

In the processing of sound, the mel-frequency-cepstrum (MFC) is one of the techniques that represents a voice and power spectrum. MFCCs are the coefficients which aggregately Structure an MFC. They were collected from a kind of cepstral form of sound representations. The distinction between the cepstrum and the MFC is that in the MFC, the frequency groups were similarly divided on the basis of mel scale, which approaches the human hearable framework's reaction very much intently more than the directly separated frequency groups utilized in the typical cepstrum. This frequency altering can be taken for consideration for the better form of representation of the sound, like the compression of sound.

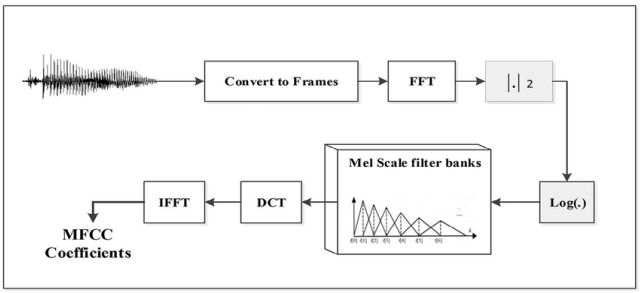


Fig. 3. MFCCs from the audio recording signals by [1]

[50] proposed a framework for fixed multiple language speech recognition technology. The goal achieved by the features like the multiple language modeling for acoustic, the dialect identification which is automatic, and the pronunciation modeling. Depending on these fresh elements, which was reasonable to understand a superior multiple language ASR framework on the resource usage stage that can manage dynamic and collection of words in multiple languages. A bunch of 12 MFCC coefficients and log-energy, along with their first-and second-order time subordinates, were extricated from a persistent time discourse signal inspected at 8 kHz in the front-end of the framework. The ASR design show the initial outcomes for five European languages.

Later [11] made the use of MFCCs for multilingual speech to text translation. Here transformation depends on the speech signal. Speech to Text framework accepts a human speech expression as an input and requires a series of words as yield. The goal of the framework was to extract, describe and perceive the data about speech. This work was executed by utilizing the Mel-Frequency-Cepstral-Coefficients (MFCCs) along with Minimum Distance Classifier and the Support-Vector-Machines (SVMs) techniques for the classification of speech. Speech expressions were pre-recorded and put in a database. This framework followed testing and training. The features were extracted by passing the samples from the database of training datasets. A feature vector was made by combining the features and it was taken as reference. Then bringing out features by utilizing the samples which were given for testing. The output of the system was given in the form of maximum similarity in comparison with both the feature vector and reference feature vector. The proposed framework accomplished the higher accuracy by using MFCC-feature extraction procedure and CDHMM-classifier.

Then [5] described a Chinese-English Speech-to-Text framework for OC16 Chinese-English Mix-ASR Challenge. This test used the recordings of Chinese-English code-mixing. The code-mixing voice was the intra-sentential exchanging of two unique languages which was in a verbally expressed expression. This proposed an exceptional testing job under ASR. For the CD-GMM-HMM acoustic models, the framework inputs 39dimension MFCC highlights in addition to 9 measurement pitch highlights to prepare the speaker-autonomous CD-GMM-HMM framework. In this test, the work

incorporated Kaldi speech recognition toolbox. The yields were consolidated from various frameworks by utilizing ROVER to accomplish a better speech. The individual subsystems worked by utilizing distinctive front-ends, acoustic models, models of languages, telephone sets etc. A final decision was taken by the ROVER. The framework gave the outputs, having the CER of 4.79% on the development set by utilizing incorporated LM while in mix with CD-LSTM-HMM framework and 4.98% CER on the development set.

Later in [44] speech to text transformation, a Turkish speech to text transformation framework was created based on "Support Vector Machines (SVM)". In the acknowledgment framework, to remove highlights of Turkish speech, SVM based classifier was utilized and "Mel Frequency Cepstral Coefficients (MFCC)" was applied. The morphological structure of Turkish, a language dependent on phonemes, has been used in the creating individual voice acknowledgments. Dissimilar to the multiclass classifiers which were utilized in the SVM-MFCC based voice acknowledgment framework, another SVM classifier framework was built up that utilized less classes in layers and expanded the quantity of multiclass layers. Another Text Comparison Algorithm was used which likewise utilized phoneme grouping to gauge closeness in word similitude estimation. Alongside these upgrades, as the preparation time frame was getting higher, execution of voice acknowledgment was improved and word acknowledgment execution was expanded. Thereafter, the exhibition of the structure was compared.

B. MBROLA Based TTS Engine

MBROLA is a speech synthesizer dependent on the link of diphones. Phonemes are taken as the input and delivers 16 bits linear speech samples, at the diphone database sampling frequency rate.

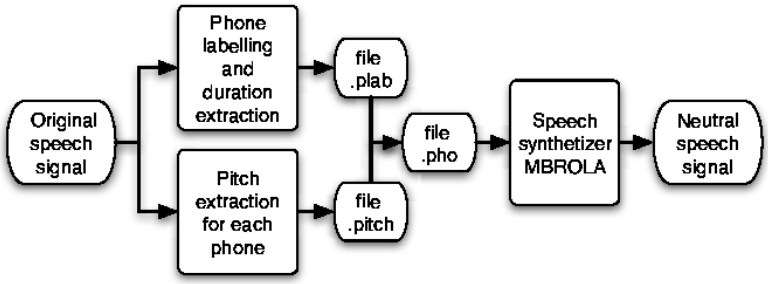


Fig. 4. Voice material creation using MBROLA by [27]

[40] introduced a framework for reading newspaper which empowers individuals to get advantage of reading a newspaper without taking the help of other person. This is a framework which gathers, divides the topics and recites online Newspapers in various kind of languages. This technology was able to handle languages such as Malayalam, Hindi, Tamil and English. The innovations such as speech recognition, synthesis of speech and the web were incorporated together under this technology. The framework utilizing natively created MBROLA based engine for TTS, which utilize di-phone as connection strategy for synthesis of the speech. The significant use of the technology was an assistive innovation for blind or handicapped, old people and even the uneducated. The application additionally encouraged the group of characters to read over the newspaper, like which gives hands free admittance to the news pages. The significant upgrades were combination of multiple languages ASR, multiple languages Text-to-Speech framework with different improved highlights like keyword identification, Intelligent/Auto customization as per client and paper free ordered headings. The coordination of ASR empowered to use this technology in a manner as complete hands-free.

C. Hidden Markov Model

This model gives a basic effective technology for demonstrating spectral vector sequences with the varying time.

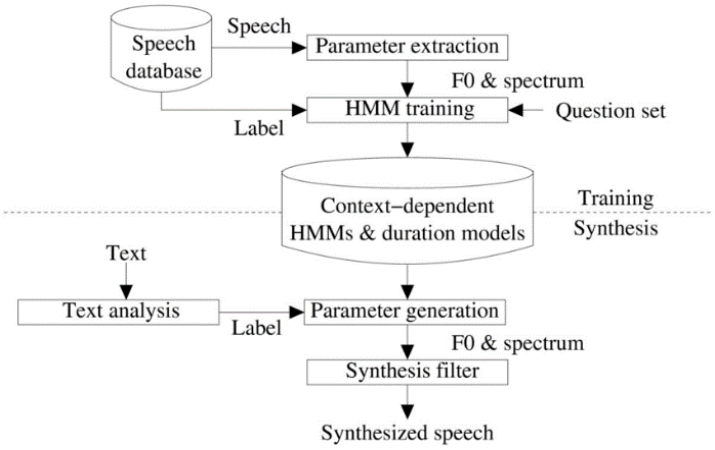


Fig. 5. Speech Synthesis Using HMM by [13]

[33] developed polyglot synthesizer speech synthesizer system which can listen like human and can overcome the language switching issue. The hidden Markov model was used in the development of polyglot. By the utilization of the model,

synthesizer was able to incorporate any content information into any of the required languages. Also, the languages which were blended with different languages was upheld by that system. In normal case the speech similarity was checked and in mixed language case, the speaker switching in between the language, switching point was focused. By using ABX Listening, test performance was evaluated and it was varied from 73% to 86%.

Then [8] studied the issue of English speech recognition to locate the most appropriate word grouping from the given section of English voice. The primary thought of this approach was to recognize English speech with Hidden Markov model. this framework proposed a novel technique to reduce boundaries in HMM. At last, exploratory outcomes presented that the proposed strategy adequately upgraded precision of English speech recognition technique. In HMMs the boundaries can be assessed naturally from a lot of information, and they are basic and computationally practical

Another way to convert speech to text for the people with benchmark disabilities by the utilization of Fast Fourier change and Hidden Markov Model system was introduced by [35]. The system also worked with advances occurring in the field of speech recognition. [4] illustrated a framework which utilizes a reproductive learning-based information. In this system, it utilizes Example Specific Hidden Markov Models to get log-probability scores of the dysarthric voice expressions to make fixed dimensional score vector representation. This is utilized as a contribution to discriminative classifier, for example, support vector machine. The framework is assessed by utilizing UA-voice information base. Also, acknowledgment precision is far superior to the regular concealed Markov model-based methodology and DNN-HMM. The productivity of the idea of the score vector illustration was demonstrated for" exceptionally low" understandable words.

D. Use of Raspberry Pi

With the new advancements in technologies, a person's life has become more refined and extremely simple. New and keen gadgets are being presented each day. As there is a probability that, there are numerous applications that can be utilized as text to speech interpreter however, it tends to be more brilliant in other way that the system can convert speech information into text. [49] designed a system that looks through web index. This application was intended to perceive the speech by an individual and enhanced information that best showed the significance of the speech structure. A similar information can be likewise changed into text by methods for GUI applications. The textual content is then shown on personal computers or other presentation gadgets associated with Raspberry Pi Board. The capacity of Raspberry Pi to connect with external environment is used for the speech processing. The objective of this [49] was to plan a gadget which simplifies the life of dazzle individuals to look out for anything. Plan of speech program with utilizing raspberry pi module and on Raspbian stage is utilized for voice based Google search and the yield of this search is given on solenoid plates and these plates are helpful for visually impaired individuals.

E. Bidirectional LSTM

It is the deep learning model which is used in speech-recognition. It advanced in the problems of neural-network design because the defects got vanished while the backpropagation was there in deep-layers.

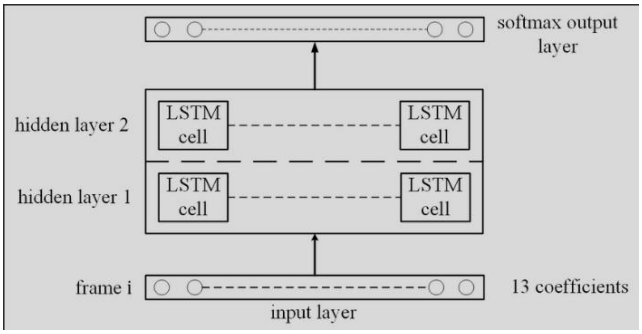


Fig. 6. LSTM RNN for speech recognition by [25]

47] broadened the speech synthesizer which was able to yield speech of numerous speakers. The multi-speaker speech synthesizer is prepared along with a corpus of the source area, to create acoustic highlights from the writings. These blended sound highlights were joined with genuine sound highlights of the source space to prepare a consideration based A2W model. Trial outcome confirms that A2W framework prepared along multi-speaker framework accomplished a critical enhancement over the gauge. The framework use multi-layer bidirectional LSTM for the encoder, and one layer unidirectional LSTM proceed with softmax layer and the sensitive area consideration component for the decoder. The mono-speaker model is illustrated in fig. 8.

The various end to end structures, and the utilization of an auxiliary connectionist temporal classification are some important aspects in speech and text transformation. [2] revealed about the various end to end structures for the interpretation of speech, where a moderate measure of speech interpretation information, for example ASR or MT sets, are accessible. Six stacked BLSTM layers fitted using the 1024 invisible dimensions were utilized for making the speech and the text encoders. The framework has shown the impact of adding an additional layer in middle for the pre-training of the system. This additional layer takes into account better joint learning and gives execution speed in high grade. Additionally, CTC misfortune can be an important factor for the end-to-end ST system since it prompts better activity just as fast assembly.

Another end-to-end technology for the translation of speech with two decoders which were designed to work with the more profound connections between the source language: sound and target language: text was proposed by [42]. The primary pass decoder creates some valuable illustrations, and the second-pass decoder coordinates with the yield of both the encoder and the principal pass decoder to produce the content interpretation in objective language. The features which are sampled furthermore gone across the bidirectional convolutional LSTM, rotating over the frequency hub with 3 sized kernels. This new arrangement of highlights was then taken care of into the bidirectional LSTM encoder. The matched source language: sound and target language: text is utilized in preparing. Some tests on a few language sets demonstrated improved execution, and offered some underlying investigation also.

Later [16] revealed that utilizing pre-prepared MT or text-to-discourse (TTS) models to transform managed information into speech and ST preparing can be more successful than performing multiple tasks of learning. Moreover, the framework exhibited that a great end-to-end ST framework could be developed utilizing just feebly managed datasets, along with artificial information from unlabeled single language content. The voice utilization can be done to upgrade execution. For the encoder, an eight layered bidirectional LSTM and for the decoder an 8 layer unidirectional LSTM with residual connections utilized by the framework. At last, the system discusses strategies for staying away from overfitting to artificial voice along a measurable removal survey.

[41] explored different avenues regarding pre-training on the datasets of changing measures, together with languages associated to the AST language, which is the source. The framework figured out best indicator of the last AST execution. It was the blundering pace of word pretrained ASR model. Three layered BLSTM is given to the yield of CNN with the hidden layer size of 512. And BLSTM also utilized for the decoding by embedding layer of 128-dimentions. The distinctions in ASR or AST execution relate on how phonetic data will be encoded in the RNN layers of the framework. The process is as demonstrated in fig. 7. They likewise show that pretraining and information expansion helps for AST.

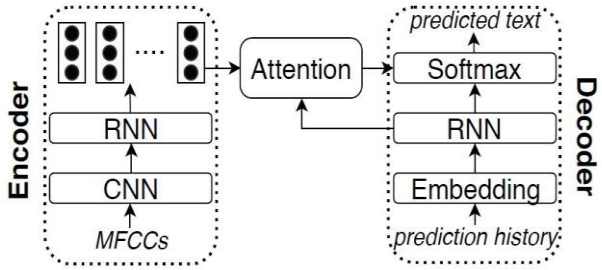


Fig. 7. Encoder-decoder architecture used for both ASR and AST. by [7]

Also, [18] proposed the principal endeavor to fabricate a direct speech to textual interpretation framework on linguistically far off language. They trained the system on English and Japanese dialect sets with general word arrangements. To manage the consideration, they build a speech interpretation with transcoding and used curriculum learning (CL) techniques that progressively trained the framework for end-to-end discourse interpretation by adjusting the decoder or encoder elements. They used TTS for information enlargement and to create relating speech expressions from the current equal content information. The approach gave huge upgraded thoughts about traditional cascade frameworks along with immediate speech interpretation technology that utilizes a solitary model without transcoding and CL with their investigation results.

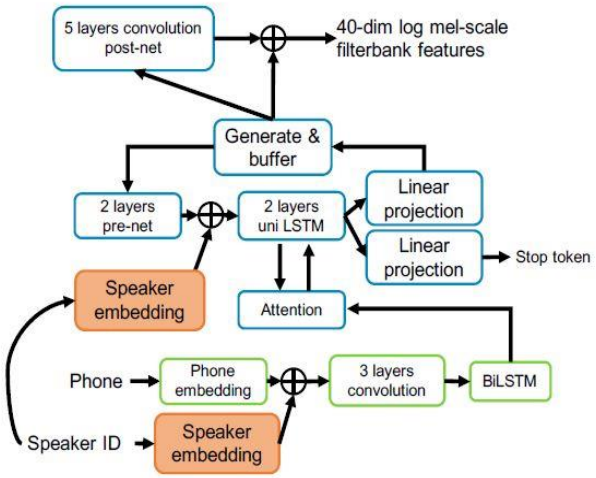


Fig. 8. The multi-speaker speech synthesizer framework overview by [47]

F. Web Speech API

Web Speech API facilitates the embedding of sound to the web applications. The Web Speech API can be divided into two sections: Speech Synthesis and Recognition. This API is used by different researchers in the speech-to-text translation field also.

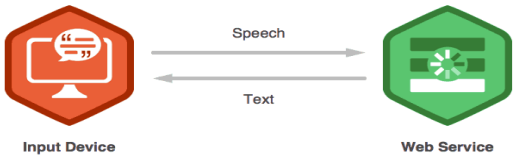


Fig. 9. Way of Web speech API working by [30]

[34] introduced a speech-to-text interface coordinated with MammoClass that permits radiologists to make some speech to a mammography report rather than typing manually in it. This new MammoClass module can take sound substance, translate it into composed words, and consequently find out the variable qualities by applying a parser to the perceived content. Whole of this interface made the use of Web Speech API to create a bridge between the written report and sound. After effects of spoken mammography reports shows that similar factors were removed for the two sorts of information: typed in or directed content.

Also, [6] designed CC Voice, a recording framework. The daily activities and speech were recorded by mobile phone. Simultaneously, it changed the recorded document into textual format by utilizing Google Speech API and saved the textual content record. This can be acquired voice proof when a client explicitly mishandled, for example, inappropriate behavior etc.

Similar kind of work was showcased [20]. It developed a methodology through half breed techniques by joining client-side and server-side administrations. The cell phone gadget was picked as a speech acknowledgment registering gadget as it is an innovative gadget that has become a need of regular daily existence while looking for data. For the speech acknowledgment the web speech API is attached within the framework. The framework supported crossover programmed speech acknowledgment framework.

Later [28] introduced an application to facilitate the ease of email composing for everybody including the disabled persons. Human voice was used instead of typing on keyboard. This application received the user input voice and compared with the database voice samples and gave outputs accordingly. The command language as the normal speech of a person was used here. The framework builds up an element that changed the speech to text for email creating and again changed text to speech for understanding the messages. Google web unit API was utilized in the framework for speech recognition. The framework was quite efficient with the comparison with different criteria like hearable distance, accent in a speech, pace, words every moment (WPM), exactness, and homophone words.

G. Deep Recurrent neural network

In the process of speech-to-text conversion, D-RNN is utilized for training the model for classification. [43] introduced a system to translate speech to text for the Bangla language by using the Deep Recurrent Neural Networks. The broken language format was carried out to decrease the training time. This particular format was based on properties of the Bangla language. Bangla Language dataset was used for the processing. For the speech recognition also, deep recurrent neural network was used. This framework reached 95% accuracy for the training data but for the testing data the accuracy was near 50%.

H. Convolutional Neural Networks and Artificial Neural Networks

ANN is a sort of computing system intended to reproduce the information handling and thoughts that a human come across. End layer of a CNN is completely associated in ANN, each and every single neuron is associated with each different neurons as demonstrated in Fig. 10.

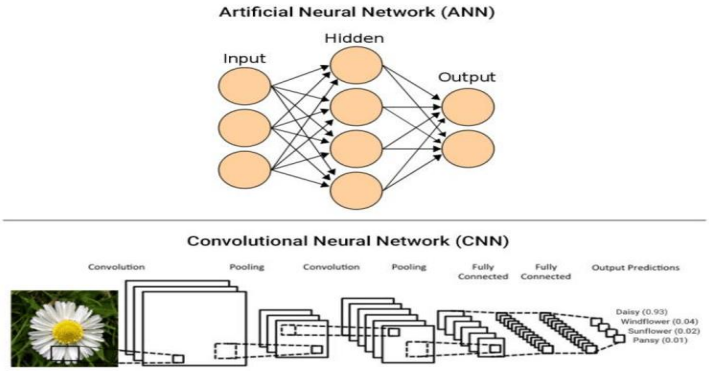


Fig. 10. The CNN and ANN difference representation by [12]

[23] outlined the spoken language works done at LIMSI laboratory, which worked with multiple languages. These exercises incorporated speech to text translation, spoken language frameworks for data recovery, the speaker and language identification, and speech reaction. The Spoken language Processing team is engaged with corpora creation and advancement. This spoken language group has taken an interest in evaluations arranged by the ARPA, in the LE-SQALE system, along with AUPELF-UREF framework for providing linguistic data. The evaluation tests were mostly done in French language. Their involvement in speech recognition for read speech has been demonstrated. The equivalent recognizer can be adjusted to various languages. It gave adequate content and the speech material were easily accessible for preparing the new language. Trained Network is simulated properly after the completion of training and check that the target output and actual output are similar. Training and testing done by utilizing the ANN. In the study it is seen that the advancements and changes in techniques kept on changing from one language to another which were not always favorable. To acquire ideal execution language specificities was must to be considered.

Later, a system which depicted sequence architecture for speech translation was given by [9]. In this the words which were spoken in source language were straightly changed into the target language on the context of sequence architecture. The experiments were conducted on publicly available data. The translation focused on one to many and many to many. The tests gave the result that MT-like objectives, utilized with no guarantees, were not viable in separating among the objective languages. In this case, for more readily uphold the yield creation was in the ideal objective language. 2D CNN layer is utilized for the processing and addition of parallel attention layers yields. The work was done for SLT from English into six languages. It showed significant enhancements while converting into comparable languages, particularly when these are upheld by lesser information. Further changes were achieved when English ASR information was utilized as an additional language.

I. CMU Sphinx tool

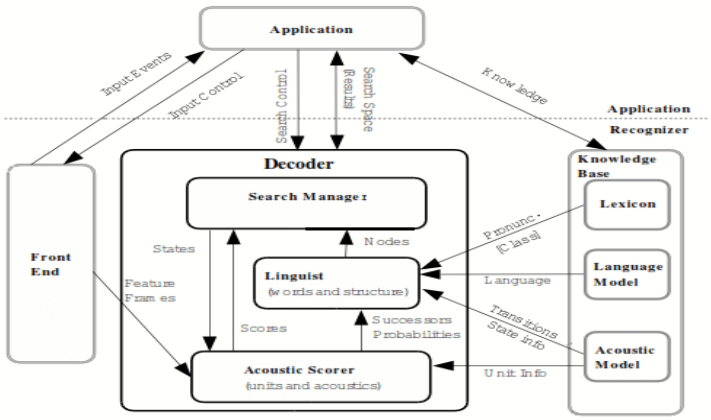


Fig. 11. CMU Sphinx design architecture by [29]

CMU Sphinx is a common term for representing collectively the speech recognition frameworks created at Carnegie Mellon University. These incorporate some series of an acoustic model trainer.

[38], made the use of CMU Sphinx structure for training and testing the sound signals. The various collection of sentences in Kannada with their four to ten-word length were studied in deep. The voice transformation framework licensed individuals to address the PC to recover data in printed structure. The quantity of letter sets in Kannada were 52. The framework explored extensibility of perceiving all alphabets along with morphological variations of communicated words in Kannada.

A similar examination of the advances utilized in little, medium, and huge glossary Speech Recognition System was introduced by [39]. The investigation showed job of language system in advancing the exactness of voice to textual format. Transformation framework tested the speech information with sentences full of noise and inadequate words. The outcomes showed an unmistakable outcome for arbitrarily picked sentences contrasted with successive arrangement of sentences.

Later, [39] gave investigation of the advances utilized in little, medium, and enormous collection of words in Speech Recognition framework. For the purpose of training and testing the framework utilized CMU Sphinx technology. The similar examination decided the advantages and liabilities of the different approaches. The analysis yielded the part of language framework in advancing the precision of voice to textual format. The framework tested the speech information with active sentences and fragmented words.

A Voice to Text framework for medical services association was proposed by [21]. This was utilized by counsellors and NGOs to record the discussion during reviews and transform it into textual format and save in storage. The framework incorporated an open-source framework. It utilized the CMUSphinx toolbox for recognition of speech. The framework allowed the recognition of multi-language. The CMUSphinx toolbox used acoustic model, phonetic word reference and

language framework. The client then stored their voice through the versatile framework. The acknowledgment and recording were done via CMUSphinx toolbox. The recorded document was stored in database as a content document.

J. Bidirectional KALMAN Filter Algorithm

It is an algorithm which gives an idea about the variables which are unknown in the time of measuring. Kalman channels have been exhibiting its handiness in different applications. Kalman channels works in a low computing power environment and is simple. Hence it is utilized in many systems and researches so far. [37] made a constant speech recognition framework that was tried continuously in noisy environment. They utilized the plan of a bidirectional nonstationary Kalman channel to improve the capacity of this real time speech recognition framework. Bidirectional Kalman channel has ended up being better in continuous motion and noisy environment. The framework presents transformation of the expressed words right away after the expression. The motivation behind this system was to present another speech recognition framework which is straightforward and stronger than the HMM based speech recognition framework. They have utilized their own database for its adaptability and TIDIGIT database for its exact correlation with the HMM based speech recognition framework. Framework was used in various noisy conditions and they got 90% Accuracy in general word.

TABLE I
APPROACHES USED BY THE AUTHORS FOR PERFORMING SPEECH-TO-TEXT TRANSLATION

Sr. No.	Approaches	List of Citations
1.	Mel-Frequency-Cepstral-Coefficients	[44], [11], [50]
2.	MBROLA Based TTS Engine	[40]
3.	Hidden Markov Model	[4], [35], [8], [33]
4.	Raspberry Pi	[5], [49]
5.	Bidirectional LSTM	[18], [41], [16], [42], [2], [47]
6.	Deep Recurrent neural network	[43]
7.	CNNs and ANNs	[9], [23]
8.	CMU Sphinx tool	[21], [39], [38]
9.	Bidirectional KALMAN Filter Algorithm	[37]
10.	Beam Forming Algorithms	[45]
11.	Web Speech API	[34], [20], [28]

K. Beam Forming Algorithms

Beamforming is utilized in sensor exhibits for directional sign transmission in signal processing method. Some of the concepts were studied for getting adequate quality for speech recognition for sounds at some distance. [45] research was carried for speech recognition where the sound capturing devices were kept at some distance. In the study two scenarios were discussed. One of them was, variety of shotgun microphones inaccessible for speakers. And the other was the use of variety of overhead mouthpieces or microphones which were kept near to the speaker. Shotgun recording Mics were normally considered to be the options in contrast to overhead microphone. In [45] considered to use shotgun microphone as components in overhead microphone. In the experimental phase, because of high directivity of microphones, they gave the enhanced Signal to Noise Ratio, very low disturbance in voice signals and high accuracy.

Other Related works

Different researchers used different methodologies or mixture of methodologies and did good work in the field of speech to text translations and recognition of speech

[52] researched on the possible ways to translate speech by utilizing the Statistical Machine Translation. The framework worked in a pipeline with speech recognition and software to create a constant voice correspondence system amongst outsiders. In this framework, TED, OPUS and the Europarl equal content corpora were utilized. These corpora were used for formative coordinating and testing. Tests including grammatical form labeling, compound parting, direct language model addition, True Casing and morphosyntactic investigation were conducted. The system assessed the impacts of assortment of information arrangements on the interpretation results by utilizing the BLEU, NIST, METEOR and TER measurements. It also gave the output in the form of metric which was reasonable for PL-EN language pair.

Voice to Text Applications as an advancement in the field of the creative writing was introduced by [17]. Results demonstrated that the utilization of these applications is developing among experts but has for more limited composition. The impacts of the utilization of these applications in the created writings showed blended outcomes.

For supporting the multiple language speaking students in cross-cultural learning project [36] came forward with a speech to text recognition and a translation system which was computer aided. To understand about the cultures and traditions the participants were engaged in interactions. As per the system, when participants talk, the STR system generated text and CAT system translated it into English language. After that it was posted on social media along with participant's native language. This system got a high accuracy for Spanish, Russian, and French languages.

To make the multilingual speech and text corpus physically is tedious task. [3] presented the general system which had the task of collecting information for text with speech for Urdu, Hindi and Manipuri which are under resourced languages. To catch the flexibility of database among the languages the content information study was done through the web crawling in three areas for example common, travel and news. In this the fundamental target of the project was to gather textual and sound information base. In all out they gathered a text corpus of 3,000,000 words along with sound corpus of hundred and fifty speakers (where fifty local speakers) of each language. Three hundred phonetically rich sentences made through text investigation were recorded by each speaker. Sound expressions were tracked at a pace of 16 kHz via the amplifier by utilizing GOLDWAVE programming instrument in a room treated by sound. This system was also accessible for improvement of multiple languages recognition techniques. Similar kind of work was also presented by [3]. In this the general procedure and encounters of textual and voice information of languages which having less material like the Hindi, Manipuri and Urdu languages were discussed. The primary targets were to gather text and speech information base which can be utilized for preparing the multiple languages distinguishing frameworks. The project work was focused on a multiple language textual and speech corpus collection for the man-machine collaboration especially in speech.

Later [26] came forward with a system to change common Bengali language to text from the sound. The technology required the utilization of the publicly released structure Sphinx 4 which has been programmed in Java and gave there the necessary procedural programming frameworks for making an acoustic framework of a custom language such as Bengali. The framework worked for word-by-word translation from the sounds of different speakers. The system used the open-source framework Audacity, to control the gathered voice recorded information through persistent recurrence profiling methods. It was used to diminish the Signal-to-Noise-Ratio (SNR), leveling of vocals, standardization etc. It ensured a 1:1 word planning ratio of every expression along with their mirror record document text which was error free. To assess the performance of framework, a sound dataset of speech was used which was already recorded information from ten separate speakers comprising of the men and women. It was custom recording using documents which they composed themself.

[26] and [7] discusses about the speech translation framework with a cross-modular bilingual word reference from the monolingual corpora. In [7] the study maps each source speech section comparing to an expressed word to its objective content interpretation. For concealed source speech expressions, the framework initially performed word-by-word interpretation on every speech fragment. The interpretation is improved by utilizing a language model and a grouping autoencoder for providing the earlier information about the objective language. Exploratory outcomes show that their unaided framework accomplished practically identical BLEU scores. It additionally gave an extraction study to look at the utility of every segment in their framework.

The possibility of utilizing a speech to-speech pipeline to encode voice tests rather than normal voice codecs, in circumstances that require high information pressure with high parcel misfortune situations were examined by [24]. A speech to-text record as a voice encoder and a book to-speech combination as a decoder was used. Furthermore, analysis was done against a standard a PCM A-law codec. It was estimated that the mistake pace of user interpreted sentences were dependent on the test which was not able to predict semantically.

TABLE II
SOME RELATED WORKS IN SPEECH TO TEXT TRANSLATION

Citation	Language sets used	Technologies and Tools Used
[16]	English-Spanish	Bidirectional LSTM encoder
[44]	Turkish	Mel Frequency Cepstral Coefficients (MFCC) LSTM encoder
[46]	English-German	ESPnet tool
[26]	Bengali	CMUSphinx framework
[20]	Bahasa Indonesia	Google Cloud Speech API, CMUSphinx API
[21]	multi-language	CMUSphinx
[45]	multi-language	beamforming algorithms
[31]	Sinhala	HMM algorithm
[35]	Multilingual	Fast Fourier Transform, Hidden Markov Model
[43]	Bangla	TensorFlow
[5]	Chinese-English	LSTM, HMM, MFCC
[22]	Bahasa Indonesia-Javanese	Fast Fourier Transform, MFCC
[2]	English-German, English-French	LSTM
[3]	Hindi, Manipuri and Urdu	GOLDWAVE software tool
[40]	Malayalam, Hindi, Tamil and English	MBROLA Based TTS Engine

[15], developed a meta learning methodology that moves information from source to target. In the meta-learning stage, values and parameters were updated to give some good initial pointers for the target. They assess the framework of meta-learning for the ST in English-German language pair and English-French language pair. The data set for these language pair was collected from MuST-C. The technique performed better than existing learning approaches. The state of art results and BLEU was improved with respective to earlier scores for English-German and English-French language pairs.

While working under machine translation, change in vocabulary is required for text translation both at source level and dictionary level. But when working for word to word and sentence to sentence translation is taking place. [46] introduces a method that can help to work out this problem. By understanding the projection between sentence-level voice encoder result along with target result, the method introduced permits change in vocabulary without any prior knowledge of dictionary or source language. Trial results showed that the technique accelerate about 20% without effecting the quality of translation.

A comparison of the Language pairs and used languages were used in related research works with their Tools and technologies are illustrated in Table II. Approaches used by the Authors for performing speech-to-text translation is explained in Table I. The usage of different approaches in speech to text translation by different authors is depicted in Fig. 14 as well. It represents the percentage of approaches used by different researchers.

III. RESEARCH PROCESS

Review paper incorporates Innovations, methodologies, evaluation and understanding the information we need, help in our queries in research and afterward making and presenting our thoughts. An important step forward in doing an intensive review or study is to comprehend the research process. The research process is carried out in a several systematic stages. Fig. 13 illustrates the stages incorporated in the research. Fig.12 depicts various databases used in searching for papers. The count of papers is mentioned in Table from year 2010-2020. Fig.15 talks about the number of papers gathered from mentioned databases.

TABLE
YEAR WISE PAPERS

Resource of Paper Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
IEEE	3	-	2	3	-	-	7	1	7	10	6
ACM	-	-	-	-	-	-	-	1	1	-	-
Springer	-	-	-	-	1	-	-	-	-	-	-
IJTRE	-	-	-	-	1	-	-	-	-	-	-

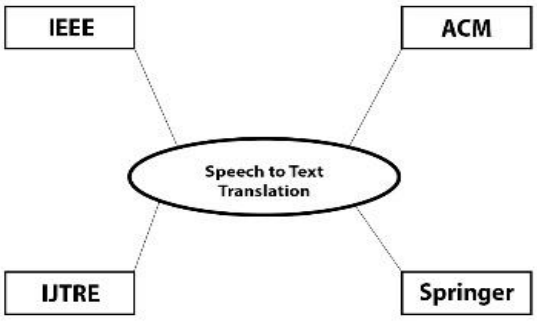


Fig. 12. An Illustration of Data Resources used in Research Process

A. Research Questions

The accompanying inquiries about the research are recorded as significant for our review:

1) How much work is done in the field of speech-to-text translation frameworks over the years 2010 to 2020 taking into account different datasets and methodologies?
2) What are the different methodologies and tools used in the field of speech-to-text translation frameworks over the years?
3) Which are the sets of languages used by the researchers for the various frameworks over the past few years?

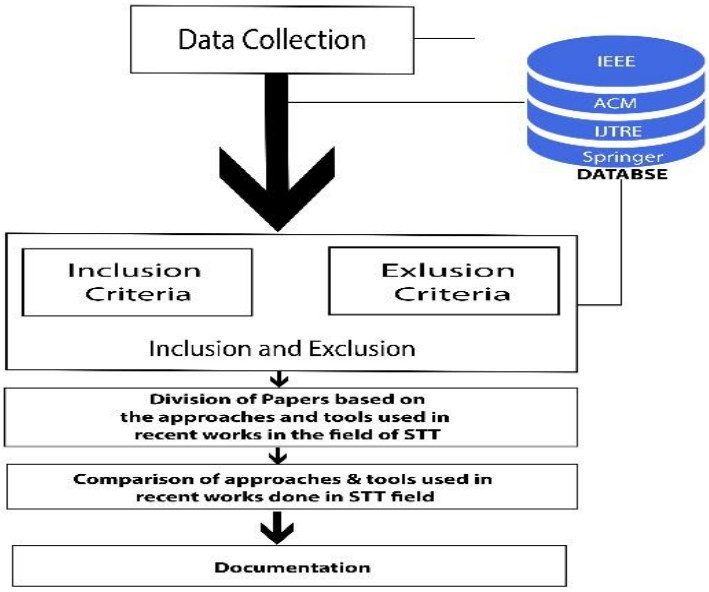


Fig. 13. Research Process

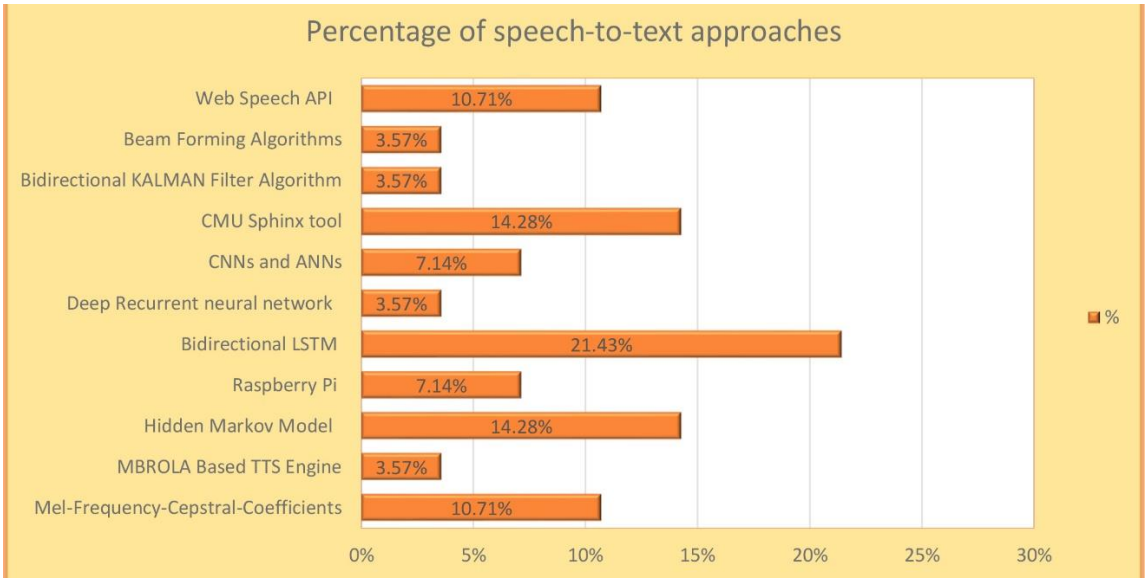


Fig. 14. Percentage of approaches used in the translation of speech-to-text

B. Inclusions and Exclusions

The various papers chosen for this review is after some filtering. Searching the relevant papers as per the topic of discussion was required. Throughout the research process, some collected papers were included and excluded according to following mentioned criteria.

1) Inclusion Criteria:
   - Papers in the IEEE, Springer, ACM and IJRTE are considered.
   - The most relevant contents are chosen by the keywords which are used mostly in the recent years.
   - The papers which are describing the various technologies related to the speech-to-text framework is considered.
2) Exclusion Criteria:
   - The papers which are not from Springer, IEEE, IJRTE and ACM were excluded.
   - The papers which are not related to computer science field is rejected.
   - Not considered the papers which having similar methodology. Tried to exclude the similarity and made each paper unique.
   - Papers published before 2010 was excluded except some important papers

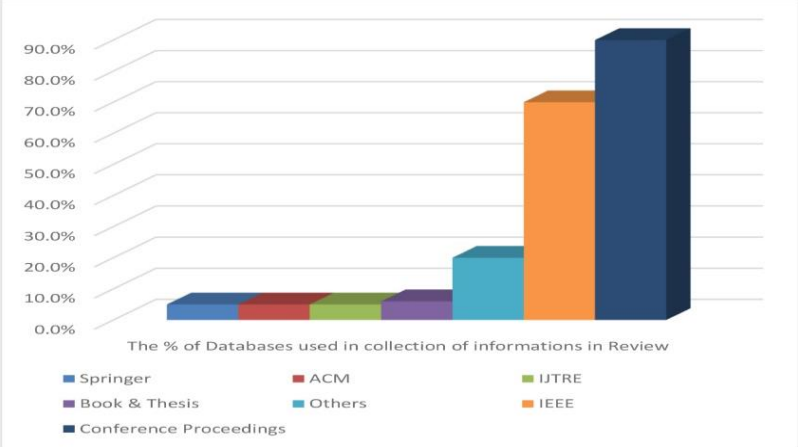


Fig. 15. The Databases Searched During Research Process

IV. Results And Discussions

The accompanying inquiries about the research are recorded as significant for our review:

1) How much works are done in the field of speech-to-text translation under the aspects of different language sets and different methodologies between 2010-2020?

A huge number of works were done in the field of speech-to-text translation under the aspects of different language sets and different methodologies between 2010 and 2020 as the Table III illustrates. Also, Table II depicts the different language sets along with tools and technologies in which speech to text translation work was carried out. Fig. 15 carries out a clear graphical representation of the work carried in the year 2010-2020 when different researches were published in different databases.

2) What methodologies and tools used in the field of speech-to-text translation frameworks over the years?

The speech-to-text is a wide area in the field of processing of speech and text. There are many queries coming in this field to solve many real-world problems. Hence it became the motivation for researchers to develop new methodologies and tools for the usage of reaching solution for the required field. Section II, Table I and Table II illustrates the kinds of approaches used by each researcher over the past years to carry out the process of speech to text translation.

3) Which are the sets of languages used by the researchers for the various frameworks over the past few years?

Every country having different languages speaking in different states. At some point people may not able to grab the things spoken by other state people. So, it is also an important thing to be considered while developing a translation framework. Hence, we can see the researchers used to develop such frame-works in appropriate language sets as illustrated in Table II.

V. Conclusion

Speech Recognition is the most important field of machine intelligence and must be considered. In this survey report, we have endeavored to give a survey of how much advancements this innovation has made in past years in the field of speech recognition and the field of speech-to-text translation. This report also gives insights of different methodologies used in speech-to-text translation for different language pairs. This paper collects and matches different approaches for speech-to-text conversions. It also provides an overview of technological perspective and advancements in speech-to-text conversions for different languages. The entire review I s carried by performing a research process for gathering the research papers. A systematic approach is followed which includes defining of research questions, inclusion exclusion criteria, predefined databases and searching the databases for the time period of 2010-2020. Multilingual speech recognition, fastest training, working environment along with the smartphone applications with more improved interfaces which will able to be mark for future work.

Acknowledgment

The authors thank Central University of Punjab, Bathinda for giving them the opportunity to carry out this study.

References

[1] Swachhata Sandesh Newsletter, Ministry of Housing and Urban Affairs (MoHUA), Govt. of India, Jan. 2020.

[2] S. Kumar, S. R. Smith, G. Fowler, C. Velis, S. J. Kumar, S. Arya, Rena, R. Kumar, and C. Cheeseman, "Challenges and opportunities associated with waste management in India," *Royal Society Open Sci.*, vol. 4, no. 3, art. 160764, pp. 1-11, Mar. 2017.

[3] S. Nanda, and F. Berruti, "Municipal solid waste management and landfilling technologies: a review," *Environmental Chemistry Lett.*, vol. 19, no. 2, pp. 1433-1456, Sept. 2020.

[4] N. Gupta, K. K. Yadav, and V. Kumar, "A review on current status of municipal solid waste management in India," *Journal of Environmental Sci.*, vol. 37, pp. 206-217, Nov. 2015.

[5] R. Joshi, and S. Ahmed, "Status and challenges of municipal solid waste management in India: A review," *Cogent Environmental Sci.*, vol. 2, no. 1, art. 1139434, pp. 1-18, Feb. 2016.

[6] A. Iravanian, and S. O. Ravari, "Types of Contamination in Landfills and Effects on The Environment: A Review Study," in *Proc. ICECAE*, 2020, pp. 1-8.

[7] Y. Pujara, P. Pathak, A. Sharma, and J. Govani, "Review on Indian Municipal Solid Waste Management practices for reduction of environmental impacts to achieve sustainable development goals," *Journal of Environmental Mgmt.*, vol. 248, art. 109238, pp. 1-14, Oct. 2019.

[8] S. Das, S. H. Lee, P. Kumar, K. H. Kim, S. S. Lee, and S. S. Bhattacharya, "Solid waste management: Scope and the challenge of sustainability," *Journal of Cleaner Prod.*, vol. 228, pp. 658-678, Aug. 2019.

[9] A. Demirbas, "Waste management, waste resource facilities and waste conversion processes," *Energy Conversion and Mgmt.*, vol. 52, no. 2, pp. 1280-1287, Feb. 2011.

[10] R. Ahirwar, and A. K. Tripathi, "E-waste management: A review of recycling process, environmental and occupational health hazards, and potential solutions," *Environmental Nanotechnology, Monitoring & Mgmt.*, vol. 15, art. 100409, pp. 1-15, May 2021.

[11] H. N. Guo, S. B. Wu, Y. J. Tian, J. Zhang, and H. T. Liu, "Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: A review," *Bioresource Tech.*, vol. 319, art. 124114, pp. 1-13, Jan. 2021.

[12] M. Mohsenizadeh, M. K. Tural, and E. Kentel, "Municipal solid waste management with cost minimization and emission control objectives: A case study of Ankara," *Sustainable Cities and Soc.*, vol. 52, art. 101807, Jan. 2020.

[13] S. A. Bini, "Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?," *The Journal of Arth.*, vol. 33, no. 8, pp. 2358-2361, Aug. 2018.

[14] R. Roohi, M. Jafari, E. Jahantab, M. S. Aman, M. Moameri, and S. Zare, "Application of artificial neural network model for the identification the effect of municipal waste compost and biochar on phytoremediation of contaminated soils," *Journal of Geochemical Explor.*, vol. 208, art. 106399, Jan. 2020.

[15] L. R. Kambam, and R. Aarthi, "Classification of plastic bottles based on visual and physical features for waste management," in *Proc. ICECCT*, 2019, pp. 1-6.

[16] M. Abbasi, and A. El Hanandeh, "Forecasting municipal solid waste generation using artificial intelligence modelling approaches," *Waste Mgmt.*, vol. 56, pp. 13-22, Oct. 2016.

[17] S. Dubey, P. Singh, P. Yadav, and K. K. Singh, "Household Waste Management System Using IoT and Machine Learning," *Procedia Computer Sci.*, vol. 167, pp. 1950-1959, 2020.

[18] A. Altikat, A. Gulbe, and S. Altikat, "Intelligent solid waste classification using deep convolutional neural networks," *International Journal of Environmental Science and Tech.*, pp. 1-8, Feb. 2021.

[19] V. Ruiz, Á. Sánchez, J. F. Vélez, and B. Raducanu, "Automatic image-based waste classification," in *Proc. IWINAC*, 2019, pp. 422-431.

[20] A. Aishwarya, P. Wadhwa, O. Owais, and V. Vashisht, "A Waste Management Technique to detect and separate Non-Biodegradable Waste using Machine Learning and YOLO algorithm," in *Proc. IEEE Confluence-2021*, 2021, pp. 443-447.

UNDERSTANDING THE APPLICABILITY AND ABILITIES OF MODERN TECHNOLOGIES FOR AUTOMATION OF WASTE MANAGEMENT

Preet Kamal Kaur[1,*], Dr. Nirvair Neeru[2]
[1,2]*Department of Computer Science and Engineering, Punjabi University, Patiala, India*
[1]preetkamalkaur06@gmail.com
[2]nirvair.ce@pbi.ac.in

ABSTRACT— India has witnessed huge rise in development of different fields since last two decades. Country has put lots of efforts to make itself grow economically. According to global goal of sustainable development, proper balance should be maintained between economic, social and environmental progress so that action performed in one field doesn't adversely affect advancement and expansion in any other field. One of the major challenges that still require an urgent attention is Waste Management. Manner to deal with daily generation of tonnes of waste (due to urbanisation) needs to be more organised and regulated. Unmanaged garbage or waste is very problematic as it can lead to contamination of environment (air, water, soil pollution) and affects the life of humans, animals as well as plants. Also, manually performing the process starting from segregation of waste to final disposal or recycling is time consuming, expensive, less accurate and risky (can cause serious respiratory and infectious diseases). This paper highlights the critical areas of concern with respect to waste management including sources, types of waste produced in country and steps involved in managing it. Utilisation of modern technologies is discussed to move in the direction of making this key process (of waste management) intelligent and automated. Applications of advanced concepts such as Artificial Intelligence, Machine Learning and Deep Learning are analysed and some of the techniques recently proposed in the area are tabulated.

KEYWORDS— Sustainable Development, Waste Management, Environment, Artificial Intelligence, Machine Learning, Deep Learning, Landfilling, Recycling, Classification

INTRODUCTION

Accomplishing the goal of sustainable development along with continuous rise in population, economic growth, industrialisation and urbanisation, is the major challenge that India is facing today. As per the statistics of 2020, India is generating near about 147,613 metric tonnes of solid waste per day and this quantity is expected to become triple by 2031 [1]. Waste is basically a substance or material that is unwanted and not usable. It is a product which is rejected and thrown away after use (has no value). Origin, content and hazard associated with the waste vary according to the source and activity/task through which it is being generated. There are various factors affecting the generation as well as composition of waste such as varying lifestyles, standard of living, variety or diversity in country with respect to traditions, cultures, religion, social behaviour, climate etc. [2]. On the basis of waste generation source, it is generally classified into certain number of categories like commercial, residential, industrial, institutional, waste from healthcare organisations, agriculture, construction or demolition [3]. Types of waste generated from different sources are summarised in the Table I below:

TABLE VIII
TYPES AND SOURCES OF WASTE

Source from which the waste is generated	Type of waste
Commercial and Industrial	Chemicals, glass, wood, plastics (thick and thin), metals, food and packaging waste, paper, ashes, cardboard, fabric, electrical and electronic waste etc.
Institutional	Waste generated from government, educational, sporting, financial (& many more) organisations which may include wood, plastic, food, electronics, metals, cardboard, paper etc.
Residential	Glass, metals, plastic, cardboard, paper, leather, food and yard wastes, ashes, tires, batteries, packaging items such as cans, miscellaneous material like old shoes, mattresses, bags, broken cooking pots, baskets.
Agriculture	Waste from forests, residues of crops, weeds, pesticide containers, leaf litter, sawdust, spoiled food, fertilizers, animal waste, litter from poultry etc.
Health Care Facilities	Bandages, syringes, masks, gloves, drugs, napkins, plastic, diapers, urine bags, paper, food waste etc.
Construction and Demolition	Copper wires, steel, concrete, bricks, dirt, rubber, glass, plastic, plasters, metal, ceramics etc.

Although, manual interference is essential to deal with different types of waste generated in the country but, for regular processing and analysis of daily basis waste related data, capabilities of machines should complement the on-going efforts of humans. In recent years, focus has already started to shift on the automation strategies, however, major problems and issues which need to be addressed more by technological solutions are: comparison of speed at which waste is being produced and at which it is being controlled (minimized), improper segregation and transportation of waste for its final disposal/recycling, lack of awareness among people for cleanliness. This study is an outline about the emerging developments in the field. It will help to understand the crucial necessity of waste management and also, role of computer science to execute the process in efficient and convenient manner.

Paper is centred on utilisation of advanced technology for automated and intelligent working of waste management system. Section 2 discusses problems faced by the country due to unmanaged waste and ideal hierarchy of waste management that needs to be followed. Steps involved in management of waste are also briefed. Section 3 highlights the role of computer science (AI, Ml and DL) for systematic completion of tasks and activities associated with waste management process. In section 4, latest techniques proposed in the field are reviewed and tabulated on the basis of different parameters. Section 5 illustrates the conclusion of paper.

WASTE MANAGEMENT

Even if there is a huge rise in the development of many sectors, waste management still remains a major issue in India that needs to be handled with much more care and awareness. Due to rapid growth in population, India is already fighting with exhaustion of natural resources. Waste generated daily through several human activities pose adverse effect on environment as well as on health of population [4]. Also, there is a lack of sanitary landfilling i.e. waste is dumped as it is, anywhere along the roads just in the outskirts of a city or town [5]. These landfills become major source and cause of pollution/contamination which in turn give rise to many more related problems. All water, air and soil are affected negatively by these landfills. Through unplanned waste decomposition, leachates and hazardous toxins are generated. For example, dumped electrical and electronic waste material when gets warmed, release dangerous chemicals. Also, many greenhouse gases are released such as methane, carbon dioxide that directly make a smooth path for environmental destruction which includes varying climatic conditions, global warming, unexpected observations of temperature etc. [6]. To achieve the global goal of sustainable development and environmental safety, critical analysis is required with respect to management of waste [7]. It can be effectively carried out on the basis of 3R principles which are Reduce, Reuse and Recycle [8]. Ideal hierarchy to be followed for proper waste management in the country is shown in Fig. 1 below:

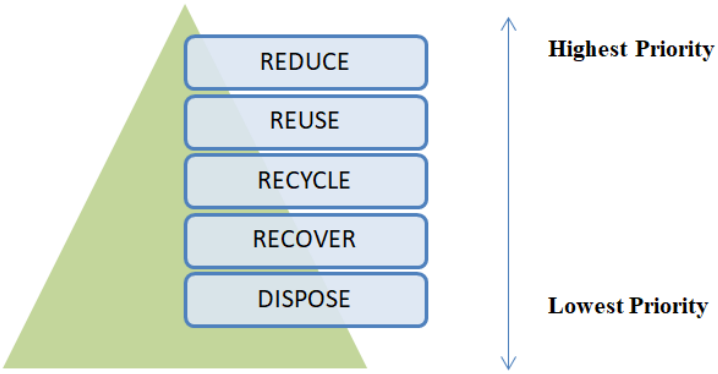


Fig. 4 Ideal Hierarchy of Waste Management

In systematic waste management, after generation of waste, it is categorised or sorted at its origin place. Then, it is stored, collected, transported, processed or treated (any type of thermal or biological reprocess). At last, final disposition takes place i.e. recycling and landfilling is there. Valuable resources can be recovered from waste using effective strategies [9, 10]. Regular monitoring of waste is required to be done using technology, planning and proper mechanisms with due regard to waste related laws in the country.

INTELLIGENCE AND AUTOMATION IN MANAGEMENT OF WASTE

Incineration, composting, recycling, treatment and landfilling processes are very crucial part and base of waste management. These mechanisms consume so much of resources in terms of manpower, money and material. Manually executing these methods at all steps of waste management (as discussed above), has many flaws like inefficiency, lesser accuracy, environmental destruction and health risk for humans. For example, separation of biomedical waste physically can lead to chronic infectious diseases. Release of harmful substances and emission of greenhouse gases are dangerous for both environment and health of an individual. Moreover, whole cycle of manual monitoring and treatment of waste becomes excessively costly and time consuming [11, 12]. To resolve the above mentioned problems, it is clear that modern technology and waste management should go hand in hand. To fasten the speed of country towards sustainable environment and to solve the complex non-linear problems, major advancements in the field of computer science are

Artificial Intelligence (AI), Machine Learning (ML) and Deep Learning (DL) [13]. Base and relation of AI, ML and DL is represented below in Fig. 2:

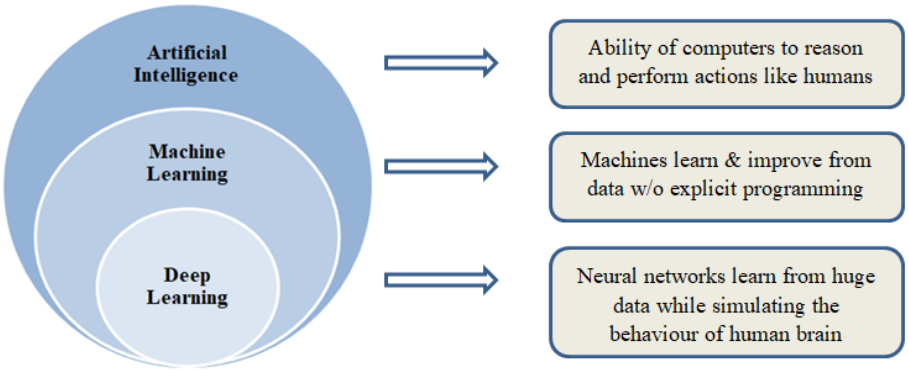


Fig. 5 Relationship of AI, ML and DL

Prediction accuracy can definitely be improved with the use of these intelligent methods for automation of waste management process. Consumption of resources can also be reduced as machines substitute manpower and time is also saved. Many complex tasks which can't be implemented otherwise can be executed conveniently by intelligent machines and algorithms. Besides, risk of disease (or any type of infection) spread through physical touch of waste or garbage is also decreased [11, 14]. Techniques depicting application of emerging technologies for waste management are discussed in the following section.

LATEST TECHNIQUES SURVEYED

Certain approaches proposed recently by researchers towards automating the implementation of various activities involved in waste management are reviewed in this module. They are listed in tabular manner (Table II).

TABLE IX
COMPARATIVE REVIEW FOR THE LATEST TECHNIQUES (AI, ML AND DL) IN THE FIELD OF WASTE MANAGEMENT

Type of waste	Task being performed	Dataset	Technique/Functionality used	Evaluation Parameters	Author(s) & Year
Plastic Bottles	Classification of plastic bottles to predict whether they can be recycled or not, on the basis of physical as well as visual features	Numeric dataset generated through extraction of physical and visual characteristics from plastic bottles	Certain Color based segmentation algorithm for color detection and k-Nearest Neighbour (kNN) classifier for predicting the color of plastic (visual feature extraction) For physical characteristics, tactual sensor is used to measure pressure and weighing machine to measure weight of plastic All extracted features combined to form training data (numeric) Supervised ML methods such as Support Vector Machine, KNN, Decision Tree and Logistic Regression applied on the data to predict the class (recycled or non-recycled)	Accuracy	Kambam and Aarthi (2019) [15]
Municipal Solid Waste (MSW)	Forecasting or prediction of future municipal solid waste generation based on time series data related to previous waste generating events	Time series data (monthly) of MSW generation for period of eighteen years (1996 to 2014), Logan city, Queensland, Australia	Based on collected real time waste generation data (time series), intelligent forecasting models are constructed 4 algorithms are used : Support Vector Machine (SVM), k-NN, Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) Ability of these 4 methods to predict the quantities of future waste generation is analysed with the help of statistical evaluation metrics	Coefficient of Determination (R^2) Root Mean Square Error (RMSE) Mean Absolute Percentage Error (MAPE) Mean Absolute Error (MAE)	Abbasi and El Hanandeh (2016) [16]

Household Waste	Framework for smart collection and decomposition of waste in order to achieve smart green society	csv file of 100 samples (sensor values) is created to simulate the working of framework	Smart dustbin (2 levels): At house level, segregation done into bio-degradable & non-biodegradable waste If poisonous gas detected or any of the dustbins found full, alert message sent to facility supervisor and dustbin moved out of the house (via messenger) At society level, non-biodegradable waste classified into different categories (message sent to municipal corporation if threshold exceeds) & bio-degradable waste turned into compost Simulation done using csv file of 100 samples (each instance with sensor values of 3 parameters depicting level of bio-degradable, non-biodegradable waste and poisonous gas) k-NN used for prediction of sending an alert message using these 3 parameters	Accuracy	Dubey et al. (2020) [17]
Solid Waste (Paper, Plastic, Glass and Organic Material)	Waste categorisation according to the contents	400 pictures of waste (100 of each category) taken using digital camera	Pictures taken from external environment are pre-processed and resized to 224*224 pixel size For feature extraction and learning process, 4 layer & 5 layer Deep Convolutional Neural Network (DCNN) applied Performance is analysed using different evaluation metrics	Accuracy Precision Sensitivity Specificity F-Score	Altikat et al. (2021) [18]
Glass, Plastic, Paper, Metal, General Trash	Sorting of Waste	TrashNet dataset	Images (for different classes of waste) taken from TrashNet are pre-processed, resized and augmented (in augmentation, size of dataset is increased by performing different operations on each image) 4 neural network architectures applied for classification: Inception, ResNet (Residual Network), VGGNet (Visual Geometry Group) and combined Inception-ResNet Performance of each classifier is evaluated and compared	Std. Deviation Accuracy Mean Accuracy	Ruiz et al. (2019) [19]
Glass, Plastic and Metal	Separation of non-biodegradable waste	Collection of 450-500 images for each category of waste (some images are self-clicked)	Approx. 500 images of each category are collected and then, labelling done using YOLO (You Look Only Once) format with the help of software Model is trained on these labelled images and files of training code are produced Now, at the first stage, testing is performed on random images Later on, testing is carried out using Laptop webcam and finally, using Raspberry Pi Camera	Accuracy	Aishwarya et al. (2021) [20]

Parameters considered for the analysis are: type of waste (e.g. plastic, glass, metal, household, solid waste etc.), task being performed (such as waste categorisation, collection, decomposition, forecasting etc.), dataset used for simulation or implementation, functionality of the technique, evaluation metrics for the assessment and year of publication [15-20]. Kambam and Aarthi [15] applied color based segmentation and k-NN for extraction of features from plastic bottles and supervised ML methods to find whether they are recyclable or not. Abbasi and El Hanandeh [16] used ML methods such as ANFIS, ANN, k-NN to forecast the future waste generation on the basis of previous waste producing events (real time series data). Dubey et al. [17] proposed a concept of smart dustbin to collect or decompose the waste and also, segregating it as bio-degradable and non-biodegradable (k-NN used to send alert messages). Alitkat et al. [18] utilized D-CNN to categorise waste based on pictures taken from external environment. Ruiz et al. [19] exploited the capabilities of neural network architectures (Inception, ResNet, VGGNet, Inception-ResNet) to sort the trash. Aishwarya et al. [20] focussed on separation of non-biodegradable waste (YOLO technique for labelling of images to train the model and laptop web cam, raspberry pi-camera to test the model). Use of intelligent approaches for solving the problems related to collection, sorting, transportation, recycling or decomposition of waste, has started creating a path towards smart green society but still, functionalities and parameters of these advanced techniques can be explored more to enhance the performance. Also, carrying out these automated tasks in real world scenarios with different types of practical challenges demands more critical and empirical perspective of thinking in research.

CONCLUSION AND FUTURE SCOPE

Due to urbanisation, quantity of waste production in the country is going higher day by day. Shifting from manual to automated waste management is basically a need of an hour. Scope and benefit of applying algorithms (Artificial Intelligence, Machine Learning as well as Deep Learning) in this field has been discussed and it can be observed that tasks or activities involved in this long process (of waste management) can be performed quickly, conveniently and safely with the help of advancement in science and technology. Latest techniques that have been proposed in the direction of automation are reviewed in this paper. In future, work could be done on balancing the trade-off between accuracy of a technique and its efficiency (in terms of time and cost). Also, data deficiency issue needs to be focussed to solve real world complex problems related to Waste Management. Country will definitely continue to remain aligned with the global goal of sustainable development and environmental protection if humans work harder to make best possible use of abilities possessed by modern technology.

REFERENCES

[1] Swachhata Sandesh Newsletter, Ministry of Housing and Urban Affairs (MoHUA), Govt. of India, Jan. 2020.

[2] S. Kumar, S. R. Smith, G. Fowler, C. Velis, S. J. Kumar, S. Arya, Rena, R. Kumar, and C. Cheeseman, "Challenges and opportunities associated with waste management in India," *Royal Society Open Sci.*, vol. 4, no. 3, art. 160764, pp. 1-11, Mar. 2017.

[3] S. Nanda, and F. Berruti, "Municipal solid waste management and landfilling technologies: a review," *Environmental Chemistry Lett.*, vol. 19, no. 2, pp. 1433-1456, Sept. 2020.

[4] N. Gupta, K. K. Yadav, and V. Kumar, "A review on current status of municipal solid waste management in India," *Journal of Environmental Sci.*, vol. 37, pp. 206-217, Nov. 2015.

[5] R. Joshi, and S. Ahmed, "Status and challenges of municipal solid waste management in India: A review," *Cogent Environmental Sci.*, vol. 2, no. 1, art. 1139434, pp. 1-18, Feb. 2016.

[6] A. Iravanian, and S. O. Ravari, "Types of Contamination in Landfills and Effects on The Environment: A Review Study," in *Proc. ICECAE*, 2020, pp. 1-8.

[7] Y. Pujara, P. Pathak, A. Sharma, and J. Govani, "Review on Indian Municipal Solid Waste Management practices for reduction of environmental impacts to achieve sustainable development goals," *Journal of Environmental Mgmt.*, vol. 248, art. 109238, pp. 1-14, Oct. 2019.

[8] S. Das, S. H. Lee, P. Kumar, K. H. Kim, S. S. Lee, and S. S. Bhattacharya, "Solid waste management: Scope and the challenge of sustainability," *Journal of Cleaner Prod.*, vol. 228, pp. 658-678, Aug. 2019.

[9] A. Demirbas, "Waste management, waste resource facilities and waste conversion processes," *Energy Conversion and Mgmt.*, vol. 52, no. 2, pp. 1280-1287, Feb. 2011.

[10] R. Ahirwar, and A. K. Tripathi, "E-waste management: A review of recycling process, environmental and occupational health hazards, and potential solutions," *Environmental Nanotechnology, Monitoring & Mgmt.*, vol. 15, art. 100409, pp. 1-15, May 2021.

[11] H. N. Guo, S. B. Wu, Y. J. Tian, J. Zhang, and H. T. Liu, "Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: A review," *Bioresource Tech.*, vol. 319, art. 124114, pp. 1-13, Jan. 2021.

[12] M. Mohsenizadeh, M. K. Tural, and E. Kentel, "Municipal solid waste management with cost minimization and emission control objectives: A case study of Ankara," *Sustainable Cities and Soc.*, vol. 52, art. 101807, Jan. 2020.

[13] S. A. Bini, "Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care?," *The Journal of Arth.*, vol. 33, no. 8, pp. 2358-2361, Aug. 2018.

[14] R. Roohi, M. Jafari, E. Jahantab, M. S. Aman, M. Moameri, and S. Zare, "Application of artificial neural network model for the identification the effect of municipal waste compost and biochar on phytoremediation of contaminated soils," *Journal of Geochemical Explor.*, vol. 208, art. 106399, Jan. 2020.

[15] L. R. Kambam, and R. Aarthi, "Classification of plastic bottles based on visual and physical features for waste management," in *Proc. ICECCT*, 2019, pp. 1-6.
[16] M. Abbasi, and A. El Hanandeh, "Forecasting municipal solid waste generation using artificial intelligence modelling approaches," *Waste Mgmt.*, vol. 56, pp. 13-22, Oct. 2016.
[17] S. Dubey, P. Singh, P. Yadav, and K. K. Singh, "Household Waste Management System Using IoT and Machine Learning," *Procedia Computer Sci.*, vol. 167, pp. 1950-1959, 2020.
[18] A. Altikat, A. Gulbe, and S. Altikat, "Intelligent solid waste classification using deep convolutional neural networks," *International Journal of Environmental Science and Tech.*, pp. 1-8, Feb. 2021.
[19] V. Ruiz, Á. Sánchez, J. F. Vélez, and B. Raducanu, "Automatic image-based waste classification," in *Proc. IWINAC*, 2019, pp. 422-431.
[20] A. Aishwarya, P. Wadhwa, O. Owais, and V. Vashisht, "A Waste Management Technique to detect and separate Non-Biodegradable Waste using Machine Learning and YOLO algorithm," in *Proc. IEEE Confluence-2021*, 2021, pp. 443-447.

STATISTICAL ANALYSIS OF COMORBIDITIES IN DECEASED AND RECOVERED COVID-19 PATIENTS USING CHI-SQUARE TEST

Amreen Ghumann[1], Dr. Brahmaleen K. Sidhu[2]
[1,2]*Department of Computer Science and Engineering,*
Punjabi University Patiala

ABSTRACT - Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. COVID-19 was declared a pandemic by the World Health Organization on 11 March 2020. There was no evidence about lasting of covid-19 for so long, and still it is incurable. In this work, we use statistical feature selection approach and test the relationship between presence of underlying medical problems like high blood pressure, heart and lung problems, diabetes, obesity etc., and seriousness of COVID-19 illness in patients. The dataset pertaining to various comorbidities and mortality in COVID-19 patients used was retrieved from the official website of the Mexican government and the non-parametric chi-square statistical test was performed using Python's sklearn library. Dependency of attributes like patient's sex, age, use of tobacco and medical conditions like pneumonia, pregnancy, diabetes, asthma, immune suppression, hypertension, cardiovascular disease, obesity, renal chronic disease, need of icu etc. was computed against the target attribute, death. The results thus obtained can be further used for a better understanding of the nature of COVID-19 illness, as well as for building machine learning based classification and prediction models.

KEYWORDS- Chi-square, COVID-19, feature selection, statistical analysis.

INTRODUCTION

Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. COVID-19 was declared a pandemic by the World Health Organization on 11 March 2020. Medical practitioners and scientists are yet to ascertain the actual cause of the breakout of disease. Research for a definite treatment of COVID-19 is also underway. A proper understanding of the effects and symptoms of the disease seen in the patients will facilitate the medical developments required to contain this pandemic. In our previous work, clustering was performed on various kinds of symptoms experienced by COVID-19 patients (Ghumann & Sidhu, 2021). The machine learning based clustering approach resulted in two clusters, thereby highlighting that the disease affected patients in different ways and different medical attention may be required in different clusters. This paper is presented in the direction of aiding the medical understanding of the effects of the disease with respect to different medical underlying conditions already present in patients encountering COVID-19.

World Health Organization has observed that patients with underlying medical problems like high blood pressure, heart and lung problems, diabetes, obesity etc., develop serious COVID illness (World Health Organization, 2021). This paper proposes the use of statistical tests for understanding the dependence of seriousness of COVID-19 in patients on the cardiovascular, respiratory and other comorbidities. Statistical tests such as chi-square are used to determine whether a predictor variable has a statistically significant relationship with an outcome variable.

Chi-square test can be used for two purposes. As a goodness of fit, chi-square test determines if sample data belongs to a population. As a test for independence, it compares two variables in a contingency table to see if they are related. It is non-parametric in nature, so can be implemented for finding out that whether the data belonging to some category is dependent on other classified feature or is it totally independent (Kothari, 1990). When data is categorical, this test can be used to distinguish between hypothetical population and the original data. Thus, chi-square test has applicability at numerous places. It is a handy, simple and quick way to determine whether data outputs are unexpected or expected. It is a strategy for inferring about dispersion of a quantity or determining if two parameters in a population have a relation. The interpretation is built on the statistical distribution, which is proportional to the number of degrees of freedom (d.f.) (Shukla, 2021).

BACKGROUND

Chi-square test has several applications, some of them are stated below (Gajawada, 2019) :-

a) The Chi-squared test can be used to assess the goodness-of-fit of trained regression model on training, validation and test data sets.
b) It can be used to decide if records follow a developed theory for probabilistic model, like the Normal or Poisson distribution.
c) The Chi-squared test is used to discover the variable domain in which the measured values would satisfy the statistical properties for sets of data that follow quantitative ranges namely the Normal, Poisson or Binomial distributions.
d) A Chi-squared test is being used to examine the relationship between two classified parameters, like Age and Income, for whom the values have been monitored in a study, are distinct from each other.
e) As under the null hypothesis the remnant inconsistencies are self-reliant, uniformly scattered Regular factors like the Pearson particles in certain Simplified Linear Regression Analysis comply a (scaled) Chi-square distribution, showing a high accuracy of the model.

f) The Deviant behaviour statistic, that could be used to draw comparisons of the log likeliness of recursive logistic regression, needs to follow a chi-squared distribution underneath the hypothesis test that introduces correlation factors which does not raise the fitness of the model. As an outcome, while selecting one out of the two approaches the most convenient may be preferable.

g) When the measured and predicted values are equal, the G-test for correlation diminishes to a Chi-squared test for predictive relevance.

CHI-SQUARE AS A TEST OF INDEPENDENCE

The chi square test helps to determine that whether the two features are related or not. For illustration, if one wants to recognize that a new medication is helpful in combating fever, the chi square test could indeed assist in determining this (Kothari, 1990).

Several steps performed for carrying out this test are as follows:-

a) Describe the Hypothesis
Null Hypothesis (H_0) states that two parameters are independent of one another.
Alternate Hypothesis (H_A) states that two parameters are dependent on one another.

b) Table of contingencies
A table presenting the allocation of one attribute in rows and another attribute in columns. It's being used to investigate the relationship among two parameters.

c) Determine the Expected Value
Since according to the null hypothesis, two factors are independent it can be specified about A and B that

$$P(A \cap B) = P(A) * P(B)$$

Where, P stands for probability of the occurrence of an event, and A and B are two events.

d) Calculation of chi-square value
Find out the Chi-Square value by first combining the observed and calculated expected values in a table and then putting the values in the formula given below:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where O = observed frequency
E = expected frequency

e) Accept or deny the Null Hypothesis
If the calculated chi-square value is less than the critical value at the given degrees of freedom, the null hypothesis is accepted, otherwise it is rejected (Gajawada, 2019).

PROPOSED METHODOLOGY

Appearance of two or more ailments in the same person is referred to as comorbidity.

A. Dataset

The data was retrieved from the official website of Mexican Government[1]. It was in Spanish, so was translated into English. The dependency of 21 attributes was tested on the target attribute. Id, sex, patient type, time to hospital, intubed, pneumonia, age, pregnancy, diabetes, copd, asthma, inmsupr, hypertension, other disease, cardiovascular, obesity, renal chronic, tobacco, contact other covid, covid_res and icu were the features whose dependency has been computed against the target attribute which is death.

B. Pseudo code of proposed approach

a) Import the required libraries. Python's pandas, sklearn, feature_selection and chi2 are the main packages imported in this work.
b) Load data from comma separated format into the data frame.
c) Perform preprocessing steps:
   a. remove null values
   b. remove duplicate records
   c. encode categorical data
d) Extract dependent and independent variables using slicing.
e) Compute the chi-square values using the chi2 function of sklearn package.
f) Compare chi-square values of the individual features and arrange them in descending order to identify top five features.

RESULTS

The pseudocode described above was implemented in Python. Python was deemed suitable ass the language permits vectorized operations on large datasets. The computed chi-square values of the attributes implemented for this proposed work are given below (Table 1). These are in the descending order.

[1] https://www.oecd.org/mexico/governmentofmexicousefullinks.htm

It was observed that hypertension, diabetes, patient's gender, time lapse between detection of COVID-19 and hospitalization, and chronic renal disease are top five features that are highly related to the seriousness of COVID-19.

Table 2: Chi-square values of the independent variables

Attributes	Chi-square Score
hypertension	1.096244e+02
diabetes	2.028661e+02
sex	5.195052e+02
time_to_hospital	1.517814e+03
renal_chronic	1.717220e+03
copd	1.992962e+03
obesity	2.010190e+03
inmsupr	2.180900e+03
tobacco	2.328618e+03
cardiovascular	2.368251e+03
asthma	2.472347e+03
covid_res	3.972340e+03
other_disease	4.028888e+03
pneumonia	6.894361e+03
patient_type	1.491166e+04
pregnancy	1.367120e+05
age	2.980655e+05
contact_other_covid	1.091823e+06
icu	2.133839e+06
intubed	2.137182e+06

CONCLUSION

The pandemic caused by COVID-19 required immediate concern in order to have better quality of life in future. This paper presents a statistical feature selection approach to test the relationship between presence of underlying medical problems like high blood pressure, heart and lung problems, diabetes, obesity etc., and seriousness of COVID-19 illness in patients. The dataset pertaining to various comorbidities and mortality in COVID-19 patients used was retrieved from the official website of the Mexican government. Chi-square test was performed using Python. Dependency of attributes like patient's sex, age, use of tobacco and medical conditions like pneumonia, pregnancy, diabetes, asthma, immune suppression, hypertension, cardiovascular disease, obesity, renal chronic disease, need of icu etc. was computed against the target attribute, death. The results thus obtained can be further used for a better understanding of the nature of COVID-19 illness, as well as for building machine learning based classification and prediction models.

REFERENCES

[1] Gajawada, S. K. (2019, October 4). *Chi-Square Test for Feature Selection in Machine Learning.* Retrieved September 8, 2021, from Towards Data Science: https://towardsdatascience.com/chi-square-test-for-feature-selection-in-machine-learning-206b1f0b8223

[2] Ghumann, A., & Sidhu, B. (2021). Machine Learning Based Clustering Of Covid-19 Symptoms. *Proceedings Of ICRITO.* Noida: IEEE.

[3] Kothari, C. (1990). *Research Methodology Methods and Techniques.* Jaipur: New Age International .

[4] Shukla, A. (2021, July 26). *Chi-Square Statistic And Chi-Squared Distribution.* Retrieved September 5, 2021, from Towards Data Science: https://towardsdatascience.com/tagged/chi-square-test

[5] Health Organization. (2021). *Coronavirus disease (COVID-19) pandemic*. Retrieved September 27, 2021, from World Health Organization: https://www.who.int/emergencies/diseases/novel-coronavirus-2019

FACE RECOGNITION TECHNIQUES: A SURVEY

Puneet Kaur[#1] and Dr. Taqdir[*2]
[#]Research Scholar
Department of Computer Science
Guru Nanak Dev University Amritsar, Punjab
puneetcsc.rsh@gndu.ac.in
And
[*]Assistant Professor
Department of Computer Science and Engineering
Guru Nanak Dev University Amritsar, Regional Campus Gurdaspur, Punjab
taqdir_8@rediffmail.com

ABSTRACT– Over the past few years, face recognition is an interesting and a vibrant research area. Machine learning, artificial intelligence and deep learning methods have been gaining good performance in area of face recognition in these days. The primary objective of a face recognition system is to recognize the human individuals from images, videos and surveillance systems etc. In this survey, we highlighted major challenges of face recognition which include pose variation, occlusion, low resolution, illumination and aging. We summed up various recent approaches of face recognition and listed their advantages and limitations along with future direction of each method. Finally, we discussed future directions to enhance the performance of face recognition system for addressing the major problems of face recognition.

KEYWORDS: Face recognition, surveillance, Deep learning, Occlusion

I. INTRODUCTION

In today‘s era, biometric has an important role to recognize an individual for authentication, security and anti-terrorism purposes. Biometric identifies physiological and behavioral characteristics of a human which involves face, fingers, iris, DNA, palm, gait, signature, voice etc. Face is a most important attribute to recognize an individual. Face recognition meant to be as a strong authentication system. Face recognition plays a major role in various security purposes like investigation of crimes, fraud votes, attendance proxy and unauthorized access etc. In the recent years, there is a rapid growth in face recognition techniques. The challenges of pose variation, occlusion, illumination, low resolution are addressed by the latest researches. Despite of achieving great performance over the face recognition challenges, there is still challenging problems and accuracy of system suffers from factors like pose variation, low resolution, occlusion, expression and aging. Moreover, there is no any common technique which addresses all the challenging issues of face recognition.

Several methods are reported in the literature for face recognition can be categorized as holistic methods, geometric methods and hybrid methods. The holistic method works on entire face as input whereas geometric methods consider the local features of a face image not the whole face. Some of the major holistic methods are PCA, LDA, IDA etc.[1][2] and the geometric methods are LBPH, HOG,LBP etc.[3][4][5]. The hybrid methods combine both geometric and holistic methods in order to achieve good results. These traditional approaches are not robust under the conditions of illumination, low resolution, occlusion and other face recognition.

With the advent of machine learning, deep learning has a great success over the challenging problems of face recognition. CNN, DeepID, DeepID2, DeepID2+,DeepID3, DeepFace, Face++, FaceNet, and Baidu are popular deep learning face recognition systems[6]. In last few years, several researches are conducted on face recognition using deep learning methods. Despite considerable success of deep learning based face recognition, there are still results can further improved under the conditions of unconstrained environment and other face recognition. To our knowledge, there is still no any common technique that gives satisfactory results by considering the all face recognition challenges.

We summarise the main contributions of this paper as (1) We introduced the face recognition and gave an overview of face recognition process. (2) Secondly, we presented some major challenging problems of face recognition. (3)Later on, we summarized the recently used face recognition techniques. (4) At last, we concluded the paper with discussing the directions of future research.

II. PROCESS OF FACE RECOGNITION

Face recognition system identifies persons in images, videos and in real time systems. Face recognition system extract features such as eyes corners, mouth, nostrils, nose tip and then locate the appropriate match in pre stored database.

Face recognition process

1. Face detection
2. Feature extraction
3. Feature selection
4. Face recognition and verification

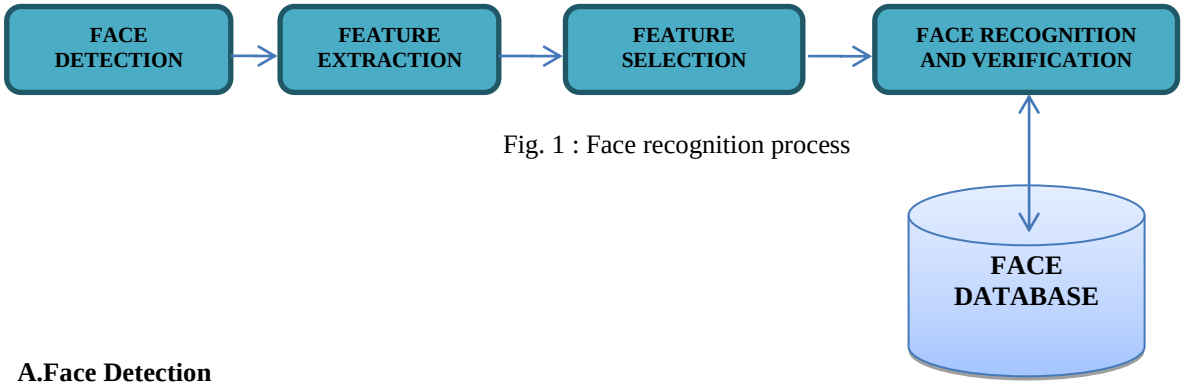


Fig. 1 : Face recognition process

A.Face Detection

The first step is pre-processing, which consists of many types of operations, such as image registration, scaling, face normalization, reducing the effect of background noise, detection and resizing, all of which affect the face recognition accuracy[1] . The pre-processing step of image involves reducing the image variability and cropping the frontal view of image and adjusts images by using some standard algorithms. After the processing of image classification techniques are applied and dataset is being trained. After that face is localized by using the knowledge of trained dataset.

B. Feature Extraction

Feature extraction is the second phase, which can be achieved by using powerful transformation approaches. The image dimension can be reduced to a smaller dimension by retaining significant features[7]. Basically there are two approaches for feature extraction i.e. holistic approach and geometric based approach. Holistic methods work on the global features of an image. It takes entire image as an input .In global (holistic)approach, entire information is extracted with a single vector from the whole face image[8]. Geometric methods use location of local features of image such as eyes, mouth, and nose are extracted from the image. This approach mainly deals with the spatial correlation uniting the profile (i.e. face) features, also we can simply that dimensional layout of the facial attributes [9].

C. Feature Selection

After the feature extraction, the next step is to select the principal features and reduce dimensions from large feature space generated by feature extraction phase. Various algorithms are used for this purpose like PCA, SVD, and SVM etc.

D. Face Recognition

After the face detection and feature extraction next step is to compare the features with the faces from database. Face recognition is divided into two steps. In first step face is compared with database to find the likely match. Then verification is done in order to make acceptance or rejection decision. For this purpose distance is measured between test face and database face. Various Techniques used for this purpose are correlation filters, CNN, KNN etc.

I. CHALLENGES OF FACE RECOGNITION SYSTEMS

As the modern face recognition systems achieved a high accuracy in face recognition, but there are still many challenges in face recognition systems in unconstrained environments like low face recognition, pose variation , illumination ,occlusion , aging etc. the major challenges in are described in this section.

A. Low Resolution

Low resolution face recognition is the key challenge in face recognition system. When the image is taken in unconstrained environments, there are many chances that resolution is image is low and blur. In surveillance systems, where image is taken from long distance, the captured images are of low resolution. Recent techniques of face recognition system achieved much high accuracy but as the resolution of probe images gets lower, the accuracy of face recognition system also decline.

B. Pose variation

In remote monitoring and in uncooperative environments, pose variation is a huge challenge for face recognition system. For example to identify a person or spy in a crowded place or to detect a terrorist and thieves in markets or public places. The solution to this problem to have all poses images of the subjected face. Despite deep learning approaches generates the images with multi-poses with good recognition rates but still the recognition performance in unsatisfactory in different datasets.

C. Occlusion

Occlusion is a key challenge in face recognition systems. Occlusion means to hide the face with masks, hats, sunglasses, hand, scarfs etc. As COVID-19 pandemic, where everyone have to wear the mask so there is an opportunity for criminals to do crimes by masked up their faces. Masked face recognition is a key challenge for face recognition systems because much of the major facial features are hidden like nose, corners of mouth, tip of chin etc. which are essential features for face recognition.

D. Illumination

The variability in lightening effects or illumination is much likely to occur in uncontrolled environments. Traditional or classical methods for face recognition are unable to cope with varying illumination conditions. Recent techniques give significance to deal with illumination factors and achieved good results.

E. Aging

Aging changes the shape, look or texture of the human faces. It is hard to recognize a face for humans as well as machines with the time passage. This is a major problem in face recognition because the ID cards or images on passport generally not updated. There are still very few techniques to cope with age invariant face recognition.

II. RECENT APPROACHES FOR FACE RECOGNITION

In last few years a high performance achieved by machine learning as well as deep learning techniques in face recognition. As per recent literature, we summarize some latest techniques in this section.

[10] Introduced a deep coupled Res-Net for low resolution face recognition. The proposed model has one trunk network and two small branch networks. The networks are trained with different resolution images. Coupled mappings of High resolution images to low resolution features performed by the branch networks.

A good accuracy is attained by this model in low resolution, but accuracy gets lower as the resolution degrades. Moreover, there is no work on occlusion, multi-pose conditions with low resolution.

[11] This work suggested an unsupervised face domain transfer for low resolution face recognition. This paper focuses on domain shift of high resolution images to low resolution probe images. The proposed framework did not use any label information for the probe set.

This work has good performance in unsupervised methods without labelling the LR probes. But still accuracy can be further improved.

[12] In this paper, a non-parametric model for low resolution face recognition in resource constrained environments is proposed. This model uses machine learning methods Pixelhop++, SSL and Active learning.

This model is only for resource constrained environment and can be generalised to resource rich environments. Accuracy can be further improved having different resolutions. In addition other factors like pose variation or occlusion in resource constrained environment can also be considered.

[13] Two methods are developed where batch training is done in first method with 50% low resolution images and in the second method, each batch is trained with specific resolution. Model is evaluated with three resolution protocols which is high, low or mid resolution. The proposed model trained with several resolutions at once and achieved a fine accuracy with lower resolutions.

The inter-class distances between LR and HR images should be further minimized. This model is only for low resolution face recognition, so other parameters occlusion, aging, multi-pose can also be added.

[14] The prime focus of this paper is multi pose face recognition using cascade alignment network and incremental clustering. Six CNN networks are used to refine facial landmarks. Six models are separately trained for facial landmarks like left eyebrow, right eyebrow, left eye, right eye, mouth and nose. This model used 21 different poses for experiment without any variation and got a good recognition rate.

There is no constraint on face pose which leads inferior performance in significant pose variation as well as there are challenges of variation in occlusion, expression and illumination.

[15] A multiview 3D face recognition is proposed in this paper which used a nose tip heuristic pose learning method where whole face is aligned through transformation using knowledge obtained from nose tip. The suggested method learnt pose variation by alignment in different planes xz , yz and xy planes.

The computational complexity of the system is further reduced and overall system performance can be enhanced.

[16] The authors employed a recurrent regression neural network for cross pose face recognition in this paper. The suggested framework is for cross pose face recognition in images and videos. This method takes on image pose and predicts the sequential poses and takes an entire sequence as input for video based face recognition.

[17] The method introduced by this paper is invariant to illumination, expression, partial occlusion and pose variation. Component based face recognition is used five statistical pattern matching tools.

As the different algorithms are used a great number of false alarms caused by combination of multiple algorithms.

[18] This paper addressed pose variation and masked face recognition. Method focused on diverse local features and discriminant facial parts for face recognition and attained an effective accuracy. Further performance of this method can be improved and other face recognition challenges can be considered.

[19] In this work, the challenges of pose variation, expressions and illumination with low resolution face recognition are addressed. A residual network is used to test the recognition accuracy with six different face resolutions. As the recognition gets decline with minimum low resolution images, the minimum resolution images are reconstructed by super resolution.

[20] The focus of attention of this paper is masked face recognition during COVID 19 pandemic. The suggested method removes the masked region of the face firstly and then uses deep CNN to extract the features from other facial regions.

Other face recognition issues pose variation; aging, low resolution with masked face can be a consideration.

Author	Techniques	Accuracy (%)	Dataset	Advantages	Disadvantages	Future scope
[10]Ze Lu et.al(2018) IEEE	DCR(Deep Coupled ResNet)	98.7 98.0	LFW SC-FACE	Better performance in low resolution face recognition	No work on occlusion, pose variation factors with low resolution.	Other FR challenges can be included.
[11]Sungeun Hong et.al(2019) IEEE	GFA(generative face augmentation)+SRA (spatial resolution adaptation)	52.25	SC Face	Robust in low resolution face recognition	Does not work on occluded faces.	image-level domain transfer can improve performance even further
[12]Mozhdeh Rouhsedagaht et.al (2021) Elsevier	Pixelhop++SSL, Active learning	89.48 83.30	Multi-PIE LFW	minimize the training sample number and achieve relatively high accuracy	Does not handle pose variation	The proposed principle could be updated to resource-rich environments and high-resolution images
[13] Martin Knoche et.al (2020)	ResNet50+CR batch training (BT) + Siamese network CR training (ST)	97.72	LFW	Trained several resolution at once and attain good accuracy in all resolutions	Accuracy decline in minimum resolutions and does not handle pose variation, occlusion, aging conditions.	The inter-class distances between LR and HR images can be further reduced.
[14]Yepeng Guan et.al (2020) Springer	LCCAN	96.82 96.58 97.88	CASPEAL-R1 CFP Multi-PIE	Superior performance in multi pose face recognition	Challenges of variations in occlusion, expression and illumination.	Challenges of variations in occlusion, expression and illumination
[15]Naeem Ratyal et.al(2019) Hindawi	d-MVAHF-SVF	100 95.4 99.3 99.8	GavaDB Bosphorous UMB-DB FRGC V2.0	3D facial feature information integrated in the d- MVAHF images significantly enhanced the face recognition accuracies	Complex computation	Reduce the number of synthesized multiview face images to decrease complexity and enhance system performance.
[16]Yang Li et.al (2018) Elsevier	RRNN	95.6	Multi-PIE	Robust in Cross pose face recognition	No work on low resolution and occlusion	Performance can be enhanced further.
[17]Sushil Kumar Paul et.al (2020) Springer	CSQ+ HuMI+AbsDiPwPs+GDVs	96	BioID	Invariant to shape, pose ,expression ,illumination and partial occlusion	Combination of multiple algorithms may lead to false alarms.	To control the false alarm with multiple databases.
[18] Qiangchang Wang et.al (2021) IEEE	DSA-Face+ PSCA+ASL	98.85 98.69 96.24	LFW CFP VGG FP	Robust for pose variation and masked face recognition	Performance can be improved further.	Other face resolutions challenges can be considered.
[19] Jianguo Shi et.al (2021) IEEE	ResNet 18	98.26(14*11) 80.87(7*6)	UMIST	Better recognition in low resolution	Recognition degrades with minimum resolution	Further enhancement in recognition rate for critical low recognition
[20] Walid Hariri	VGG-16+ AlexNet+ ResNet-50	91.3 88.9	RMFRD SMFRD	Efficient recognition during COVID 19 pandemic	Pose variation and low resolutions factors are not included.	Pre-trained model can be used to improve accuracy

Table1: Summary of recent face recognition approaches

III. CONCLUSION

At this juncture, face recognition is a prevalent research area due to its enormous digital and scientific applications. Face recognition is one of the major research areas in machine learning, pattern recognition, artificial intelligence and deep learning. In this paper, we discussed various challenges of face recognition like pose variation, aging, occlusion, low resolution and aging. We summarize various recent methods to rectify these challenges and limitations and future scope of each method. After examining previous studies, we found that there are still unsatisfactory and less consistent

results and further performance can be improved. The literature survey shows that there is no any common single technique which addresses all the challenging issues of face recognition. Future research is needed to rectify these challenges and to increase the overall efficiency of system. All challenges considered by a single face recognition system can also be a future direction of research. Future studies should also consider factors like face recognition in resource constrained and unrestricted environments.

REFERENCES

[1] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face Recognition Systems : A Survey," 2020.

[2] M. K. Halidu, P. Bagheri-Zadeh, A. Sheikh-Akbari, and R. Behringer, "PCA in the context of Face Recognition with the Image Enlargement Techniques," *2019 8th Mediterr. Conf. Embed. Comput. MECO 2019 - Proc.*, no. June, pp. 10–14, 2019, doi: 10.1109/MECO.2019.8760162.

[3] A. Ahmed, F. Ali, and A. Ahmed, "LBPH Based Improved Face Recognition At Low Resolution," *2018 Int. Conf. Artif. Intell. Big Data*, pp. 144–147, 2018.

[4] W. Yang, X. Zhang, and J. Li, "A Local Multiple Patterns Feature Descriptor for Face Recognition," *Neurocomputing*, 2019, doi: 10.1016/j.neucom.2019.09.102.

[5] H. Ta, M. Nhat, and V. T. Hoang, "Feature fusion by using LBP , HOG , GIST descriptors and Canonical Correlation Analysis for face recognition," *2019 26th Int. Conf. Telecommun.*, pp. 371–375, doi: 10.1109/ICT.2019.8798816.

[6] C. Rosenberger, *Deep Biometrics*.

[7] M. A. Abuzneid and A. Mahmood, "Enhanced human face recognition using LBPH descriptor, multi-KNN, and back-propagation neural network," *IEEE Access*, vol. 6, no. c, pp. 20641–20651, 2018, doi: 10.1109/ACCESS.2018.2825310.

[8] S. K. Paul, S. Bouakaz, C. M. Rahman, and M. S. Uddin, "Component-based face recognition using statistical pattern matching analysis," *Pattern Analysis and Applications*, vol. 24, no. 1. pp. 299–319, 2021, doi: 10.1007/s10044-020-00895-4.

[9] G. Singh and A. K. Goel, "Face Detection and Recognition System using Digital Image Processing," *2nd Int. Conf. Innov. Mech. Ind. Appl. ICIMIA 2020 - Conf. Proc.*, no. Icimia, pp. 348–352, 2020, doi: 10.1109/ICIMIA48430.2020.9074838.

[10] Z. Lu, X. Jiang, S. Member, and A. Kot, "Deep Coupled ResNet for Low-Resolution Face Recognition," vol. 14, no. 8, 2018, doi: 10.1109/LSP.2018.2810121.

[11] S. Hong and J. Ryu, "Unsupervised Face Domain Transfer for Low-Resolution Face Recognition," *IEEE Signal Process. Lett.*, vol. PP, no. c, p. 1, 2019, doi: 10.1109/LSP.2019.2963001.

[12] M. Rouhsedaghat, Y. Wang, S. Hu, S. You, and C. J. Kuo, "Low-resolution face recognition in resource-constrained environments R," *Pattern Recognit. Lett.*, vol. 149, pp. 193–199, 2021, doi: 10.1016/j.patrec.2021.05.009.

[13] G. Rigoll, "I MAGE R ESOLUTION S USCEPTIBILITY OF F ACE R ECOGNITION," 2021.

[14] Y. Guan, J. Fang, and X. Wu, "Multi-pose face recognition using Cascade Alignment Network and incremental clustering," *Signal, Image Video Process.*, vol. 15, no. 1, pp. 63–71, 2021, doi: 10.1007/s11760-020-01718-z.

[15] N. Ratyal *et al.*, "Deeply Learned Pose Invariant Image Analysis with Applications in 3D Face Recognition," vol. 2019, 2019.

[16] Y. Li, W. Zheng, Z. Cui, and T. Zhang, "US CR," *Neurocomputing*, 2018, doi: 10.1016/j.neucom.2018.02.037.

[17] S. K. Paul, S. Bouakaz, C. M. Rahman, and M. S. Uddin, "Component-based face recognition using statistical pattern matching analysis," *Pattern Anal. Appl.*, vol. 24, no. 1, pp. 299–319, 2021, doi: 10.1007/s10044-020-00895-4.

[18] Q. Wang, G. Guo, and S. Member, "DSA-Face : Diverse and Sparse Attentions for Face Recognition Robust to Pose Variation and Occlusion," vol. 6013, no. c, pp. 1–11, 2021, doi: 10.1109/TIFS.2021.3109463.

[19] J. Shi, T. Liu, N. Chen, J. Liu, Y. Dou, and Y. Zhao, "Low Resolution and Multi-pose Face Recognition based on Residual Network," vol. 2021, pp. 1587–1593, 2021.

[20] W.Hariri, "Efficient Masked Face Recognition Method during the Covid 19 Pandemic"2021.

DIABETES MELLITUS DETECTION USING MACHINE LEARNING TECHNIQUES

Manbir Singh,Nirvair Neeru

Department of Computer Science and Engineering Punjabi University, Patiala 147002, India
Email - Manbir98@gmail.com, nirvair.ce@pbi.ac.in

ABSTRACT— Diabetes mellitus(DM) is a disease due to high blood sugar. Diabetic retinopathy is a common complication of diabetes and the leading cause of blindness in the working-age populace. Early detection using AI could be a cost-effective alternative. The adoption of these techniques could help patients for early referral to Specialized Care and improve life quality by tackling conditions as soon as possible. Machine learning techniques help medical professionals by remote monitoring as well as predicting the possibility of diabetes. Using technology for diagnosis in such a manner can drastically improve case discovery rates. This review is a detailed rundown of diabetes mellitus detection and techniques which are in a phase of development and implementation.

KEYWORDS— DM-Diabetes mellitus, ML- Machine learning, ANN Artificial Neural Networks, CNN-Convolutional Neural Networks, DME- Diabetic Macular Edema, FFNN Feedforward neural networks, KNN -K-Nearest Neighbour, RF-Random Forest, NAFLD Nonalcoholic fatty liver disease, EMR- Electronic Medical Records

I. INTRODUCTION

Experts estimate that 463 million individuals over the world have diabetes. This number is expected to increase to 700 million by 2045, and yet we only have records of around 200 million cases, meaning about 50% of the patients remain undiagnosed [1]. Diabetes contributed at least 760 billion dollars in health expenditure in 2019 [2].

One in six people diagnosed with Diabetes is from India. The number of cases in 2019 was 77 million. While in another time this may have only been a grave public health issue, today it also presents as an opportunity to collect a vast amount of data in the form of electronic medical records to be used in various applications of Artificial Intelligence. Some ML techniques are proficient in finding patterns in data, such as case-based reasoning, neural networks, random forest, etc. These can help in population risk control.

Products are being designed to bring a measure of self-diagnosis to people. Products such as glucose sensors and insulin pumps integrated with smartphone applications are on the verge of release in the free market. Artificial Intelligence-based applications provide accuracy comparable to the human expert with the ease-of-use that allows laymen patients to operate them.

Diagnosis and prediction are not the only applications for artificial intelligence. The study of management strategies for diabetes, and diseases in general, with the aid of artificial intelligence, is sorely neglected. Singla et. al. have reported on the advancement of an insulin delivery system with inbuilt ML algorithms to predict hypoglycemic and hyperglycemic excursions and take automatic action by varying the dose.[3] Treating type 2 diabetes is very complicated as there are different types of treatment options and have to be used carefully. choice of medication and doses depends upon numerous factors such as wait insulin resistance, body mass, and other complicated situations of the patient.

II. Diabetes Mellitus

It is a disorder regarding how the body converts food into energy. Normally the body converts sugar into glucose and then transfers it to the bloodstream. The pancreas produces insulin which is needed for the transfer of glucose from the blood to cells. Types of diabetes are as follows:

A. *Type I Diabetes*

Type 1 Diabetes is a condition in which the body's immune system attacks insulin-producing cells in the pancreas.

B. *Type 2 diabetes*

In type 2 diabetes pancreas releases some insulin but it is not enough to be used by the body or the body doesn't utilize it. Overweight individuals have a high chance of getting type 2 diabetes as more insulin is required due to insulin resistance.

III. Artificial Intelligence Approaches Used In The Detection Of Diabetes

A. *Convolutional neural network*

There are multiple layers of neurons with a convolutional layer of neurons which locate small inputs like image and filter and use the whole image and share parameters across the image. Each CNN layer detects the presence of features and makes a model of the relationship between those features. It usually learns by using a backpropagation method. It requires a large amount of data to train and it is very resource-intensive and many parameters require proper adjustment while using the model for training.

B. *Random forest*

Random forest creates an ensemble of decision trees and in each tree random features are considered. The root node is determined and split. It generally produces really good results. Important features can be easily identified and it usually

avoids overfitting and outliers given a sufficient data set. It can only be used for discrete outcome classification and regression problems.

C. *K-Nearest Neighbour*

KNN algorithm is a supervised machine learning algorithm that uses k-nearest neighbors to categorize data into several classes without making any assumption about distribution. It is very easy to implement and understand and can be applied to classification as well as regression problems. It is very resource-intensive and can be susceptible to outliers.

IV. Literature Survey

In selected studies of diagnosing and detecting diabetes, researchers have used various techniques as given below:

A. *Deep Learning*

Gadekulla et al.(2020) developed a deep neural network model in concurrence with Principal Component Analysis (PCA) and firefly algorithm. The model can classify the diabetic retinopathy set with an accuracy of 96% and was trained using the UCI machine learning repository.

Singh and Gorantla (2020) propose a novel DMENet algorithm using Hierarchical Ensemble of Convolutional Neural Networks (CNNs). The two stages of the algorithm operate on preprocessed colored fundus images (of the eye). The first stage detects patterns of Diabetic Macular Edema (DME) in the eye, and in the next stage collates the positive cases and groups the images based on severity. The dataset used for training and classification were IDRiD and MESSIDOR. Observations include Accuracy of 96.12%,

Ghani et al. 2019 developed Retinal Image Classification Using Artificial Neural Networks (ANNs). Instead of analyzing the full spectrum color images, only the green channel is extracted as it provides more relevant details in the context of the patterns hoped to be detected. Optical disk segment area shape features are computed for accelerating the classification. Feedforward neural networks (FFNN) is used for classification. HRF image database is used. The accuracy observed is 100%.

Varun Gulshan et al. (2016) created a deep learning algorithm that identifies diabetic retinopathy and diabetic macular edema in the retinal fundus images. The learning model used was Deep CNN photographs. The data set included 128,175 retinal images graded each by 3 to 7 members of a 54 member panel of ophthalmologists for suitability in signs of DR and diabetic macular edema. Test data used EyePACS-1 data set and Messidor-2 data set. Study outcomes include Sensitivity of 97.5% for EyePACS-1, 96.1% for Messidor-2 .

Carson Lam et al. (2015) develop a Deep Learning based method for grading retinal images with diabetic retinopathy severity using a limited set of training data without hard-coding. Data set is a collection of manually created image patches from the public image dataset Kaggle Retinopathy. Data set includes 243 images which after labeling by 2 ophthalmologists yielded 1050 patches for training and 274 patches for validation. The study resulted in Accuracy of 98% and ROC 99% with GoogLeNet validation by using the patch images.

Hsiao-Hsien Rau et al. (2016) developed a model for predicting the event of liver cancer within 6 years of diagnosis with type 2 Diabetes. Learning models are ANN and Logistic Regression. Data set of 2060 diabetic patients from the Nation Health Insurance Research Database (NHIRD) of Taiwan was used, out of which 1442 cases were used for training the model. The study concluded that LR performance was inferior to Artificial Neural Network for predicting which diabetes mellitus patients would be diagnosed with liver cancer within the next 6 years. Sensitivity: 75.7%, Specificity: 75.5%, Area Under The Curve: 87.3% were found.

B. *K-Nearest Neighbou*r

Ali et al. (2020) utilized multiple KNN variations to identify and classify diabetes mellitus. The dataset is provided by the American Diabetes Association. A total of 5000 samples were divided into 4900 for the training and 100 for testing. Certain variations of KNN were found to be more precise than others (Coarse and Cosine). Fine KNN had the highest precision.

Aminah and Saputro (2019) designed the iris-based diabetes mellitus prediction framework utilizing ML algorithms. The image processing framework comprises techniques that aid in the development of images and further algorithms. Extraction of features texture from the image done by Gray Level Co-Occurrence Matrix (GLCM) approach. Groups are categorized using the K fold cross-validation process and confusion matrix. There are 27 subjects of which 11 are diabetic. The precision was found to be 85.6%.

Carter et al. (2019) discovered the usage of elemental analysis of diabetic toenails and machine learning techniques for the diagnosis of type 2 diabetes. The concentration of various metals such as aluminum, caesium, vanadium, nickel, zinc, etc is shown to be different between diabetic patients and healthy patients. Several artificial intelligence algorithms are used to utilize toenails concentration details like gender, age, and other background features.

Samant and Agarwal (2019) provide a comparative analysis of various machine learning vision techniques used in the computer-assisted diagnosis of type 2 diabetes. Physiological parameters and fundus image features are used. 334 subjects are split into two classes diabetic and non-diabetic. The diabetic group is further split into three exclusive subgroups based

on the duration of the diabetes mellitus condition. 85% precision was achieved by the use of three ensemble classifiers: the bagged tree, boosted tree, and subspace KNN.

C. *Random Forest*

Datta et al. (2019) proposed a machine learning-based system for the initial diagnosis of diabetes mellitus. The data set is collated from a German body of 2314 subjects (1396 female). Of these, 941 subjects (500 females) have metabolic syndrome. Random forest, gradient boosters, logistic regression, etc., and an ensemble model were used. Curve field values of up to 0.90 were obtained by the ensemble classifier.

Xu and Wang (2019)[15] created a prediction model for adult-onset diabetes. Feature Weighting Random Forest is used for variable selection. Classification is done by XG Boost. Pima Indians Diabetes Database is used for validation. The precision of the model is 93.75%

Beatriz Lopez et al. (2017) implemented artificial intelligence techniques to aid the identification of relevant Single Nucleotide Polymorphisms according to Type-2 diabetes. A tool for evidence-based recommendations on risk prediction was also built. Random Forest algorithm (RF) is used for the SNPs search. K Nearest Neighbor is used to measure similarity according to the relevance of attributes in RF. The data set contained 96 SNPs regarding type 2 diabetes each for 677 subjects (248 diabetics). AUC of 0.89 for risk prediction is observed. Random forest is found to be very useful for doctors for prediction of type 2 diabetes and Single Nucleotide Polymorphisms.

Wei-Hsuan Lo-Ciganic et al. (2015) applied ML techniques to determine the association between oral hypoglycemic medications and the avoidance of hospitalizations and threshold for hospitalization risk. The data set included 33,130 unique Medicaid enrolled patients with type 2 diabetes with 10% of the set used for validation and the rest for training. Features of the dataset included measures of service use, diabetes treatment intensity, sociodemographic, and health status. The adherence threshold of 80% is widely used. Study results conclude that the thresholds that best signify the risk of hospitalization were not a stable 80%, instead varying between 46% to 94%.

Longfei Han et al. (2015) proposed a hybrid system for diabetes prognosis that provided a 93.9% accuracy on a nutritional survey. It has an option for a second opinion using SVM and Random forest rule extraction learning. The dataset provided by China Health and Nutrition Survey contained 79,13,646 diabetics. 90% of the set was used for training, and the rest was used for validation using 10-fold cross-validation. The data model was narrowed to 15 features picked using univariate Linear Regression, χ^2 tests, information gain-based method, and Random Forest. Study outcomes for the positive case were Precision, Recall, F-Score: 89.6%, 44.3%, 0.593% respectively. Overall, Weighted Precision is found to be 94.2% and Weighted average recall is 93.9%.

D. *Other Technique*

Mercedes Rigla et al. (2018) developed a decision support system for Gestational Diabetes Management (GDM) within the framework program MobiGuide project. The project has access to EMR data as well as current information from glucose, blood pressure, and activity sensors to aid in decision making. The learning model used was the Mobile telemedicine system. GMD examined 20 patents using a dataset which contains 4561 glucose measurements, 369 BP values, 184 physical activities, and 997 ketonuria values. Patients using the system had reduced blood pressure.

Kathleen E. Corey et al (2016) design a classification algorithm for Nonalcoholic fatty liver disease (NAFLD) for the development of a large-scale longitudinal cohort. Linear Regression with adaptive LASSO was used on a data set that contained EMRs from 620 patients. The application performance resulted in proving that this classification algorithm is better than the ICD-9 billing data. The simplicity of this approach allows it to be applied across different institutions and a cohort of patients with NAFLD can be created based on Electronic Medical Records. This simple approach can be pursued across different organizations to create Electronic Medical Record-based groups of individuals with Nonalcoholic Fatty Liver Disease. Study outcomes included Specificity of 91% and Sensitivity of 51%.

Shankaracharya et al. (2012) developed a tool for diagnosing prediabetes and type 2 diabetes using ML techniques. A mixture of Experts (ME) was developed which subdivides the learning task into subtasks, each of which is solved by a simple expert network and their output combined to provide the global output. Data from over 1500 patients from a hospital in Delhi and the local population of Ranchi, India was processed into a data set of 1415 samples (947 diabetics). The ME model was implemented and trained in MatLab. Best results achieved were Sensitivity, Specificity, Accuracy of 99.5%, 99.07%, 99.36% respectively. The proposed tool's accuracy makes it a good candidate for large-scale screening for prediabetes.

V. Review of Machine Learning model and database used in DM

TABEL I Deep Learning models

S.No	Author/Year	Learning model	Database	Study outcomes
1	Gadekallu et al.(2020)[4]	Deep neural network	PIDD	96.00% Accuracy
2	Singh and Gorantla (2020)[5]	HE-CNN	IDRiD dataset	96.12% Accuracy
3	Ghani et al., 2019[6]	FFNN	HRF image database	100% Accuracy
4	Varun Gulshan et al. (2016)[8]	ANN, LR	EyePACS-1 data set Messidor-2 data set	EyePACS-1 sensitivity, specificity: 97.5% and 93.4% Messidor-2 sensitivity, specificity :96.1% and 93.9%
5	Carson Lam et al. (2018)[19]	Deep Learning	Kaggle retinopathy data subset, n = 243 Training subset include 1050 patches Validation data 274 patches	Weighted average precision and recall: 93.9% , 94.2% respectively
6	Hsiao-Hsien Rau et al. (2016)[20]	ANN and Logistic	data from 2060 NHIRD of Taiwan Training data: 1442	Sensitivity: 0.757 Specificity: 0.755 AUC: 0.873

TABEL II K-Nearest Neighbor models

S.No	Author/Year	Learning model	Database	Study outcomes
1	Ali et al. (2020)[10]	Fine KNN	a data set delivered from Diabetes 130-US hospitals	99.9% Accuracy
2	Aminah and Saputro (2019)[11]	KNN	Iris dataset from 15 diabetic and 11 diabetic persons	85.6% Accuracy
3	Carter et al. (2019)[12]	KNN	Self-created	AUC of 0.90
4	Samant and Agarwal (2019)[13]	subspace KNN (SKNN)	Self-created dataset of 338 (180 diabetics)	85% precision

TABEL III Other Technique

S.No	Author/Year	Learning model	Database	Study outcomes
1	Mercedes Rigla et al. (2018)[7]	Mobile telemedicine system	20 patients	Accuracy 98%,
2	Kathleen E. Corey et al.[9]	Linear regression with lasso	EMRs from 620 patients picked randomly from the high-risk patients in PHG Trust	Specificity; 91% Sensitivity: 51%
3	Shankaracharya et al.(2012) [21]	Mixture of experts	health profiles of patients from hospitals in India Data set: 1415 subjects, 947 diabetic Training data used 1104	Sensitivity,Specificity, Accuracy 99.5%,99.07%, 99.36%respectively

TABEL IV Random forest models

S.No	Author/Year	Learning model	Database	Study outcomes
1	Datta et al.(2019)[14]	Random forest	Self-created diabetic data from a mobile research center in Germany	AOC 0.90 for the ensemble classifier
2	Xu and Wang (2019)[15]	Random forest-based (RFWFS)	PIDD	93.75% Accuracy
3	Beatriz López et al. (2017)[16]	Random forest	Data of 677 subjects out of which 248 diabetic	AUC: 0.89
4	Wei-Hsuan Lo-Ciganic et al. (2015)[17]	Random survival forests	Data set of 33,130 patents with type 2 diabetes	thresholds Level hospitalization ranged from 46% to 94%
5	Longfei Han et al. (2015)[18]	Random survival forests,	China Survey data total 7913, 646 diabetic	Weighted average precision: 94.2%

Conclusion

This paper focuses on automatic diabetes detection and diagnostic techniques. There have been various studies and aspects such as databases. machine learning, Diagnostic methods were analyzed. In this paper, the studies that used the deep learning model using CNN provide better classification outcomes while measuring accuracy, specificity, AUC. Artificial intelligence will be very helpful in the detection and diagnosis of diabetes and will improve exponentially with new studies.

References

[1] International Diabetes Federation (IDF). IDF diabetes atlas, 7th edition. Brussels, Belgium: International Diabetes Federation, 2015.

[2] International Diabetes Federation. IDF Diabetes Atlas, 9th edn. Brussels, Belgium: 2019. Available at: https://www.diabetesatlas.org

[3] Singla R, Singla A, Gupta Y, Kalra S. Artificial Intelligence/Machine Learning in Diabetes Care. Indian J Endocrinol Metab. 2019;23(4):495-497. doi:10.4103/ijem.IJEM_228_19

[4] Gadekallu, T.R.; Khare, N.; Bhattacharya, S.; Singh, S.; Maddikunta, P.K.R.; Ra, I.-H.; Alazab, M. Early Detection of Diabetic Retinopathy Using PCA-Firefly Based Deep Learning Model. Electronics 2020, 9, 274. https://doi.org/10.3390/electronics9020274

[5] Singh, R. K., Gorantla, R., 2020. DMENet: Diabetic Macular Edema diagnosis using Hierarchical Ensemble of CNNs. Plos one. 15(2), e0220677.

[6] Ghani, A., See, C. H., Sudhakaran, V., Ahmad, J., Abd-Alhameed, R., 2019. Accelerating Retinal Fundus Image Classification Using Artificial Neural Networks (ANNs) and Reconfigurable Hardware (FPGA). Electronics. 8(12), 1522

[7] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;

[8] Corey KE, Kartoun U, Zheng H, Shaw SY. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. Dig Dis Sci 2016

[9] Rigla M, Martı´nez-Sarriegi I, Garcı´a-Sa´ez G, Pons B, Hernando ME. Gestational diabetes management using smart mobile telemedicine. J Diabetes Sci Technol 2018

[10] Ali, A., Alrubei, M. A., Hassan, L. F. M., Al-Ja'afari, M. A., Abdulwahed, S. H., 2020. diabetes classification based on KNN. IIUM Engineering

[11] Aminah, R., Saputro, A. H., 2019. Application of Machine Learning Techniques for Diagnosis of Diabetes Based on Iridology. In IEEE 2019 International Conference on Advanced Computer Science and Information Systems (ICACSIS). 133-138.

[12] Carter, J. A., Long, C. S., Smith, B. P., Smith, T. L., Donati, G. L., 2019. Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes. Expert Systems with Applications. 115, 245-255.

[13] Samant, P., Agarwal, R., 2018. Machine learning techniques for medical diagnosis of diabetes using iris images. Computer methods and programs in biomedicine. 157, 121-128.

[14] Datta, S., Schraplau, A., da Cruz, H. F., Sachs, J. P., Mayer, F., Böttinger, E., 2019. A Machine Learning Approach for Noninvasive Diagnosis of Metabolic Syndrome. In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE). 933-940.

[15] Xu, Z., Wang, Z., 2019. A Risk Prediction Model for Type 2 Diabetes Based on Weighted Feature Selection of Random Forest and XGBoost Ensemble Classifier. In 2019 IEEE Eleventh International Conference on Advanced Computational Intelligence (ICACI). 278-283.

[16] Lo´pez B, Torrent-Fontbona F, Vin˜as R, Ferna´ndez-Real JM. Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction. Artif Intell Med 2018;

[17] Lo-Ciganic WH, Donohue JM, Thorpe JM, et al. Using machine learning to examine medication adherence thresholds and risk of hospitalization.Med Care 2015;

[18] Han L, Luo S, Yu J, Pan L, Chen S. Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. IEEE J Biomed Health Inform 2015;

[19] Lam C, Yu C, Huang L, Rubin D. Retinal lesion detection with deep learning using image patches. Invest Ophthalmol Vis Sci 2018

[20] Rau H-H, Hsu C-Y, Lin Y-A, et al. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Comput Methods Programs Biomed 2016

[21] Shankaracharya, Odedra D, Samanta S, Vidyarthi AS. Computational intelligence-based diagnosis tool for the detection of prediabetes and type 2 diabetes in India. Rev Diabet Stud 2012

[22] Sumit Saha " Convolutional Neural Networks " link Dec 2020

[23] S., G., Gopi, V. P., & Palanisamy, P. (2020). A lightweight CNN for Diabetic Retinopathy classification from fundus images. Biomedical Signal Processing and Control, 62, 102115. doi:10.1016/j.bspc.2020.102115

[24] Ian H.; Frank, Eibe; Hall, Mark A.; Pal, Christopher J. (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco (CA)

A STUDY ON SOFT COMPUTING APPROACHES FOR IMAGE SEGMENTATION

Ramanjot Kaur [#1], Baljit Singh [*2]
[#]*Computer Science & Engineering, I.K.Gujral Punjab Technical University, Jalandhar Kapurthala, India*
[*]*Computer Science & Engineering, Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, India*
[1]jot_klair@yahoo.co.in
[2]baljitkhera740@gmail.com

ABSTRACT— Image modality types such as X-ray, CT (Computer Tomography) scan MRI (Magnetic Resonance Imaging), PET (Positron emission tomography), etc. are used for the diagnosis process. The main problem during the diagnostic process is the accurate detection of a particular disease and image segmentation provides the exact solution to solve this problem. This survey paper addresses a brief introduction to different methods of image segmentation and soft computing approaches like an artificial neural network, Genetic algorithm, fuzzy logic and nature-inspired algorithm. The various techniques and applications of soft computing that are used for the segmentation of medical images and other images such as satellite images, pant disease, standard test images, etc. are discussed. It has been concluded that soft computing-based algorithms give results with an accuracy rate of 92% to 99% (classification) and a precision of 93% for image segmentation. Algorithms such as Particle Swarm Optimization (PSO) and novel black widow algorithms of soft computing approach to multilevel thresholding also take less time to segment an image. Soft computing approaches for grayscale and color image segmentation are very effective and robust. This paper will provide the context to researchers in the field of image segmentation.

KEYWORDS— Image segmentation, filter, soft computing, artificial neural network, genetic algorithm, fuzzy logic.

INTRODUCTION

Image Segmentation methods

Image analysis is a crucial part of image processing. Types of images used for analysis are Ultrasound, Computer Tomography (CT) scan, PET (Positron emission tomography) and Magnetic Resonance Imaging (MRI) images, etc. [1]. Image segmentation is used to partition an image into a number of parts so that each and every part gives a proper view [2]. Before image segmentation, there is a need for some preprocessing means to enhance an image using filters.

1) Thresholding: Thresholding is the most desired technique used for segmentation. The thresholding method converts the whole image into a binary image based upon one numeric value which is also known as the threshold value. Let us suppose W(x, y) is an image having light objects on a dark background. The value that segments the whole image into two dominant modes object and background is called as threshold value denoted by T. Any point (x, y) satisfies the condition W(x, y)>T, is belonged to the object point otherwise, that point belonged to the background[3]. Thresholding is of two types one is a global threshold and the second is the local threshold. The quality of segmentation depends on the value of the selected thresholds.

2) Edge detection: *The second method used for segmentation is edge-based. The segmentation is done by detecting the edge of any object. This method is further divided into two classifications one gradient and the second is a gray histogram-based method [4]. Several edge detection operators are used. Some of them are canny, Sobel, Roberts, Prewitt, etc***.**

3) K means Clustering: *Clusters mean making groups of one or more elements having the same features. In K mean clustering algorithm, the first step is to decide the number of clusters i.e. K. The next step is to calculate the distance between various points and k clusters. Those are nearer to any k clusters added to that k cluster. After that mean is calculated for each k cluster so that reallocates the center of each cluster. The distance is again calculated between points and centers for making homogeneous clusters. This process continues till no change[5].*

4) Fuzzy C means Clustering: The fuzzy c mean algorithm is used for clustering [6]. In Fuzzy C-Means (FCM) clustering one part has belonged to one or more clusters. It is mostly recommended for pattern recognition [7]. This procedure groups homogeneous portions into clusters in an unsupervised manner.
5) Region Growing: Region growing is based on some predefined criteria in which grouping of homogeneous pixels into larger regions. The main method is that to select the seed point then grow according to some predefined criteria based upon color or intensity amends pixels similar to the seed to make larger regions [3].
6) Region splitting and merging: The Second method based upon region growing is to split the region into a number of disjoint regions. After that merge or split again disjoint regions to satisfy the predefined criteria which mean grouping based upon the color or intensity. Fig. 1 shows the region splitting process. This technique is developed for the segmentation process of grayscale and color images [8].

<table>
<tr><td>S1</td><td colspan="2">S2</td></tr>
<tr><td rowspan="2">S3</td><td>S41</td><td>S42</td></tr>
<tr><td>S43</td><td>S44</td></tr>
</table>

Fig.1 Region Splitting

Soft Computing Approaches

Soft computing basically deals with an approximation, uncertainty, partial truth and imprecision. It tolerates all conditions. Soft computing approaches are like artificial neural networks, genetic algorithms, fuzzy logic, nature inspired approaches[9]. Many applications are there for image segmentation purposes. A short description about main approaches of soft computing are discussed here.

1) Artificial Neural Networks (ANN): *This approach is basically artificially trained a network. It mimics the human brain process means neurons are interconnected with each other with the help of networks. There is one, two, or more layers namely the input, hidden and output layers in each network that are interlinked with weighted connections. Each element that is processed get value from the input layer and after performing some calculation it gives the output. The learning procedure of the neural network decides the value-connected weights so that correct output is produced. In an artificial neural network, a network is trained in a supervised manner to get the desired results. Fig.2 describes the basic neural network process. ANN having several training algorithms like perceptron, feedforward, backpropagation, convolutional neural network, etc.*

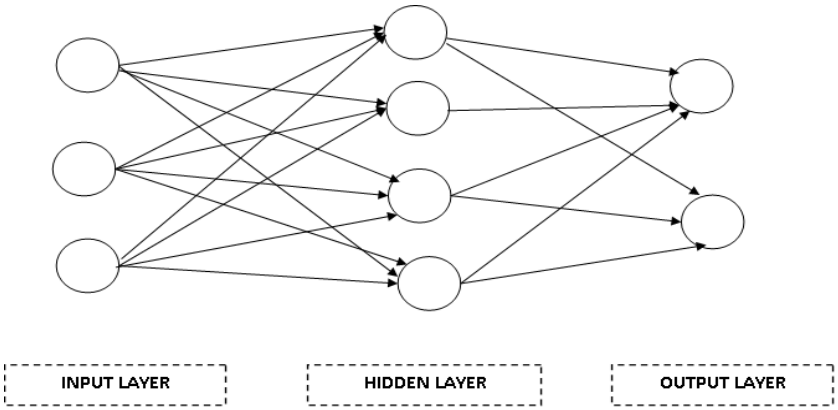


Fig.2 Artificial neural network

2) Genetic Algorithm (GA): *The Genetic algorithm is another soft computing method that is used for image segmentation. The image analysis process based upon genetic algorithms is increased day by day. The genetic algorithm is based upon the search method which mimics the natural biological evolution. In this algorithm initial population is created at random. After that selection of individuals is based upon the at their fittest survival rate[10]. The various methods of selection of fittest value like tournament selection, roulette wheel method, SUS (stochastic universal Sampling) method, ranking method, and random method, etc. The Best fitness values are selected for reproduction. The crossover and mutation process is applied for getting better results. Fig.3 explains the complete process of Genetic algorithm.*

3) Fuzzy logic based approach: *In this soft computing method, fuzzy logic means handling uncertainty. In fuzzy logic, the main steps are a selection of membership function, fuzzification and defuzzification. A fuzzy inference system is used to implement the fuzzy logic concept. The range of values used in fuzzy logic is between 0 and 1. The data point having some membership value belongs to more than one cluster[11][12].*

4) Deep learning: *Deep learning is like a machine learning approach. The basic idea behind Deep learning concept is an artificial neural network with high level of abstraction processing layer. Due to this behavior learning ability increases [13].*

5) Nature Inspired Metaheuristic: *Nature-inspired mean based upon the behavior of different nature birds, animals that how they can collect their food as an individual or in groups. Some of the algorithms are based on human behavior. Here the mentioned category of algorithms is called a metaheuristic algorithm mostly used for optimization purposes. From these behaviors, numerous algorithms are developed in recent years for the extraction of useful features from images. Few of them are like PSO (Particle Swarm intelligence)[14], ant colony optimization, Crow search algorithm[15], Bat algorithm, Island bat algorithm[16], Cuckoo search algorithm[17], Krill herd algorithm[18], Ant lion optimization, whale optimization algorithm, Teaching learning-based optimization algorithm, and many more. The unsupervised learning method is applied for the segmentation process. The PSO algorithm is basically a group of particles that fly in the sky for the search process. Each particle tries to find out the best solution. After that, all particles follow that route to get the optimal result.*

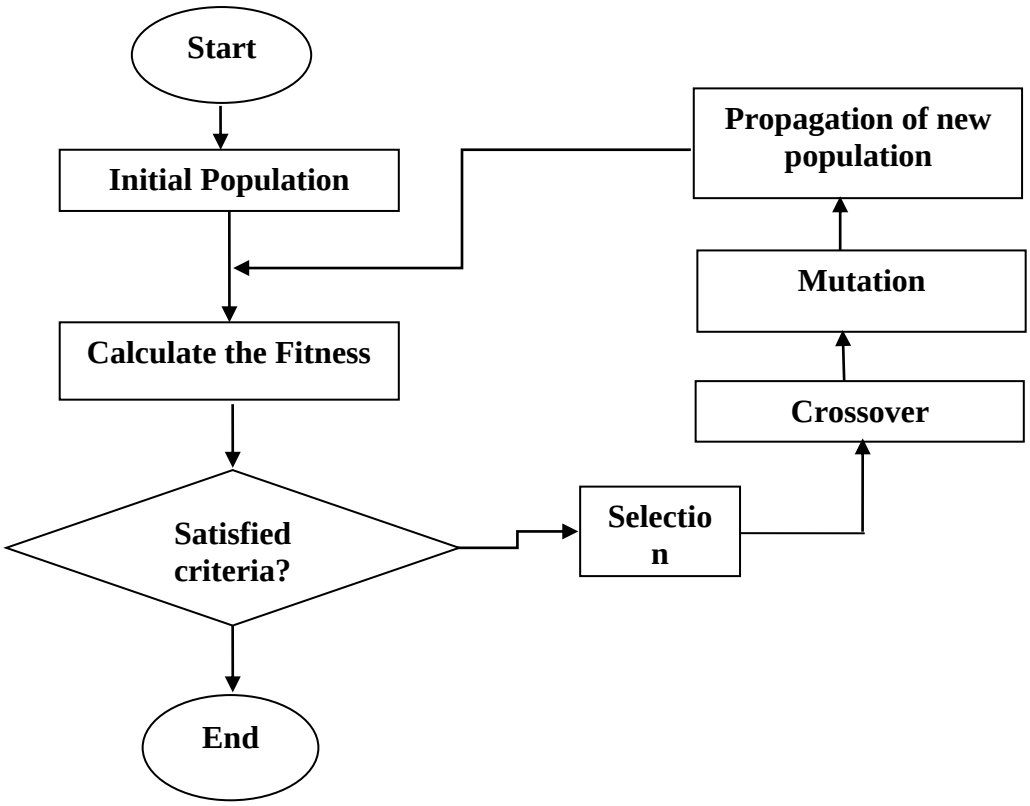


Fig.3 Flow chart for Genetic Algorithm

PREPROCESSING

In the case of preprocessing, noise removal is the main function. To improve the vision of an image filters are used. Filtering is considered an important task of image processing [19]. Various types of filters are used like mean, median, Gaussian, and many more. In preprocessing, an adaptive mean filter gives better performance after comparisons with different filters like median, adaptive median and average filters. Lung cancer CT scan images are segmented using an evolutionary algorithm with 95.80% accuracy and 90% precision [20]. The impulse noise is removed by an adaptive median filter with higher efficiency. The adaptive median filter clears the noise with low intensity and also the extreme noise with 90% to 99 % intensity [21]. The median filter has the ability to remove the noise while keeping edges. The application of the median filter is to remove the noise named salt and pepper [22]. The segmentation is done by numerous methods.

IMAGE SEGMENTATION TECHNIQUES

A. Thresholding

Mousavirad et al. describe the multilevel threshold based human mental search algorithm that consumes less time to segment the images as the level of thresholds increases [23]. The thresholding based upon the color index method proposed by Castillo-Martínez et al. for plant image segmentation, in which the foreground and background part is identified. This method does not require complex calculations. It concludes the segmentation error 6.62±5.85 % and classification ratio 1.93± 0.05% [24]. In multilevel thresholding, Otsu's method plays a significant role. Srikanth et al. investigated an approach that recovers the problem faced during histogram based approach. The rate of peak signal to noise ratio was 96.86% and also consumes less time to select a number of threshold levels used to segment an image [25].

B. Edge detection

A novel edge-based segmentation approach was proposed by Padmapriya et al. for identifying the thickness of the bladder wall [4]. Fig. 4(a) shows the original image of Lena. Lena image is a standard test image that is generally used for segmentation purposes. The segmentation result obtained by an edge-based method is shown in Fig. 4(b).

Fig.4 (a) Original image of Lena (standard test image)

Fig.4 (b) Edge based Segmentation(source: zafari(2014) p.12) [11]

C. K means clustering

K means clustering gives good results for the segmentation of color images. In terms of reducing pixel uncertainty, K means the algorithm was used with the help of neutrosophic logic. In the case of indeterminacy minimization of the pixels, K means algorithm used with the help of neutrosophic logic. The proposed method gives good results [27]. Nithya et al. proposed a new method for the detection of kidney stone using K means clustering and an Artificial neural network [28]. Dynamic particle swarm optimization and k mean clustering-based algorithm was proposed for getting good results [5].

D. Fuzzy c means clustering

Mishro et al. proposed FCM based approach to segment the MR brain tissue. Membership value calculated with the type II fuzzy approach and provides an accurate result as compared to the standard clustering technique. Images with lesions taken in the future for segmentation purposes [29]. Lie et al. introduced the concept of morphological reconstruction for the segmentation of real images. This proposed approach does not require any parameter value set. The objective function was used in terms of minimization or maximization of a problem. The vigorous patch and fuzzy logic based method imposed for concluding similar portions in an image. Two main features of Fuzzy c-means (FCM) were its effectiveness and simplicity [18-19]. As compared to hard c mean clustering, FCM produces better quality information in terms of segmentation[18].To overwhelm the drawback of over-segmentation fuzzy c means clustering based algorithm developed by Jia et al. [31].

E. Region based

Kaushik et al. discussed the advantages and disadvantages of different methods of segmentation. Also, Image segmentation using genetic algorithm done by various factors like texture, homogeneity, image continuity and content of image [32]. Evolutionary maximum entropy based algorithm proposed by Merzougui et al. This helps to estimate the starting point (seeds), for the segmentation process. The minimum variance of Red, Green, Blue components are calculated so that the seed region initial stage is recognized. The 4-connected processes are used to grow the seed region until no change [33]. Lie et al. imposed the Watershed transforms and morphological reconstruction for seeded segmentation of an image. The main purpose of this algorithm was to decrease the problem of over-segmentation. This proposed approach compared with other state of the algorithms. Also, filtration of useless regional minima done by this method with less computation [34].

SOFT COMPUTING APPLICATIONS

A. Artificial neural network (ANN)

Han et al. implemented a new method for segmentation using backpropagation neural networks with best recognition rate i.e. 99%. A concept cat chaotic mapping included in addition to combination of BP network and Gravitational search to avoid the local minima [35]. Numerous methods were proposed for the segmentation process using neural networks. For color image segmentation, the canny operator and pulse coupled neural network-based algorithm developed by Jiang et al. 0.93 best Precision value and 0.83 F-measure was achieved through this approach. For multichannel images parameters were simplified in this approach [36]. A new method proposed by Lang et al., to extract the brain tumor from multimodal images using convolution neural network with high accuracy and less time consumption. This method provide good dice ratio and also includes the concept of De noise which helps the recognition of brain tumor voxels [37]. Arumugadevi et al. implemented a neuro-fuzzy approach that gives good results for segmentation and also minimizes the time taken to train the network. The main aim of the proposed approach was to overcome the problem of the selection of number of clusters and validation of a measure of clusters. The co-occurrence based approach was applied to solve the said problem. The neuro-fuzzy approach provides a 99.3 % accuracy rate [38]. Millions of deaths occurred due to lung cancer. Deep learning-based U-net architecture proposed by Skourt et al. for accurate lung cancer detection from CT images with an 0.9502 dice-coefficient index [13]. Single or multiple brain tumors from MRI images are (95% accuracy rate) clearly segmented with the help of fuzzy entropy and Convolution neural networks. This method was developed by Sert et al. [39]. The double-layer pulse-coupled network gives highly effective results of segmentation. He et al. deployed this method to handle the difference in the color of kiwi fruit and their background during sunny and cloudy day clicked picture. It achieved a 4.75% micro classification rate of kiwifruit images [40]. Liu et al., by using neural network methodology an innovative approach developed for accurate results of segmentation of heart (ventricles 93.14%, septum 92.58%, Apex

96.21 % accuracy). At initial step region of interest (ROI) included. The proposed approach deals the situations where the contrast of images not clearly defined and suppress noise [41].

B. Genetic Algorithm (GA)

Sheta et al. proposed the edge detection and the Genetic algorithm based algorithm used for efficient image segmentation on natural images. Genetic algorithm segmented the image without the threshold values. Different type of edge detection operators like Prewit, Sobel and Robert used in proposed approach which helps to provide more accurate results of the segmentation [42]. Singh et al., to recognize the plant disease at initial stage, the segmentation process is done with the help of a genetic algorithm. In this various diseases of plants are surveyed. Almost ten species of tested with the help of proposed approach. The accuracy in detection is 93.63% and improved the SVM classification rate by 95.71% [43]. The modified approach of genetic algorithm and FCM was developed by Satapathy et al. This approach gives optimal results for MRI segmentation by improving the initialization of population and crossover process. Local and global search abilities combined in this approach to get accurate results with an optimum number of clusters [44]. To optimize the result and extracting features from the brain image Bahadure et al. proposed a genetic algorithm based approach which includes various methods like Berkley watershed transformation(BWT), Discrete Cosine transformation(DCT) and Fuzzy c means clustering(FCM).This algorithm results in 92.03% accuracy , Average dice coefficient index of 93.79% and about 0.82 to 0.93 segmentation of tumor. This method is helpful in the diagnosis process [45]. An optimized method proposed by Khan et al. for the classification of apple disease. In this optimized method four step processes involved like preprocessing, second was spot segmentation, the third extraction of feature and last was classification. Scab, Blackrot and Rust named disease classes of apple were detected using an optimized method based on a genetic algorithm. In this the multiclass support vector machine (MSVM) classifier achieved a 94.8% accuracy rate [46]. Genetic algorithm is mostly used in recent researches. Abdel-khalek et al. investigated the performance of Tsallis entropy based GA and Renyi entropy based GA measured by the peak signal to noise ratio. The function for fitness of Tsallis entropy is more complex in the case of optimization than Renyi's entropy. By using Renyi's entropy, the genetic algorithm gives good results for two dimensional image segmentation and the PSNR value is between 2.64 DB to 9.11 DB [47].

C. Fuzzy logic

Fuzzy c means clustering gives good segmentation results as compared to hard clusters and it was also suitable for the overlapped datasets. As compared to the k means clustering method proposed by Chakraborty et al. fuzzy c means method provides high performance. To improve the efficiency of biomedical image segmentation an extension and modified model of the EMO (Electromagnetism optimization) method was developed. The type II fuzzy C-Means algorithm was combined with this algorithm to increase the accuracy. The drawback of the proposed approach was not efficient to give the optimum number of clusters and also it does not handle multi-objectives [48]. To find out the breast cancer symptoms i.e white spots in X-ray mammograms, FCM with morphological top hat algorithm was applied by Bhattacharya et al. The intensity-based accurate detection of one or more micro classifications was done with the help of fuzzy c mean clustering algorithm. This method provides better segmentation results as compared to the conventional thresholding method [49]. Ren et al. proposed an unsupervised and robust algorithm. The kernel-based and weighted fuzzy c means clustering discards the unimportant information to enhance the quality of segmentation of brain MRI images. The misclassification rate of proposed approach (6.67 %) lower than the kernel based approach (9.03%) [50]. Radha et al. give accurate segmentation of brain tissues done with the help of the fuzzy level set intelligent and improved quantum based PSO method. This improved method gives a 15 percent more accurate result than the original method. In this, the approach improved segmentation results by contour initialization enhancement [51]. In an era of Geographical and remote sensing, color segmentation has an efficient part. Classification of color is needed by many color vision systems. Three color space components HSV used in the human perception based approach developed by Kyi et al. The Takagi-Sugeno fuzzy method implemented to segment a color image. The classification was based on pixel color. The computation time increases with the increase in the size of the image [52]. The combination of Tsallis entropy, fuzzy adaptive and gravitational search based method proposed by Tan et al. to provide segmentation results in less amount of time with optimal multilevel thresholds. Various types of noise like Gaussian noise and salt and pepper noise were applied on the standard test images for segmentation [11]. A new approach was developed by Singh et al. to monitor the large farms. Different disease classification methods were surveyed in this paper. Also, the sunflower disease detection was extracted with a 98% accuracy rate using the PSO approach as compared to other state of art methods [53]. The fuzzy and swarm intelligence approach proposed by Anter et al., gives accurate results to classify hepatic liver cancer lesions from CT images of the liver. The fuzzy and swarm intelligence approach provides an optimal solution with a high convergence speed in less than 50 iterations. The jaccard index and dice coefficient having higher accuracy rate in the classification of hepatic liver cancer lesions [54]. A median filter was applied for preprocessing to increase the contrast level of the original image of the liver. The improved approach based upon PSO and Fast fuzzy c means give accurate results of segmentation of the liver tumor. The algorithm proposed by Anter et al. consumes less time [55]. Khera et al. proposed the fuzzy entropy and teaching learning-based optimization algorithm for the segmentation of standard test images. The proposed approach computes a better threshold value which segments an image in a better way. This proposed approach finds out the optimal number of thresholds to segment the image [56]. To overcome the drawback of consumption of more time for more than two threshold selections a novel black widow algorithm was proposed by Houssein et al.This approach used the fitness function in terms of Ostu or Kapur's entropy to evaluate the performance. Ostu's method was good in all measures as compared to Kapur's entropy [57].

D. Metaheuristic

The wind-driven and cuckoo search-based approach segments an accurate object from satellite images. The segmentation was done by multilevel thresholding becomes expensive, to overcome this problem, approach was proposed by Bhandari et al. [58]. The whale optimization algorithm was proposed to extract a liver portion from liver MRI images. Mostafa et al. proposed an algorithm to make clusters based on the distance between two points to extract useful features. After applying the proposed approach morphological operations are applied to enhance the segmented results to achieve overall accuracy in terms of Structured similarity index (SSIM) 96.75% and Similarity index(SI) 97.5% [59]. Vijh et al., The support vector machine and Whale optimization algorithm based approach helps the doctors to diagnose the lung tumor from CT images of the Lungs. Total nineteen features are extracted using this approach. Whale optimization algorithm used to find best optimal feature. The results were obtained with a 95% accuracy rate, 100% sensitivity and 92% specificity [60]. To measure the performance of different methods, many quantitative measures are used. A List of a few of them is shown in Table 1.

TABLE I
LIST OF PARAMETERS USED TO MEASURE THE PERFORMANCE OF DIFFERENT APPROACHES

Sr. no.	Performance measure	Abbreviations	Sr. no.	Performance measure	Abbreviations
1	Dice ratio	DR	11	Jaccard Similarity	JS
2	Peak signal to Noise ratio	PSNR	12	Feature similarity Index	FSIM
3	Structured Similarity Index	SSIM	13	Uniformity	U
4	Similarity index	SI	14	Universal quality index	UQI
5	Mean Square error	MSE	15	Normalized Root Mean Square Error	NRMSE
6	Jaccard Index	JI	16	Dice Similarity Coefficient	DSC
7	Dunn index	DI	17	Accuracy	-
8	Xie- Beni Index	-	18	Sensitivity	-
9	Davies Bouldin Index	-	19	Specificity	-
10	B-index	-			

As per the above discussion, several research papers have been implemented for medical image segmentation based on soft computing approaches. Numerous diseases like brain tumors, breast cancer, liver tumors, Kidney, COVID-19, etc have been discussed. Some more soft computing approaches are given below in Table 2 for medical image segmentation.

TABLE II
DIFFERENT APPROACHES FOR MEDICAL IMAGE SEGMENTATION

Sr. No.	Year	Technique	Dataset/ Type of image	Purpose	Result and Future scope
1	2015 [61]	Support Vector Machines and Meta-Heuristic Method, Genetic algorithm.	442 MRI brain images.	Feature extraction and 4 classes of Brain Tumor using multiclass classification in abnormal brain images.	96.8% accuracy rate. In future, terms named feature extraction, segmentation and classification can be combined to make a CAD system.
2	2016 [62]	Multi-scale 3D Otsu thresholding algorithm.	10 MR-T2 brain slices.	Proposed approach shows accurate segmentation in case of bi-level and multilevel thresholding. Simple fusion rule for noise reduction.	Optimal value of scale=2. In future, should design a more sophisticated fusion rule for noise reduction also find out the algorithm to reduce the time complexity.
3	2018 [63]	Fuzzy c means and artificial neural networks	75 images breast images	Breast tumor detected and classifications of them like normal, benign and malignant.	Accuracy rate: higher than 90%.
4	2019 [64]	Feedback mechanism convolution neural network	MRI and CT	Feedback neuron and feedback layer for effective feature extraction on medical image segmentation.	Accuracy rate 85.9%, Dice value=0.8579, In future, should provide different approaches for adaptive medical image segmentation.

5	2019 [65]	Firefly Algorithm, Otsu's and K means clustering	Harvard Whole Brain Atlas, fdg-PET, titc-SPECT, MRI.	Brain tumors: glioma, metastatic adenocarcinoma, metastatic bronchogenic carcinoma and sarcoma were detected.	NRMSE=0.1559, PSNR=27.9912 and SSIM=0.8541,In future, automatic determination of number of clusters and fitness function will be adjusted for segmentation purpose.
6	2020 [66]	Improved version of Whale optimization	204 MR brain (Siemens Medical Systems)	Feature extraction and feature selection method was employed for three types of brain tumors.	Accuracy rate =96.5% for wine dataset, dermatology dataset =97.1%, celevand dataset=62.5%.
7	2020 [67]	Marker-controlled watershed segmentation and Ostu's method	Singles image, MRI	To eliminate the noise median filter used. In Proposed method, pre and post processing done for accurate detection of the liver tumor	Efficient and accurate, consume less amount of time or segmentation
8	2020 [28]	Artificial neural network and multi-kernel k-means clustering, Crow search algorithm	100 images from Internet source, Ultrasound images	To overcome the stone delineation done manually, median filter was used to remove the noise, classification of normal, tumor and stone in case of Kidney.	Accuracy rate 99.61%. In future, segmentation of different types of diseases and classification with different methods.
9	2020 [68]	Unsupervised Fuzzy c mean clustering, membership entropy	Weizmann , Berkeley , Benchmark , WANG and MSRA Salient Object, Standard images	Tissue classification (whole tumor, hyper-dense, and hypo-dense region contours). To remove sensitivity for initial cluster concept membership entropy introduced.	Object boundaries are clearly identified, Not sensitive to initial clusters, as the amount of noise increases to 0.10, the segmentation performance decreases. In future, more sophisticated color model will be implemented to enhance the segmentation accuracy.
10	2020 [69]	Slime mould and Whale optimization	12 chest X-ray images	Hybrid approach based on thresholding detectect the features of COVID -19 chest segmentation.Dealt with ISP (image segmentation problem). X-ray images are not sufficient to detect infection of COVID-19 due to that machine learning is applied for exact classification of infected patients.	Higher performance values in case of PSNR, SSIM, UQI and consume less time. In future, Performance will be evaluated on number of test images from Berkeley segmentation dataset and Benchmark.
11	2020 [70]	Convolution Neural Networks, Deep learning	COVID-19(Kagggle, sirm, radiopedia),CT	5-fold cross validation implemented for Binary and multiclass level classification (normal, pneumonia and COVID-19 cases).Two type architecture used named VGG16 and Resnet50 based architecture.	At binary level accuracy in classification more than 99%, At multiclass 86.74% and 88.52. In future, Large training data should used to reduce generalization error, Preprocessing measures should be included.

CONCLUSIONS

The main aim of this paper is to study various soft computing approaches used for image segmentation. In an image processing field, the image analysis phase plays an important role. Many medical image segmentation techniques are categorized under soft computing techniques, such as a neural network, genetic algorithms, fuzzy logic, and nature-inspired algorithms. Soft computing approaches are very effective in the process of segmenting the medical images. The accurate classification of normal and defective images is done by various soft computing approaches. It has been observed that above 90% accuracy rate result for segmentation achieved in different Imaging modalities like X-ray, CT, MRI, PET etc. The soft computing based image segmentation process is helpful for doctors in the process of diagnosing various diseases in different parts of the human body such as brain, liver, lungs, kidneys, COVID-19, etc. Except medical images, soft computing approaches are very effectively segmenting the standard test images, satellite images, leaf distribution of the plant images and plant diseases. Some algorithms are not providing optimal number of clusters during the segmentation. Therefore more research work is required for selection of optimal number of clusters.

ACKNOWLEDGEMENT

Authors are thankful for the guidance and support of I.K. Gujral Punjab Technical University, Jalandhar, Kapurthala.

REFERENCES

[1] M. G. Mavilia, T. Pakala, M. Molina, and G. Y. Wu, "Differentiating Cystic Liver Lesions: A Review of Imaging Modalities, Diagnosis and Management," *J. Clin. Transl. Hepatol.*, vol. 6, no. 2, pp. 1–9, 2018, doi: 10.14218/jcth.2017.00069.

[2] S. S. Chouhan, A. Kaul, and U. P. Singh, *Soft computing approaches for image segmentation: a survey*, vol. 77, no. 21. Multimedia Tools and Applications, 2018.

[3] Rafael C. Gonzalez and Richard E.woods, "Digital Image Processing," *3rd Edition,Pearson publication*, 2008. .

[4] B. Padmapriya, T. Kesavamurthi, and H. W. Ferose, "Edge based image segmentation technique for detection and estimation of the bladder wall thickness," *Procedia Eng.*, vol. 30, no. 2011, pp. 828–835, 2012, doi: 10.1016/j.proeng.2012.01.934.

[5] W. Xiaoqiong and Y. E. Zhang, "Image segmentation algorithm based on dynamic particle swarm optimization and K-means clustering," *Int. J. Comput. Appl.*, vol. 42, no. 7, pp. 649–654, 2020, doi: 10.1080/1206212X.2018.1521090.

[6] Y. Tang, F. Ren, and W. Pedrycz, "Fuzzy C-Means clustering through SSIM and patch for image segmentation," *Appl. Soft Comput. J.*, vol. 87, p. 105928, 2020, doi: 10.1016/j.asoc.2019.105928.

[7] M. Alata, M. Molhim, and A. Ramini, "Optimizing of Fuzzy C-Means Clustering," *Int. J. Comput. Electr. Autom. Control Inf. Eng.*, vol. 2, no. 3, pp. 670–675, 2008.

[8] K. Plataniotis and M. Zervakis, "Region growing and region merging image segmentation," no. August, 1997, doi: 10.1109/ICDSP.1997.628077.

[9] K. M. Saridakis and A. J. Dentsoras, "Soft computing in engineering design - A review," *Adv. Eng. Informatics*, vol. 22, no. 2, pp. 202–221, 2008, doi: 10.1016/j.aei.2007.10.001.

[10] G. Nagarajan, R. I. Minu, B. Muthukumar, V. Vedanarayanan, and S. D. Sundarsingh, "Hybrid Genetic Algorithm for Medical Image Feature Extraction and Selection," *Procedia Comput. Sci.*, vol. 85, no. Cms, pp. 455–462, 2016, doi: 10.1016/j.procs.2016.05.192.

[11] Z. Tan and D. Zhang, "A fuzzy adaptive gravitational search algorithm for two-dimensional multilevel thresholding image segmentation," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 11, pp. 4983–4994, 2020, doi: 10.1007/s12652-020-01777-7.

[12] A. Tiwari, S. Srivastava, and M. Pant, "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019," *Pattern Recognit. Lett.*, vol. 131, pp. 244–260, 2020, doi: 10.1016/j.patrec.2019.11.020.

[13] B. Ait Skourt, A. El Hassani, and A. Majda, "Lung CT image segmentation using deep neural networks," *Procedia Comput. Sci.*, vol. 127, pp. 109–113, 2018, doi: 10.1016/j.procs.2018.01.104.

[14] L. Li and D. Li, "Fuzzy entropy image segmentation based on particle swarm optimization," *Prog. Nat. Sci.*, vol. 18, no. 9, pp. 1167–1171, 2008, doi: 10.1016/j.pnsc.2008.03.020.

[15] M. S. Turgut, O. E. Turgut, and D. T. Eliiyi, "Island-based Crow Search Algorithm for solving optimal control problems," *Appl. Soft Comput. J.*, vol. 90, p. 106170, 2020, doi: 10.1016/j.asoc.2020.106170.

[16] M. A. Al-Betar and M. A. Awadallah, "Island bat algorithm for optimization," *Expert Syst. Appl.*, vol. 107, pp. 126–145, 2018, doi: 10.1016/j.eswa.2018.04.024.

[17] X. S. Yang and S. Deb, "Cuckoo search: Recent advances and applications," *Neural Computing and Applications*, vol. 24, no. 1. Springer, pp. 169–174, Jan. 09, 2014, doi: 10.1007/s00521-013-1367-1.

[18] K. P. Baby Resma and M. S. Nair, "Multilevel thresholding for image segmentation using Krill Herd Optimization algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 33, no. 5, pp. 528–541, 2021, doi: 10.1016/j.jksuci.2018.04.007.

[19] A. Joao, A. Gambaruto, and A. Sequeira, "Anisotropic gradient-based filtering for object segmentation in medical images," *Comput. Methods Biomech. Biomed. Eng. Imaging Vis.*, vol. 8, no. 6, pp. 621–630, 2020, doi: 10.1080/21681163.2020.1776642.

[20] K. S. Kumar, K. Venkatalakshmi, K. Karthikeyan, and M. E. Fantacci, "Lung Cancer Detection Using Image Segmentation by means of Various Evolutionary Algorithms," 2019, doi: 10.1155/2019/4909846.

[21] P. A. Lyakhov, A. R. Orazaev, N. I. Chervyakov, and D. I. Kaplun, "A new method for adaptive median filtering of images," *Proc. 2019 IEEE Conf. Russ. Young Res. Electr. Electron. Eng. ElConRus 2019*, pp. 1197–1201, 2019, doi: 10.1109/EIConRus.2019.8657050.

[22] Q. Liu, Z. Liu, S. Yong, K. Jia, and N. Razmjooy, "Computer-aided breast cancer diagnosis based on image segmentation and interval analysis," *Automatika*, vol. 61, no. 3, pp. 496–506, 2020, doi: 10.1080/00051144.2020.1785784.

[23] S. J. Mousavirad and H. Ebrahimpour-Komleh, "Human mental search-based multilevel thresholding for image segmentation," *Appl. Soft Comput.*, vol. 97, no. 572086, 2020, doi: 10.1016/j.asoc.2019.04.002.

[24] M. Castillo-Martínez, F. J. Gallegos-Funes, B. E. Carvajal-Gámez, G. Urriolagoitia-Sosa, and A. J. Rosales-Silva, "Color index based thresholding method for background and foreground segmentation of plant images," *Comput. Electron. Agric.*, vol. 178, no. September, p. 105783, 2020, doi: 10.1016/j.compag.2020.105783.

[25] R. Srikanth and K. Bikshalu, "Multilevel thresholding image segmentation based on energy curve with harmony Search Algorithm," *Ain Shams Eng. J.*, no. xxxx, 2020, doi: 10.1016/j.asej.2020.09.003.

[26] S. Zafari, "Segmentation of Overlapping Convex Objects," no. August, pp. 5–6, 2014, doi: 10.13140/RG.2.1.1339.0803.

[27] M. N. Qureshi and M. V. Ahamad, "An Improved Method for Image Segmentation Using K-Means Clustering with Neutrosophic Logic," *Procedia Comput. Sci.*, vol. 132, pp. 534–540, 2018, doi: 10.1016/j.procs.2018.05.006.

[28] A. Nithya, A. Appathurai, N. Venkatadri, D. R. Ramji, and C. Anna Palagan, "Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images," *Meas. J. Int. Meas. Confed.*, vol. 149, p. 106952, 2020, doi: 10.1016/j.measurement.2019.106952.

[29] P. K. Mishro, S. Agrawal, R. Panda, and A. Abraham, "A Novel Type-2 Fuzzy C-Means Clustering for Brain MR Image Segmentation," *IEEE Trans. Cybern.*, pp. 1–12, 2020, doi: 10.1109/tcyb.2020.2994235.

[30] T. Lei, X. Jia, Y. Zhang, S. Member, L. He, and S. Member, "Significantly Fast and Robust Fuzzy C-Means Clustering Algorithm Based on Morphological Reconstruction and Membership Filtering," pp. 1–15, 2017.

[31] X. Jia, T. Lei, X. Du, S. Liu, H. Meng, and A. K. Nandi, "Robust Self-Sparse Fuzzy Clustering for Image Segmentation," *IEEE Access*, vol. 8, pp. 146182–146195, 2020, doi: 10.1109/ACCESS.2020.3015270.

[32] D. Kaushik, U. Singh, P. Singhal, and V. Singh, "Medical Image Segmentation using Genetic Algorithm," *Int. J. Comput. Appl.*, vol. 81, no. 18, pp. 10–15, 2013, doi: 10.5120/14222-2220.

[33] M. Merzougui and A. El Allaoui, "Region growing segmentation optimized by evolutionary approach and maximum entropy," *Procedia Comput. Sci.*, vol. 151, pp. 1046–1051, 2019, doi: 10.1016/j.procs.2019.04.148.

[34] T. Lei, X. Jia, T. Liu, S. Liu, H. Meng, and A. K. Nandi, "Adaptive Morphological Reconstruction for Seeded Image Segmentation," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5510–5523, 2019, doi: 10.1109/TIP.2019.2920514.

[35] X. H. Han, X. Xiong, and F. Duan, "A new method for image segmentation based on BP neural network and gravitational search algorithm enhanced by cat chaotic mapping," *Appl. Intell.*, vol. 43, no. 4, pp. 855–873, 2015, doi: 10.1007/s10489-015-0679-5.

[36] W. Jiang, H. Zhou, Y. Shen, B. Liu, and Z. Fu, "Image segmentation with pulse-coupled neural network and Canny operators," *Comput. Electr. Eng.*, vol. 46, pp. 528–538, 2015, doi: 10.1016/j.compeleceng.2015.03.028.

[37] R. Lang, L. Zhao, and K. Jia, "Brain tumor image segmentation based on convolution neural network," *Proc. - 2016 9th Int. Congr. Image Signal Process. Biomed. Eng. Informatics, CISP-BMEI 2016*, pp. 1402–1406, 2017, doi: 10.1109/CISP-BMEI.2016.7852936.

[38] S. Arumugadevi and V. Seenivasagam, "Color image segmentation using feedforward neural networks with FCM," *Int. J. Autom. Comput.*, vol. 13, no. 5, pp. 491–500, 2016, doi: 10.1007/s11633-016-0975-5.

[39] E. Sert, F. Ozyurt, and A. Dogantekin, "A new approach for brain tumor diagnosis system: Single image super resolution based maximum fuzzy entropy segmentation and convolutional neural network," *Med. Hypotheses*, vol. 133, no. September, p. 109413, 2019, doi: 10.1016/j.mehy.2019.109413.

[40] F. He, C. Fu, H. Shao, and J. Teng, "An image segmentation algorithm based on double-layer pulse-coupled neural network model for kiwifruit detection," *Comput. Electr. Eng.*, vol. 79, 2019, doi: 10.1016/j.compeleceng.2019.106466.

[41] D. Liu *et al.*, "Cardiac magnetic resonance image segmentation based on convolutional neural network," *Comput. Methods Programs Biomed.*, vol. 197, 2020, doi: 10.1016/j.cmpb.2020.105755.

[42] A. Sheta, M. S. Braik, and S. Aljahdali, "Genetic Algorithms: A tool for image segmentation," *Proc. 2012 Int. Conf. Multimed. Comput. Syst. ICMCS 2012*, no. May 2012, pp. 84–90, 2012, doi: 10.1109/ICMCS.2012.6320144.

[43] V. Singh and A. K. Misra, "Detection of plant leaf diseases using image segmentation and soft computing techniques," *Inf. Process. Agric.*, vol. 4, no. 1, pp. 41–49, 2017, doi: 10.1016/j.inpa.2016.10.005.

[44] S. C. Satapathy, V. Bhateja, S. K. Udgata, and P. K. Pattnaik, "A Modified Genetic Algorithm Based FCM Clustering Algorithm for Magnetic Resonance Image Segmentation," *Adv. Intell. Syst. Comput.*, vol. 515, pp. v–vii, 2017, doi: 10.1007/978-981-10-3153-3.

[45] N. B. Bahadure, A. K. Ray, and H. P. Thethi, "Comparative Approach of MRI-Based Brain Tumor Segmentation and Classification Using Genetic Algorithm," *J. Digit. Imaging*, vol. 31, no. 4, pp. 477–489, 2018, doi: 10.1007/s10278-018-0050-6.

[46] M. A. Khan *et al.*, "An Optimized Method for Segmentation and Classification of Apple Diseases Based on Strong Correlation and Genetic Algorithm Based Feature Selection," *IEEE Access*, vol. 7, no. c, pp. 46261–46277, 2019, doi: 10.1109/ACCESS.2019.2908040.

[47] S. Abdel-Khalek, A. Ben Ishak, O. A. Omer, and A. S. F. Obada, "A two-dimensional image segmentation method based on genetic algorithm and entropy," *Optik (Stuttg).*, vol. 131, pp. 414–422, 2017, doi: 10.1016/j.ijleo.2016.11.039.

[48] S. Chakraborty and K. Mali, "Fuzzy Electromagnetism Optimization (FEMO) and its application in biomedical image segmentation," *Appl. Soft Comput. J.*, vol. 97, p. 106800, 2020, doi: 10.1016/j.asoc.2020.106800.

[49] M. B. and A. Das, "Fuzzy Logic Based Segmentation of Microcalcification in Breast Using Digital Mammograms Considering Multiresolution," *Int. Mach. Vis. Image Process. Conf. IMVIP 2007*, pp. 169–176, 2007, doi: 10.1109/IMVIP.2007.33.

[50] T. Ren, H. Wang, H. Feng, C. Xu, G. Liu, and P. Ding, "Study on the improved fuzzy clustering algorithm and its application in brain image segmentation," *Appl. Soft Comput. J.*, vol. 81, p. 105503, 2019, doi: 10.1016/j.asoc.2019.105503.

[51] R. Radha and R. Gopalakrishnan, "A medical analytical system using intelligent fuzzy level set brain image segmentation based on improved quantum particle swarm optimization," *Microprocess. Microsyst.*, vol. 79, no. September, p. 103283, 2020, doi: 10.1016/j.micpro.2020.103283.

[52] T. M. Kyi, K. Chan, and M. Zin, *Color Segmentation Based on Human Perception Using Fuzzy Logic*. Springer Singapore, 2019.

[53] V. Singh, "Sunflower leaf diseases detection using image segmentation based on particle swarm optimization," *Artif. Intell. Agric.*, vol. 3, pp. 62–68, 2019, doi: 10.1016/j.aiia.2019.09.002.

[54] A. M. Anter, S. Bhattacharyya, and Z. Zhang, "Multi-stage fuzzy swarm intelligence for automatic hepatic lesion segmentation from CT scans," *Appl. Soft Comput. J.*, vol. 96, p. 106677, 2020, doi: 10.1016/j.asoc.2020.106677.

[55] A. M. Anter and A. E. Hassenian, "Computational intelligence optimization approach based on particle swarm optimizer and neutrosophic set for abdominal CT liver tumor segmentation," *J. Comput. Sci.*, vol. 25, pp. 376–387, 2018, doi: 10.1016/j.jocs.2018.01.003.

[56] B. S. Khehra and A. S. Pharwaha, "Image Segmentation Using Teaching-Learning-Based Optimization Algorithm and Fuzzy Entropy," in *Proceedings - 15th International Conference on Computational Science and Its Applications, ICCSA 2015*, Jul. 2015, pp. 67–71, doi: 10.1109/ICCSA.2015.10.

[57] E. H. Houssein, B. E. din Helmy, D. Oliva, A. A. Elngar, and H. Shaban, "A novel Black Widow Optimization algorithm for multilevel thresholding image segmentation," *Expert Syst. Appl.*, p. 114159, 2020, doi: 10.1016/j.eswa.2020.114159.

[58] A. K. Bhandari, V. K. Singh, A. Kumar, and G. K. Singh, "Cuckoo search algorithm and wind driven optimization based study of satellite image segmentation for multilevel thresholding using Kapur's entropy," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3538–3560, 2014, doi: 10.1016/j.eswa.2013.10.059.

[59] A. Mostafa, A. E. Hassanien, M. Houseni, and H. Hefny, "Liver segmentation in MRI images based on whale optimization algorithm," *Multimed. Tools Appl.*, vol. 76, no. 23, pp. 24931–24954, 2017, doi: 10.1007/s11042-017-4638-5.

[60] S. Vijh, D. Gaur, and S. Kumar, "An intelligent lung tumor diagnosis system using whale optimization algorithm and support vector machine," *Int. J. Syst. Assur. Eng. Manag.*, vol. 11, no. 2, pp. 374–384, 2020, doi: 10.1007/s13198-019-00866-x.

[61] A. Jayachandran and G. Kharmega Sundararaj, "Abnormality Segmentation and Classification of Multi-class Brain Tumor in MR Images Using Fuzzy Logic-Based Hybrid Kernel SVM," *Int. J. Fuzzy Syst.*, vol. 17, no. 3, pp. 434–443, 2015, doi: 10.1007/s40815-015-0064-x.

[62] Y. Feng, H. Zhao, X. Li, X. Zhang, and H. Li, "A multi-scale 3D Otsu thresholding algorithm for medical image segmentation," *Digit. Signal Process. A Rev. J.*, vol. 60, pp. 186–199, 2017, doi: 10.1016/j.dsp.2016.08.003.

[63] Y. Hamad, K. Simonov, and M. B. Naeem, "Breast cancer detection and classification using artificial neural networks," in *Proceedings - 2018 1st Annual International Conference on Information and Sciences, AiCIS 2018*, Feb. 2019, pp. 51–57, doi: 10.1109/AiCIS.2018.00022.

[64] A. Feng-Ping and L. Zhi-Wen, "Medical image segmentation algorithm based on feedback mechanism convolutional neural network," *Biomed. Signal Process. Control*, vol. 53, p. 101589, 2019, doi: 10.1016/j.bspc.2019.101589.

[65] R. Capor Hrosik, E. Tuba, E. Dolicanin, R. Jovanovic, and M. Tuba, "Brain image segmentation based on firefly algorithm combined with k-means clustering," *Stud. Informatics Control*, vol. 28, no. 2, pp. 167–176, 2019, doi: 10.24846/v28i2y201905.

[66] B. Yin, C. Wang, and F. Abza, "New brain tumor classification method based on an improved version of whale optimization algorithm," *Biomed. Signal Process. Control*, vol. 56, p. 101728, 2020, doi: 10.1016/j.bspc.2019.101728.

[67] S. Thulasidass, D. V. Soundari, S. Chinnapparaj, and R. Naveen, "Liver tumor diagnosis by using hybrid watershed segmentation method," *Mater. Today Proc.*, vol. 37, no. 2, pp. 2848–2857, 2020, doi: 10.1016/j.matpr.2020.08.660.

[68] S. K. Choy, T. C. Ng, and C. Yu, "Unsupervised fuzzy model-based image segmentation," *Signal Processing*, vol. 171, p. 107483, 2020, doi: 10.1016/j.sigpro.2020.107483.

[69] M. Abdel-basset, V. Chang, and R. Mohamed, "HSMA _ WOA : A hybrid novel Slime mould algorithm with whale optimization algorithm for tackling the image segmentation problem of chest X-ray images," *Appl. Soft Comput. J.*, vol. 95, p. 106642, 2020, doi: 10.1016/j.asoc.2020.106642.

[70] R. Lokwani, A. Gaikwad, V. Kulkarni, A. Pant, and A. Kharat, "Automated detection of COVID-19 from CT scans using convolutional neural networks," *ICPRAM 2021 - Proc. 10th Int. Conf. Pattern Recognit. Appl. Methods*, vol. 41, no. 2, pp. 565–570, 2021, doi: 10.1016/j.bbe.2021.04.006.

ANALYTICAL REVIEW OF COMMUNITY BASED INFLUENCE MAXIMIZATION MODEL

Ms. Sneha[1] Dr. Anupam Bhatia[2]
Department of Computer Science and Application, Chaudhary Ranbir Singh University
Shining.jind2yahoo.com
anupambhatia@crsu.ac.in

ABSTRACT— Influence Maximization (IN-MAX) is a major analytical issue in social influence research that chooses a group of k individuals (named seed set) from a social media platform to maximise the predicted number of impacted users (known as influence spread). We examine and synthesise a wide range of previous research on IN-MAX from an algorithmic viewpoint in this work, with a particular focus on the Algorithms Taxonomy at Community Level. Influence analysis is a critical tool for comprehending real-world events. Motivated by these facts, we present a state-of-the-art overview of influence analysis approaches for tackling these issues in this work. The major goal of this thorough study is to explore and compare research methodologies and procedures, with a particular focus on the following essential aspects: (1) an overview of diffusion models that describe the process of information dissemination and constitute the basis of the IN-MAX issue, (2) Taxonomy of existing research on IN-MAX algorithms based on underlying objectives, and (3) a sound conceptual comparison of underlying IN-MAX algorithms.

KEYWORDS: Influence Maximization, Information Diffusion, Social Networks, Community Based Algorithms, Community Structure

I. INTRODUCTION

Social media has been a popular medium for product advertising in recent years (e.g. viral marketing). Previous research has shown that viral marketing is more successful than television or print advertising. Artificial intelligence (AI) is currently playing a major role in social media marketing, thanks to the advancement of new technology. IN-MAX is the key problem behind viral marketing in social networks, which has been extensively studied recently [1][2][3]. With the rise of geo-social networks (such as Foursquare and Facebook), location-based product marketing in real-world applications is becoming increasingly important. Due of its potential economic value, Influence Maximization has recently received a lot of attention as a significant algorithmic challenge in ID research. In an online Social Network, IN-MAX tries to pick a group of k users with the largest influence spread, i.e., the predicted number of impacted users through the seed set is maximised in information diffusion. Viral marketing is a well-known application of IN-MAX, in which a corporation may seek to spread the acceptance of a new product from a small group of early adopters via social ties between users. IN-MAX is used in a variety of other significant applications, including network monitoring, rumour management, and social recommendation, in addition to viral marketing.

Despite its wide range of applications, IN-MAX poses significant research hurdles. The first problem is determining how to represent the information diffusion process in a social network, which has a significant impact on the propagation of effect of any seed planted in IN-MAX. Second, the IN-MAX problem is inherently conceptually difficult. Under most diffusion models, obtaining an optimum IN-MAX solution has been proved to be NP-hard [4]. Further, because information diffusion is unpredictable, even determining the impact diffusion of a single seed set is computationally intractable. These theoretical findings demonstrate that retrieving an optimum seed set while scaling to huge social networks is extremely difficult. Third, online social networks have lately been enhanced with new features such as location-based services, topical analysis, streaming material, and so on. This has given rise to the possibility of integrating IN-MAX with multiple circumstances, such as location, time, and subject information, in order to boost IN-MAX 's efficacy.

Li et al. [2] study the location-aware IN-MAX problem (i.e., finding k users in a geo-social network who have the most influence spread over a group of users in a specified region) and propose algorithms with 1-1/e approximation ratio and algorithms with (1-1/e) approximation ratio for any (0, 1] for online queries to meet the location-aware requirement in IN-MAX. However, they make the unrealistic assumption that each user has a known fixed position. In reality, users have diverse preferences depending on where they are. Wang et al. [3] establish the distance-aware IN-MAX issue, which considers the distance between users and the advertised location, and present a priority-based method to solve the problem with a 1-1/e approximation ratio. To derive the location aware propagation probability in LBSN, Zhu et al. [5] propose two user mobility models, namely the Gaussian based and distance-based mobility models. Zhou et. al. [1] investigate IN-MAX in an O2O environment, taking into account users' previous mobility patterns. They also offer a two-phase model, which is an enhanced influence diffusion model that incorporates both the online and offline product adoption processes.

Community detection is the technique of identifying related groups or clusters in a network. There are two sorts of communities: disjoint and overlapping. If the intersection of two communities is empty, the community is disjoint; otherwise, the community is overlapped. Figure 1 depicts a fragmented community, with dotted lines representing communities such as C1, C2, C3, and C4, whereas Figure 2 depicts an overlapping community, with two or more communities sharing nodes such as E, D, and J.

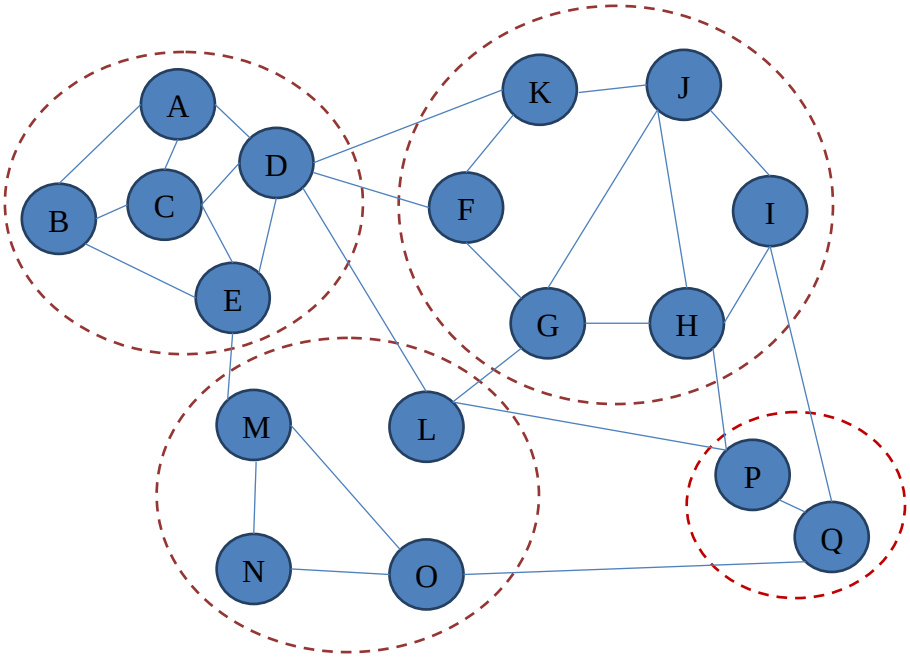


Fig 1: Disjoint Community

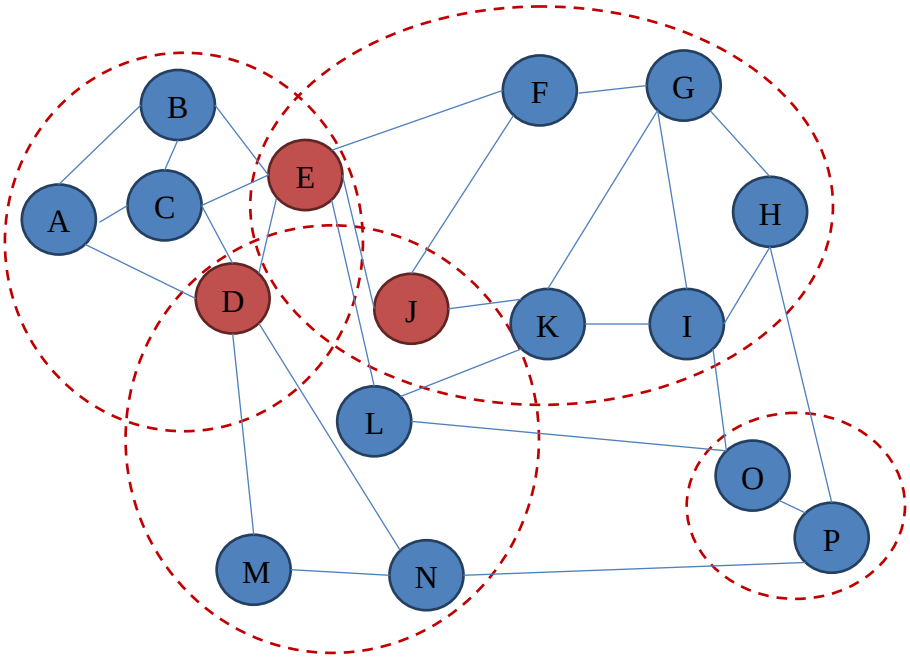


Fig 2: Overlapped Community

II. RELATED WORK

Social-Networks are social entity sets such as organisations, people, and crafts that interact or have ties with one another that may be tabulated in the form of relational databases. The social network is either directed or undirected. It is dynamic in nature and comprises of intricate connections. People have begun to combine the most popular social-networks in recent years, and these social-networks have the potential to influence community behaviour, communication, and information nature. Social networks may be quantified in the same way as graphs can, with nodes containing information about people and edges representing the connections between them. There are two sorts of connections in a social media platform: active and passive. This process continues when units get input from its neighbours and alter their status to active. Weak relationships were accountable for knowledge creation, conservation, and judgement, whereas strong ties were responsible for knowledge generation, preservation, and decision making [6]. Information is disseminated on social networks through wall postings, messages, and one-bit pokes.

The Static Greedy Algorithm was suggested by Cheng et. al. [7]. Edges were chosen based on related diffusion probability after a number of Monte Carlo simulations were done. Initially, an empty set was obtained, and the procedure was repeated until k nodes with the greatest average marginal spread in all sampled snapshots were chosen. Instead of upgrading greedy

algorithms, Chen[8] advised investing in heuristic search for influence spread since it may be a million times faster. Heuristic algorithms, on the other hand, might be far more efficient and scalable for large-scale social networks. Because of the high computing complexity of greedy algorithms, several excellent heuristic techniques have been presented to tackle this problem efficiently and with fewer iterations.

People trusted their relatives the most, according to Hossenpour et al.[9], and they were the most useful nodes for direct suggestion. The line graph has a proclivity for preserving all of a network's information, including node relationships, indirect neighbours, and direct neighbours. The major goal of the author was to reduce computational burden. Due to the line graph, overlapping vertices might also be examined.

The communication between social network users and their neighbours, whether online or offline, was sporadic. Information spreads as a result of discrete communication steps, and the frequency with which these events occur has a significant influence on communication routes, forming a dynamic communication backbone [10].

The storing of user information within the nodes is the underpinning idea of tree-based algorithms. A binary tree was deconstructed from a series parallel graph, with each node representing a subgraph's edges and the leaf node representing the subgraph's edges. In the specific situation of directed graph, the influence spread was calculated. Zhang et al. [11] presented a hybrid inverted R tree (HIR) based on R and inverted trees for tackling the multi-location influence maximisation problem and improving the offline phase search efficiency. The HIR-tree index structure was disk-resistant, and the page size was 8kb. HIR-tree concurrently computed three offline phase factors, including query speed for each user's potential consuming location, search space pruning, and time complexity of O(|V| log|M||N|) and space complexity of O|N|.

III. DIFFUSION MODELS

Scholars in fields as diverse as cloud computing, economics, advertising, and mathematics are all interested in methodological approaches for describing information spread. Many of the problems that arise in real-world applications may be modelled to better understand and address them. Predicting individual behaviour to anticipate the result of some crowd-based processes, such as polling and idea spread, has been stressed by Granovetter [12]. Various models employ various techniques to represent how a user transitions from passive to active condition, which is impacted by its surroundings. This section solely looks at typical models for the IN-MAX issue, such as the Independent Cascade (IC) model, Linear Threshold (LT) model, Path-Based model, Community-Based model among others.

Independent Cascade Model

The likelihood of a node u impacting a node v is a better way to express cascading models. The probability is denoted by the letter p. (u, v). To comprehend the process of information dissemination, the independent cascade model [4,8,10] is employed. The following is how ICM operates. Assume node u is affected (that is, turns active) at time t. Then, with p(u,v) frequency, u has a chance to affect each of its neighbours v. If u activates v, v becomes active at time t+1. If you don't, you'll never be able to influence v in the future. At the termination of the diffusion phase, no new nodes become active.

Granovetter [12] introduced the **linear threshold model (LTM)** to better explain information dispersion in crowds. Unlike the ICM, the LTM relies on the threshold parameter to propagate information. This threshold reflects a number of impediments to a user's adoption of the information. The LTM has been widely utilised in researching the spread process in social networks since its introduction.

In the LTM model, each node v has a threshold θ_v and every $u \in N(v)$ has a non-negative edge weight $w_{u,v}$, so that $\Sigma u \in N(v) wu,v \leq 1$. The spread advances in predictable discrete steps with a threshold as well as a starting set of active nodes. If node v turns active at period t, then

$$\Sigma u \in N_a(v) wu,v \leq 1 \quad (1)$$

where $N_a(v)$ denotes the set of v's already-active neighbours. Until there are no more nodes to activate when the procedure finishes.

The state of the targeted nodes is taken into account by the **Mixed Diffusion Model**. The primary set of seed nodes was used to start this model. Similar to ICM, this primary set now only had one chance to activate its neighbours. Instead of using an arbitrary propagation probability to influence their neighbours, LTM considers just a portion of the targeted node's active neighbours. The targeted nodes change their status from inactive to active when the combined weight of these nodes and their neighbours exceeds the threshold [8].

When there were many product locations, including online and offline phases, **Multi Factor Propagation (MFP)** functioned. To evaluate if a user visits a product location in the offline phase, many criteria were considered: distance, user interest, and friend evaluation [14]. The edge weight in a graph is reciprocal to the node degree in the Weighted Cascade Model (WC). Kempe et al. utilised this model as well [4].

Influence-based similarities are determined between users using the **Maximum Influence Arborescence (MIA)** model, which is based on spectral clustering. Under the MIA model, Location Aware Influence Maximization is NP-hard, and its influence spread is sub-modular and monotone [6]. There is no ambiguity in the MIA model, and nodes are activated only through the maximal influence channel with the highest activation probability. Furthermore, MIA uses a threshold to remove unimportant paths.

The shift between offline and online behaviour is represented by the **Two-Phase** model. Before deciding on customer satisfaction, customers will have an offline experience if they are affected online through other customers. Online and offline diffusion stages exist, as well as four user states: inactive, online-active, offline-active, and closed. [15][16]

Path-based Influence Maximization (PB-IN-MAX): To address micro-level problems, it took into account the community's strong interactive connectedness. To assess the spreading of a node's impact in the community, weights were applied to routes from a node to other accessible nodes. It conducts a basic route traversal, which increases diffusion speed. Unit Community Detection and Community Merging are two steps that come with it. [11]

Community-based Influence Maximization (CB-IN-MAX): The distribution of influence in the community was re-evaluated for just those nodes from which seed was chosen. It decreased the frequency of re-evaluations by addressing macro-level problems. [11]

Hybrid Influence Maximization: Ko, Cho, Kim [13] proposed a hybrid influence maximisation diffusion model to address orthogonal challenges at the macro and micro levels. There are two steps to this process: community detection and seed selection.

IV. Existing IN-MAX Algorithms Taxonomy at Community Level

In general, the IN-MAX issue has been demonstrated to be NP-hard in the setting of several diffusion models. To get approximate answers, several works have been offered. Because a person with a large number of friends is likely to be important, a frequently used heuristic to solve the IN-MAX issue is to choose seeds based on their degree, which is known as degree centrality. Members of large communities, on the other hand, frequently have a higher degree than members of small groups. As a result of degree centrality, seeds from the same huge community may readily appear. Another often used heuristic is distance centrality, which takes seeds in order of increasing average distance to other nodes since influence spreads (i.e., the number of impacted nodes) of seeds in the same community tends to overlap. Distance centrality, on the other hand, results in seeds in the same big community since nodes in large communities generally have a short average distance. In conclusion, seed clustering occurs as a function of both degree and distance centrality, resulting in a significant decrease in influence spread.

IN-MAX is still difficult to solve, despite the fact that the aforesaid approach has a decent approximation ratio of (1-1/e), since evaluating $\sigma(\cdot)$ is a #P-hard task even under simple models. In recent years, theoretical difficulty has prompted significant study towards developing efficient IN-MAX algorithms. A community is a collection of seeds that are densely linked to one another but sparsely connected to others. The majority of real-world social networks have a community-like structure. Community Detection is a highly significant topic in Network Analysis, and it has captivated the interest of researchers from different disciplines. Community was described by Li et. al. [17] as a collection of users with similar behaviours who communicate regularly and are likely to influence one another inside the group. Ko, Cho & Kim [13] utilised the community structure feature to tackle macro-level problems. Influence Spread in a single community is comparable to that of a full social network; neighbouring communities have greater Influence Spread, which impacts community nodes further away; nodes within a community are strongly linked.

Recently, various research papers have been presented to address the problem of influence maximising utilising community data. The split of network nodes into groups, within which nodes are highly connected while they are sparsely connected, is referred to as community structure [11, 17-21]. The topic of community identification has been widely researched in recent years, and community structure generally reveals basic characteristics of networks. The fundamental concept behind community-based influence maximisation algorithms is that, because various communities are sparsely linked, we may estimate a node's impact on the whole network by using its influence inside its own community. Because the size of a community is typically significantly less than the size of the entire network, the impact of a node inside its own community can be estimated more quickly. The first community-based influence maximisation algorithm, OASNET, was proposed by Cao et al. [22]. (Optimal Allocation in a Social Network). They believe that separate groups are self-contained and that influence cannot travel between them. The CNM (Clauset–Newman–Moore) [21] method was used to detect the community structure. Two sentences appear in the seed node selection. In the first step, the method picks k nodes from each community using a standard greedy approach, yielding a total of C k candidate nodes. Using dynamic programming, the algorithm picks k nodes as the seed set S from the C k candidates in the second phase.

Zhang et al. [20] investigated the challenge of finding prominent nodes in community-structured networks. From the weighted network, the authors first created an information transmission probability matrix. The k-medoid clustering technique was then used to find the k seed (influential) nodes. They tested their method on numerous real-world networks with ground truth community structure as well as LFR [17] synthetic networks. Their approach can successfully locate the most influential nodes, even on networks with an imbalanced community structure, according to experimental data. In terms of influence scope, a new evaluation metric proposed by the authors, the method even beats the standard greedy algorithm. The efficacy of these algorithms, however, has only been tested on tiny networks. The algorithm's ability to handle large-scale networks with millions of nodes and edges remains unknown.

Chen et al. [18] used the HD (heat diffusion) model to investigate the community-based influence maximisation issue and presented the CIM (Community-based Influence Maximization) method. There are three steps to the algorithm: (i) community detection, (ii) candidate node generation, and (iii) seed node generation. To get the community structure, the authors suggested a hierarchical community detection technique called HClustering in the first step. Candidate nodes are

chosen in the second phase based on their network structure and community characteristics. In the third step, k nodes from the candidates are chosen as the final seed set S depending on various scoring measures, such as whether the node is a hub or how many communities it connects. Under the heat diffusion model, the experimental findings demonstrate that this method functions effectively.

Within a community that relied on label propagation, the IN-MAX-PLA algorithm identifies the important nodes. The method, however, has a lower influence spread than the Greedy algorithm and simply takes into account the degree of each node. Wang et. al. [23] simplified the original graph by removing edges with weights greater than a pre-set threshold. Only in terms of running duration did it properly predict the influence spread. Live edges were used to transfer influence amongst groups. In mobile social networks, they discovered a seed set of prominent nodes; nevertheless, this method was shown to be less efficient than the Community Based Seed Selection Algorithm [17].

The community structure is of extremely high quality if the impact is very near to the whole network. In terms of accuracy and performance, path-based community detection surpasses [11]. It discovered communities that were more favourable for seed selection by relying on edges rather than living edges. A node's impact spread is calculated by summing the weights of all routes in a single path traversal.

Li et al. [24] calculated the effects using MIA dispersion between usage and proposed a CSS method. The method successfully discovered seeds using offline PR-tree based indexes that precomputed user's community-based inferences and the marginal influence of individuals who would be picked as seeds with high probability online preferentially. They use the Spectral Clustering Algorithm for Directed Weighted Graph algorithms and define the social influence-based similarity metric as part of this category, which reduces the problem at the community level by employing a community detection strategy on the inherent social networking sites at the moderate level. These methods do not provide a worst-case bound on the propagation of influence. The selected approach was shown to be far more effective than alternative ways with similar spreads of impact. On the underlying social network, communities were also discovered between nodes. The majority of the algorithms were based on topological structures and sought to discover communities that did not overlap (Modularity Maximization, Random Walk based method, Spectral Clustering). However, some research focused on groups that overlapped (Bai algorithm, Ma et. al. algorithm).

Algorithm	**Authors**	**Paper Titled**	**Complexity**	**Benchmark**	**Findings**
OASNET	Cao T. et. al. *[20]*	OASNET: An Optimal Allocation Approach to Influence Maximization in Modular Social Networks	O(kms)	Non-Overlapped Clustering	It is assumed that the network was static and does not change over time. It was considered that social network groups were disjointed
Hierarchical Agglomeration Algorithm	Clauset A. et. al. *[19]*	Finding Community Structure in very large Network	O (md log n)	Modularity	It has the potential to extend community structure research to networks that were previously thought to be too vast to be tractable.
SNMF-SS	Ma X. et. al. *[23]*	Semi-Supervised Clustering Algorithm for Community Structure Detection in Complex Networks	O ($\|V\|^2m^4\Gamma$)	SS Clustering	It is also critical for community identification to make the best use of applicable measurements in order to grab as much network structure information as feasible.
OCDDP	Bai X. et. al. *[27]*	An Overlapping Community Detection Algorithm based on Density Peaks	O (αn $[m+(1+3n/2\alpha)]$ + $n^2\log_2 n$	Overlapped Clustering	Core was a more constrained way of selection. The membership vectors of all other nodes are used to assign them to distinct communities. It was tested on both real and artificial networks.

CINEMA	Li H. et. al. *[26]*	Conformity-aware Influence Maximization in Online Social Networks	O (k'm'n' + kTRm')	Non-Overlapped Clustering	CINEMA isn't bound to any particular partitioning or conformance computing approach. This improves CINEMA's universality and portability. It may be implemented on a decentralised network.
SVDCNMF and SVDCSNMF	Lu H. et. al. *[24]*	Community Detection Algorithm based on Non-Negative Matrix Factorization and Pairwise Constraints	O ($n^3 + s^2 + n^2$)	Hierarchical Clustering	To determine the number of communities from a single run, singular value decomposition was used. The suggested technique applied logical inference-based paired restrictions to the NMF model.
k-Medoid Clustering Algorithm	Zhang X. et. al. *[18]*	Identifying Influential Nodes in Complex Networks with Community Structure	O ($N_{Gr} + L_{Gr)}$	Partitioning Clustering	Reveals a new Influence scope maximization-based metric for evaluating the influence impact, which complements the existing measure of the predicted number of influenced nodes.
CoFIM	Shang J. et. al. *[25]*	CoFIM: A Community-based Framework for Influence Maximization on Large-Scale Networks	O ($k^2.n.k_{max}$)	Non-Overlapped Clustering	Community structure was combined with Influence Diffusion Models

I. CONCLUSION

As we can see from the preceding discussion, community-based influence maximisation algorithms are typically quicker than standard greedy algorithms. Furthermore, because these algorithms generally assume that distinct communities are separated, they naturally allow parallelization. The drawbacks of existing community-based algorithms, on the other hand, are clear. First, determining a node's marginal gain inside its own community necessitates Monte-Carlo simulations, which takes time and restricts the use of these methods on large networks.

Second, because the algorithms estimate a node's effect on the entire network based on its influence inside its own community, the algorithm's accuracy will be heavily reliant on the underlying community structure. The approximation will be good if the links between various communities are relatively scarce. The approximation, on the other hand, will be poor, and the algorithm may produce erroneous results. Third, certain community-based algorithms rely on a specific diffusion model; however, it is unknown if these algorithms will function well under the classis IC and LT models.

REFERENCE

[1] Tao Zhou, Jiuxin Cao, Bo Liu, Shuai Xu, Ziqing Zhu, and Junzhou Luo. Location-based influence maximization in social networks. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, page 1211–1220, New York, NY, USA, 2015. Association for Computing Machinery.

[2] Guoliang Li, Shuo Chen, Jianhua Feng, Kian-lee Tan, and Wen-syan Li. Efficient location-aware influence maximization. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14, page 87–98, New York, NY, USA, 2014. Association for Computing Machinery.

[3] Au Wang, au Zhang, au Zhang, and au Lin. Distance-aware influence maximization in geo-social network. In 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pages 1–12, 2016.

[4] David Kempe, Jon Kleinberg, and Eva Tardos. Maximizing the spread of influence through ´ a social network. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, page 137–146, New York, NY, USA, 2003. Association for Computing Machinery.

[5] Wen-Yuan Zhu, Wen-Chih Peng, Ling-Jyh Chen, Kai Zheng, and Xiaofang Zhou. Exploiting viral marketing for location promotion in location-based social networks. ACM Trans. Knowl. Discov. Data, 11(2), nov 2016.

[6] Zhen Zhang, Xiangguo Zhao, Guoren Wang, and Xin Bi. Multi-location influence maximization in location-based social networks. In Leong Hou U and Haoran Xie, editors, Web and Big Data, pages 336–351, Cham, 2018. Springer International Publishing.

[7] Suqi Cheng, Huawei Shen, Junming Huang, Guoqing Zhang, and Xueqi Cheng. Staticgreedy: Solving the scalability-accuracy dilemma in influence maximization. In Proceedings of the 22nd ACM International Conference on Information amp; Knowledge Management, CIKM '13, page 509–518, New York, NY, USA, 2013. Association for Computing Machinery.

[8] Chen Wei, Wang Yajun, and Yang Siyu. Efficient influence maximization in social networks. KDD'09, Paris, France. ACM 978-1-60558-495-9/09/06, 06 2009.

[9] Mohammadreza Hosseinpour, Mohammad Reza Malek, and Christophe Claramunt. Sociospatial influence maximization in location-based social networks. Future Gener. Comput. Syst., 101:304–314, 2019.

[10] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne VanBriesen, and Natalie Glance. Cost-effective outbreak detection in networks. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, page 420–429, New York, NY, USA, 2007. Association for Computing Machinery.

[11] Zhang Zhen, Zhao Xiangguo, Wang Guoren, and Bi Xin. Multi-location influence maximization in location-based social networks. Springer International Publishing, 2018.

[12] Mark Granovetter. Threshold models of collective behavior. American Journal of Sociology, 83(6):1420–1443, 1978.

[13] Yun Yong Ko, Kyung Jae Cho, and Sang Wook Kim. Efficient and effective influence maximization in social networks: A hybrid-approach. Information Sciences, 465:144–161, October 2018. Publisher Copyright: © 2018 Elsevier Inc.

[14] Tao Zhou, Jiuxin Cao, Bo Liu, Shuai Xu, Ziqing Zhu, and Junzhou Luo. Location-based influence maximization in social networks. pages 1211–1220, 10 2015.

[15] Yadati Narahari. A shapley value-based approach to discover influential nodes in social networks. Automation Science and Engineering, IEEE Transactions on, 8:130 – 147, 02 2011. [16] Chi Wang, Wei Chen, and Yajun Wang. Scalable influence maximization for independent cascade model in large-scale social networks. Data Mining and Knowledge Discovery, 25, 11 2012.

[17] Xiao Li, Xiang Cheng, Sen Su, and Chenna Sun. Community-based seeds selection algorithm for location aware influence maximization. Neurocomput., 275(C):1601–1613, jan 2018.

[18] Yi-Cheng Chen, Wen-Yuan Zhu, Wen-Chih Peng, Wang-Chien Lee, and Suh-Yin Lee. Cim: Community-based influence maximization in social networks. ACM Trans. Intell. Syst. Technol., 5(2), apr 2014.

[19] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. Physical review. E, Statistical, nonlinear, and soft matter physics, 78:046110, 11 2008.

[20] Xiaohang Zhang, Ji Zhu, Qi Wang, and Han Zhao. Identifying influential nodes in complex networks with community structure. Know.-Based Syst., 42:74–84, apr 2013.

[21] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. Physical Review E, 70(6), Dec 2004.

[22] Tianyu Cao, Xindong Wu, Song Wang, and Xiaohua Hu. Oasnet: An optimal allocation approach to influence maximization in modular social networks. In Proceedings of the 2010 ACM Symposium on Applied Computing, SAC '10, page 1088–1094, New York, NY, USA, 2010. Association for Computing Machinery.

[23] Yu Wang, Gao Cong, Guojie Song, and Kunqing Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, page 1039–1048, New York, NY, USA, 2010. Association for Computing Machinery.

[24] Guoliang Li, Shuo Chen, Jianhua Feng, Kian-lee Tan, and Wen-syan Li. Efficient locationaware influence maximization. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD '14, page 87–98, New York, NY, USA, 2014. Association for Computing Machinery.

APPLICATION OF MACHINE LEARNING IN THE HEALTH SECTOR AND AGRICULTURE: A REVIEW

Manpreet Kaur[#1], Sikander Singh Cheema[*2]
[#]*Department of Computer Science Engineering, Punjabi University,Patiala,*Department of Computer Science Engineering, Punjabi University,Patiala*
[1]lectmanpreet1993@gmail.com
[2]cheemasikander8@gmail.com

ABSTRACT— In this research paper we had defined the multiple use of machine learning in various fields. Machine learning is a study that can allow the model to learn from the given dataset and train itself. Machine learning is used in health search so that the diabetic disease can be predicted with the help of machine learning algorithms. Google's machine learning applications are trained to detect breast cancer and give us accurate results. Machine learning is not only used in the health sector it is even very popular in the agriculture field, the crop yield prediction can be done using various machine learning algorithms so that the farmers can predict their growth and take maximum benefits from their production.

KEYWORDS— Machine Learning, Crop yield prediction, Agriculture, Supervised Learning, Applications.

INTRODUCTION

Machine learning is the subset of artificial intelligence; it is the study of an algorithm that can improve itself without being programmed by humans. The algorithm can improve itself through experience and by using the input data, machine learning algorithm generate a model based on some sample dataset, these dataset are known as "training data", the model use this dataset to make predictions and give appropriate results[1]

The primary aim of machine learning is to make an algorithm that learns through experience and provides better solutions every time without human assistance. [2]

METHODS

A. Machine learning methods

Machine learning methods can be classified into three categories.

1) *Supervised machine learning*

Supervised learning as its name implies this learning technique required input data fed into the model; the algorithm is trained by using the labeled dataset so that it can predict the appropriate outcome. As shown in Fig. 1 clearly defined the supervised machine learning. There are various methods that are used in supervised learning these are neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM),etc[3]

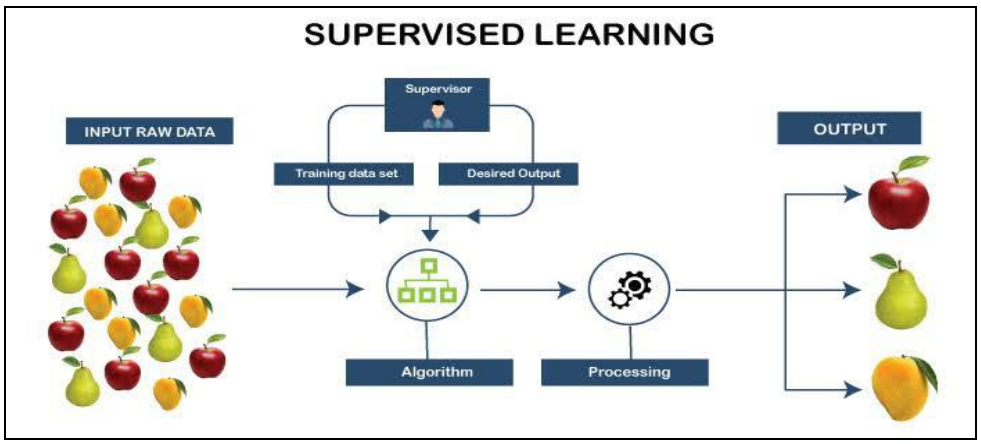


Fig. 1 Supervised machine learning

2) *Unsupervised machine learning*

In unsupervised machine learning techniques the algorithm does not provide any training, the model has to learn by itself. The algorithm analyzes the unlabeled dataset and makes a cluster or group of similar types of dataset. Fig. 2 defines the unsupervised machine learning. The algorithm searches for similarities and differences in the dataset without the help of human being. Various algorithms used in unsupervised learning are neural networks, k-means clustering, probabilistic clustering methods etc. [3]

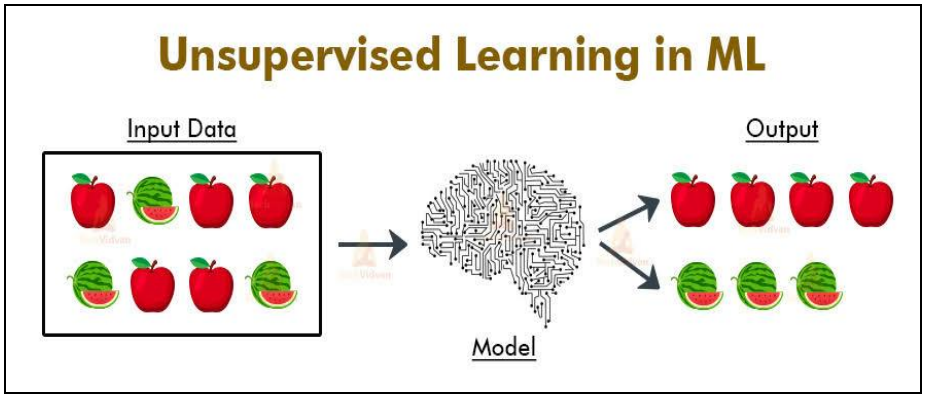


Fig. 2 Unsupervised machine learning.

3) *Reinforcement machine learning*

This machine learning technique does not train its model with a sample dataset, but the model has to learn through it's experience and past events. It can use trial and error methods to learn and give the desired outcomes. The system improves itself again and again to give an appropriate solution for a problem. Fig. 3 defines the reinforcement machine learning. [3]

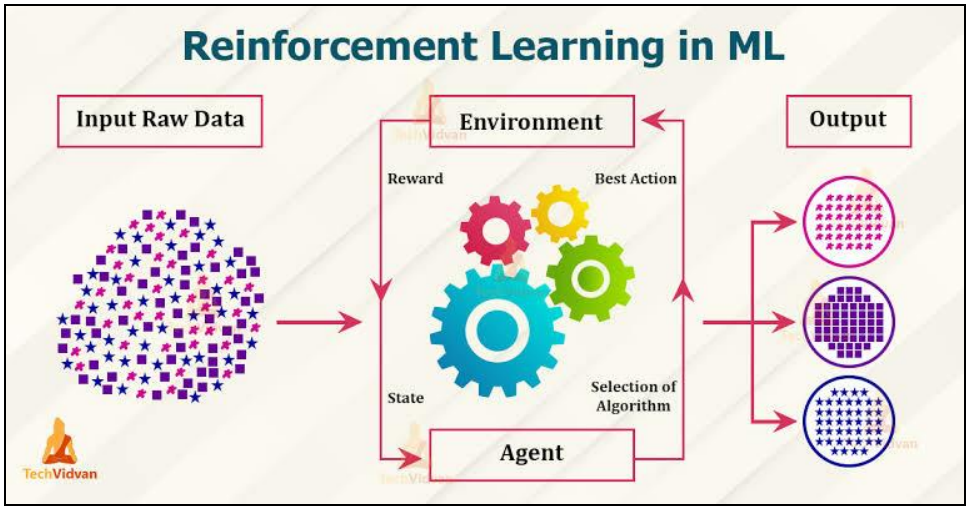


Fig. 3 Reinforcement machine learning.

APPLICATIONS OF MACHINE LEARNING IN HEALTH SECTOR

Various research papers can be studied so that the objective of application of machine learning should be clear, there description is given below:-

In this research paper the author defines the three main elements that are required in medical imaging, as well as natural language processing system to generate the medical documents and to get the information, the author examine history of machine learning, and various technique and latest technology that is used in health sectors [4]

In this research paper, the author has developed a machine learning algorithm that can predict acute toxicities in patients by performing radiation therapy for head and neck cancers. The author defines that deep learning is very useful in the health sector. It identifies difficult patterns by itself and even helps the doctors to diagnose the disease and take appropriate action by studying various radiology reports. Google's machine learning application are trained and detect breast cancer even it gives 89% accuracy rate. [5]

Amine Rghioui al. [6] defines a machine learning algorithm for continuous monitoring of the diabetic patients. The author combines machine learning with the internet of things, sensors etc. A smart system can collect data from the patients and then classify dataset using machine learning algorithms so that disease can be diagnosed. The author use various machine learning algorithms such as sequential minimal optimization The author uses WEkA tool with six different machine learning algorithms, by comparing various algorithm, he found that sequential minimal optimization (SMO) gives us accurate results its accuracy and precision is 99.66% whereas sensitivity is 99.85% [6]

APPLICATIONS OF MACHINE LEARNING IN AGRICULTURE SECTOR

In this research paper [7] the author defines various machine learning techniques that can extract new knowledge through decision rules to get the best approach. The model is based on inductive and iterative processes the model again and again searches for the pattern and modifies its previous knowledge. The result of this research is to create an effective decision rules that can predict the plants and their states even unwanted impacts of water on plants can be prevent

In this research paper[8], the author can categorize the data into four sections it includes (a) crop management, disease detection,(b) livestock management (c) water management (d) soil management. By applying many types of machine learning techniques to the sensor data the farmers can get the right decision and support for their crops. The author concluded that in crop management, soil management, water management the artificial neural networks (ANN) is used whereas in livestock management support vector machine (SVM) provide us best results. [8]

Parth khunteta al. [9] considers agriculture to play an important role in the Indian economy. Different changes in market, season and pattern of the crop may be a reason for loss for the farmers; the author said this can be overcome by prediction of the dataset of weather and crop's types that can even help the farmers to get maximum profit. In this article various algorithms are used so that predictions of crop production can be predicted. The author applies the Bayesian network for statistical analysis and then artificial neural networks to compare the hidden pattern between them. [9]

In this research paper is mainly based on machine learning and image processing, the author defines that many researcher had used various machine learning method such as support vector machine and artificial neural networks and also image Processing is applied on the photos of crops, seed, soil etc so that the graphical image can be formed. By applying various technique the growth of the crops can be increased and get maximum productivity with minimum equipment. [10]

CONCLUSION

Machine learning is the fastest and quickly growing field of computer science. Machine learning algorithms are used in every field of study, machine learning can solve many difficult problems or problems which take too much time to solve a particular problem that humans cannot solve easily. The machine learning algorithms simply works on learning patterns in the data and make predictions related to the pattern and take accurate results

In this research paper we had studied various research papers related to the application of machine learning in health sector, the various algorithms are used to study the accurate result but for diabetes diagnosis google's machine learning algorithms provide 89% results [5], and the sequential minimal optimization (SMO) gives 99.66% results for the diabetic patients[6], similarly in agriculture we studied different research papers and we found that support vector machine and artificial neural networks are best for getting result for the prediction of the crops.

REFERENCES

[1] *Machine learning*. (2003, May 25). - Wikipedia. [Online] Available: https://en.m.wikipedia.org/wiki/Machine_learning

[2] [Online] Available https://www.expert.ai/blog/machine-learning-definition

[3] [Online] Available https://www.ibm.com/in-en/cloud/learn/machine-learning

[4] [Christopher Toh] and [James P. Brod] Applications of Machine Learning in Healthcare. smart-manufacturing-when-artificial-intelligence-meets-the-internet-of-things

[5] [Online] Available https://www.foreseemed.com/blog/machine-learning-in-healthcare

[6] Amine Rghioui , OrcID,Jaime Lloret , OrcID,Sandra Sendra 2OrcID and Abdelmajid Oumnad *A Smart Architecture for Diabetic Patient Monitoring Using Machine Learning Algorithms*

[7] Savvas Dimitriadis and Christos Goumopoulos *Applying machine learning to extract new knowledge in precision agriculture applications* 2008 Panhellenic Conference on Informatics, 100-104, 2008

[8] Konstantinos G. Liakos , Patrizia Busato , Dimitrios Moshou , Simon Pearson and Dionysis Bochtis Machine Learning in Agriculture: A Review

[9] *Smart farming prediction using machine learning* .International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7 May, 2019

[10] Ishma Mohiuddin1* and Mirza Mohtashim Alam. *A Short Review on Agriculture Based on Machine Learning and Image Pre Processing* ACTA SCIENTIFIC AGRICULTURE (ISSN: 2581-365X) Volume 3 Issue 5 May 2019

COMPARATIVE STUDY OF MACHINE LEARNING MODELS ONHEART FAILURE DETECTION

Vikram Balaji[1], Nirogi Surya Priyanka[2], N Ganesh[3], Deepankur Kansal[4] PrathmeshChandwade[5], and Siddhant Manoj Wange[6]
College of Engineering, Guindy, India. vikrambalaji2k@gmail.com
Geethanjali College of Engineering and Technology, Hyderabad, India.
nsuryapriyanka@gmail.com
Mahatma Gandhi Institute of Technology, Telangana, India.
ganeshrohitnirogi@gmail.com
Indian Institute of Technology, Kanpur, Meerut, India. deepank@iitk.ac.in
Veermata Jijabai Technological Institute, India. prathmesh111999@gmail.com
Government College of Engineering, Aurangabad, India. siddhantwange@gmail.com

Abstract—— Diseases like cardiovascular can kill nearly 17 million people globally and the only common cause is heart related problems and in myocardial infarctions. The main reason behind heart failure is that the blood required for the body can not be pumped by the heart. To make an accurate system that can detect heart failure depends on quantitative symptoms, lab results, etc. Machine learning can help in this task and predict the survival of the human from the given points in data. For this paper we have selected a study of 299 patients on which we have applied different machine learning algorithms to foretell the survival of patients. We discuss about certain machine learning techinques and compare their performance on this data.

Keywords—— SVM, KNN, Decision Tree, RFC, Logistic Regression, Cardiovascu- lar, Heart

1 INTRODUCTION

Cardiovascular diseases are problems of heart along with blood vessels, apart from this heart failure, cerebrovascular diseases contribute to the passing away of over 17 million people globally[5]. The reason in most cases of heart failure is the reason of high blood pressure, diabetes or heart diseases or conditions.[1]. The HFrEF is the heartfailure caused by the left ventricular systolic heart problem and the range of fraction ofejection is less than 40%[18]. The HFpEF is also known as heart failure having normalfraction of ejection where the left ventricle being able to normally contract, but is stiffwhich causes failure to relax systemmatically during diastole, which results in filling impairing[19][17].

Heart being the most vital organ the heart failure diagnosis (HFD) has become a toppriority for physicians and doctors, but clinical practice has shown that the diagnosis has not reached a high accuracy[13].

Machine Learning techniques applied to medical records can help predict the sur- vival of patients with heart problems [9][8]. And certain other studies also indicate the important features that can cause a heart failure [14][15]. Researchers have taken advantage for of machine learning for predictions [10][25], and at the same time for ranking of features[23]. Image processing has also shown good results when it comes to medical reports[7][21]. Currently deep learning methods have also been applied in thisfield [16][12].

The survival rate prediction using these techniques is still quite low in terms of driving factors and accuracy. Several models that are developed reach only a moderate accuracy[24]. Studies are not showing the effect of features on the survival rate ofpatients[22]. Sakamoto in [20] tells us that the reason is reproducibility, which hindersin forming opinions about the factors and at the same time troubles model performance.

The capacity to correctly classify observations is particularly useful for a variety of business usecases, such as forecasting whether or not a user would buy a serviceor forecasting whether or not a given loan will fail. We have a comparative analysisas to which model is able to perform for this specific dataset. The paper is parted into 9 parts where section 1 is introduction, section 2 covers Logistic regression, section 3 covers decision trees, section 4 covers random forest classifier, section 5 covers supportvector machines, section 6 covers K-nearest neighbour, section 7 covers the dataset used, section 8 discusses the approach and result and finally section 9 has the conclusion.

2 Logistic Regression

The oversaw learning game plan computation vital backslide is used to predict the likelihood of a goal variable. Since the possibility of the target or ward variable is di- chotomous, there are only two orders. A determined backslide model predicts P(Y=1) as a component of X mathematically. It's perhaps the most fundamental AI calculation,and it may be used to handle a grouping of request issues like spam acknowledgment, diabetes figure, harmful development assurance, and so forth Straight backslide is rou-tinely parted between three sorts; Binomial, Multinomial and Ordinal. For our usage case we will use binomial straight backslide as it oversees matched portrayal.

3 Decision Tree

Decision tree investigation is a prescient displaying approach that can be utilized in an assortment of circumstances. An algorithmic methodology that can part the dataset in various ways dependent on various conditions can be utilized to make decision trees. The most impressive calculations in the space of managed calculations are decision trees.

They can be utilized for characterization just as relapse. The decision hubs, where the information is parted, and the leaves, where we get the outcome, are the two essential substances of a tree. A decision tree is likea cascade structure in which each inward hubisfeature test (e.g., regardless of whether a flip of coin will land one of heads or tails), eachhub addresses a mark (decision is relating subsequent to processing all elements), and branches address include blends that lead to those class names. With names (Rain(Yes), No Rain(No)), the fundamental progression of a decision tree is portrayed in the image underneath. Decision trees are made utilizing a calculation that decides different ways of dividing an informational dependent on certain elements.

4 Random Forest

Random forest is/are comprised of enormous number of identical number of trees thatare decision trees which cooperate as outfit. Every tree here in the forest that is randomcreates a class expectation, and the class with the greatest number of votes turns into theforecast.

The key is the very low association between's models that is achieved. Uncorrelated models are capable of giving outfit measures that are ultimately more careful than any before of the solitary assumptions, similar to previously how low-association adventures (such bonds adn stocks) join to collect a bag that is very much bigger than the measureof its parts. The only explanation is that, this surprising effect is that the individual trees protect every other within their own slips up. A couple of trees may be very accuratelymistaken, various different will be correct, enabling them, the trees to move in the direction which is correct way all things considered. Likewise, all together for irregularwoodland to perform adequately, the going with conditions ought to be met:

1. Our highlights should have some authentic sign all together for models made withthem to beat random speculating.
2. Individual tree projections (and in this manner mistakes) should have low relation-ships with each other.

5 Support Vector Machines

Support Vector Machine is conceivably the widely well known Supervised Learning calculation, which is targetted for classification at the same time as regression tasks. Nevertheless, fundamentally, it is widely targettted to use for classification targets.

The usecase of the SVM computation is to ultimately try to settle on the best line along with the limit decision that can disconnect the space which is n-dimensional intovarious different classes so that they can without a very remarkable stretch place the brand new data point that we get in the right characterization after on. A hyperplane isknown as the best decision limit.

6 k-Nearest Neighbors

K-Nearest Neighbor is computations reliant upon Supervised Learning. Computation in KNN anticipates the equivalence between the available and new cases and put the new case into the characterization is for the part like the open classes.It stores all of the available point and gatherings another point reliant upon the equivalence. Suggesting when a brand new point appears then it will in general be very easily organized intoa good suite characterization. K-NN computation can be used for Regression similarlywith respect to Classification anyway generally it is used for the Classification issues.

K-NN is a non-parametric computation, which says that it doesn't necessarily make anyassumption on principal data. KNN computation at the arrangement stage essentially holds the points important and when it is given brand new points, it then tries to club them into a similar class.

7 Dataset

For the dataset we have utilized the cardiovascular breakdown dataset gave to us by theUCI AI storehouse. Which has the clinical records of 299 cardiovascular breakdown patients which are gathered from Faisalabad Institute of Cardiology and Allied Hospitalin Faisalabad (Punjab, Pakistan), in 2015 (AprilDecember) [6] . There are a sum of 105ladies and 194 men. There are a sum of 13 elements on which the forecast can be made. A portion of these highlights are twofold similar to hypertension, paleness, diabetes, smoking, sex. The information in the sickliness area is thought of if the patient had haematocrit level lower than 36% [6]. CPK will in general stream into the blood when the muscle tissue gets harmed, so the significant degrees of CPK may show heart failure[2].The serum creatinine can be considered as a side-effect which is produced by the creatine when a specific muscle separates, these levels help in location of kidney work, taking everything into account significant degree of serum creatinine it demonstrated renal dysfuntion[3]. Another component launch part is the level of measure of blood that theleft ventricle can siphon out in every constriction. Right muscle work is benefited by sodium mineral, serum sodium is one more element in the dataset which is a typical blood test which tells us if the sodium levels in the blood. Low degree of sodium can cause heart failure[4]. Passing occasion is an element of the dataset that is utilized an objective for the grouping, shows if the patient endure or kicked the bucket before thesubsequent period which was 130 days[6]. According to the dataset the quantity of survivors is 203 (passing occasion =0) and there are 96 dead patients (demise events=1).From this we can say genuinely that 32.11% are positive and 67.89% are negative.

8 Proposed Approach and Results

For our comparative analysis we applied all the above described machine learning models to compare the results on heart failure detection dataset.

Logistic Regression model yielded 73% test accuracy when no hyper parameter tuning was applied to it. With the help of GirdSearchCV and setting the random state tobe 101, the following parameters were tuned :

1. Sover
2. Penalty
3. C value

With the help of corrected parameters, the underlying accuracy that of Logistic Re- gression model was improved to 87%. To further increase the test accuracy without compromising overfitting, the train and test attributes were scaled. Out of MinMax, Robust and Standard scalers, the use of MinMax scaler resulted in an accuracy of 90%.

Class	Precision	Recall	F1
0	0.93	0.93	0.93
1	0.80	0.30	0.80

Table 1. Logistic Regression Evaluation Results

The KNN algorithm gave an accuracy of 48% when no hyper parameter tuning wasapplied. Using GridSearchCV and setting the random state as 101, the number of nearest neighbours leaf size and p value was optimised. The number of nearest neighbours wasset to 5. After this the attributes were scaled which resulted in 75% accuracy through MinMax scaler, 78% through Standard scaler, and 85% through Robust scaling.

Class	Precision	Recall	F1
0	0.86	0.96	0.91
1	0.80	0.53	0.64

Table 2. KNN Evaluation Results

Using Decision Tree algorithm without hyper parameter tuning, the test accuracy was noted as 75%. By fixing the random state and gradient boosting techniques, the accuracy increased to 82%. This was further improved by scaling techniques. All 3 scaling methods resulted in the same test accuracy of 92%.

Class	Precision	Recall	F1
0	0.93	0.96	0.95
1	0.86	0.80	0.83

Table 3. Decision Tree Evaluation Results

Random Forest Classifier with a randoms state of 101 and with no hyper parametertuning gave 85% test accuracy but the maximum depth and the number of estimators were determined using GridSearchCV. These optimal values were 4 and 28 respectively.Adding further scaling to the attributes yielded 93% accuracy in MinMax scaler, 88% in Standard scaling and 92% in Robust scaling.

Class	Precision	Recall	F1
0	0.93	0.96	0.95
1	0.86	0.80	0.92

Table 4. Random Forest Classifiers Evaluation Results

Support Vector Machines with the help of only linear kernel and the default settingwas able to give us about 83% accuracy. Different kernels were used to bump up the accuracy but the model was not able to perform any better than this.

Class	Precision	Recall	F1
0	0.83	0.84	0.82
1	0.76	0.60	0.64

Table 5. Support Vector Machines Evaluation Results

Table 6. Caption

Paper	Accuracy
[11]	82.22%
Our approach	93.4% (Random Forest)

We have also plotted a Region over curve (ROC) for the comparison of different algorithm to evaluate which model was able to outperform other models. From Figure we are able to say that random forest classifiers were able to outperform other models with ROC value of 0.821. We can attribute this performance due to the ensemble of many decision trees.

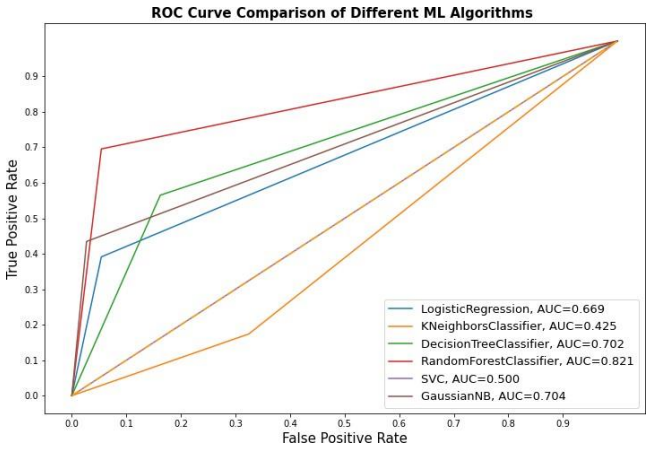


Fig. 1. ROC comparision for all models.

9 Conclusion

Heart diseases have formed to be the most usual and common factor of death in humans. There was a need to evaluate if there were factors which could make people aware about the heart problem in due time to prevent heart attacks and other cardiovascular problem. Machine learning have been used to predict patterns in data and hence, this paper dwells upon different machine learning models to help understand their performance on heart dataset[6]. From our analysis we were able to say that random forest classifier were able to outperform previous methods due their ensemble decision trees with and ROC valueof 0.821. And this method was able to outperform previous approaches.

References

1. National heart lung and blood institute (nhlbi). heart failure. Accessed 10 May 2021.
2. Johns hopkins rheumatology. creatine phosphokinase (cpk). Accessed 15 May 2021.
3. Stephens c. what is a creatinine blood test? Accessed 20 May 2021.
4. Case-lo c. what is a sodium blood test? Accessed 25 May 2021.
5. World health organization, world heart day. Accessed 7 May 2021.
6. Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza. Survival analysis of heart failure patients: A case study. *PloS one*, 12(7):e0181001, 2017.
7. Tariq Ahmad, Lars H Lund, Pooja Rao, Rohit Ghosh, Prashant Warier, Benjamin Vaccaro, Ulf Dahlström, Christopher M O'connor, G Michael Felker, and Nihar R Desai. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal of the American Heart Association*, 7(8):e008081, 2018.
8. Subhi J Al'Aref, Gurpreet Singh, Alexander R van Rosendael, Kranthi K Kolli, Xiaoyue Ma, Gabriel Maliakal, Mohit Pandey, Bejamin C Lee, Jing Wang, Zhuoran Xu, et al. Determi- nants of in-hospital mortality after percutaneous coronary intervention: a machine learning approach. *Journal of the American Heart Association*, 8(5):e011160, 2019.
9. Subhi J Al'Aref, Khalil Anchouche, Gurpreet Singh, Piotr J Slomka, Kranthi K Kolli, Amit Kumar, Mohit Pandey, Gabriel Maliakal, Alexander R Van Rosendael, Ashley N Beecy, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *European heart journal*, 40(24):1975–1986, 2019.
10. Bharath Ambale-Venkatesh, Xiaoying Yang, Colin O Wu, Kiang Liu, W Gregory Hundley, Robyn McClelland, Antoinette S Gomes, Aaron R Folsom, Steven Shea, Eliseo Guallar, et al. Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclero- sis. *Circulation research*, 121(9):1092–1101, 2017.
11. Saba Bashir, Zain Sikander Khan, Farhan Hassan Khan, Aitzaz Anjum, and Khurram Bashir. Improving heart disease prediction using feature selection approaches. In *2019 16th inter- national bhurban conference on applied sciences and technology (IBCAST)*, pages 619–623. IEEE, 2019.

12. Jan Walter Benjamins, Tom Hendriks, Juhani Knuuti, Luis E Juarez-Orozco, and Pim van der Harst. A primer in artificial intelligence in cardiovascular medicine. *Netherlands Heart Journal*, 27(9):392–402, 2019.TA Buchan, HJ Ross, M McDonald, F Billia, D Delgado, JG Duero Posada, A Luk, GH Guyatt, and AC Alba. Physician prediction versus model predicted prognosis in ambulatory patients with heart failure. *The Journal of Heart and Lung Transplantation*, 38(4):S381, 2019.
13. Warwick B Dunn, David I Broadhurst, Sasalu M Deepak, Mamta H Buch, Garry McDowell, Irena Spasic, David I Ellis, Nicholas Brooks, Douglas B Kell, and Ludwig Neyses. Serum metabolomics reveals many novel metabolic markers of heart failure, including pseudouridine and 2-oxoglutarate. *Metabolomics*, 3(4):413–426, 2007.
14. Joe Gallagher, Darren McCormack, Shuaiwei Zhou, Fiona Ryan, Chris Watson, Kenneth McDonald, and Mark T Ledwidge. A systematic review of clinical prediction rules for the diagnosis of chronic heart failure. *ESC heart failure*, 6(3):499–508, 2019.
15. Chayakrit Krittanawong, Kipp W Johnson, Robert S Rosenson, Zhen Wang, Mehmet Aydar, Usman Baber, James K Min, WH Wilson Tang, Jonathan L Halperin, and Sanjiv M Narayan. Deep learning for cardiovascular medicine: a practical primer. *European heart journal*, 40(25):2058–2073, 2019.
16. Gavin A Lewis, Erik B Schelbert, Simon G Williams, Colin Cunnington, Fozia Ahmed, Theresa A McDonagh, and Christopher A Miller. Biological phenotypes of heart failure with preserved ejection fraction. *Journal of the American College of Cardiology*, 70(17):2186– 2200, 2017.
17. Fanqi Meng, Zhihua Zhang, Xiaofeng Hou, Zhiyong Qian, Yao Wang, Yanhong Chen, Yilian Wang, Ye Zhou, Zhen Chen, Xiwen Zhang, et al. Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in china. *BMJ open*, 9(5):e023724, 2019.
18. Jan F Nauta, Xuanyi Jin, Yoran M Hummel, and Adriaan A Voors. Markers of left ventricular systolic dysfunction when left ventricular ejection fraction is normal. *European journal of heart failure*, 20(12):1636–1638, 2018.
19. Mari Sakamoto, Hiroki Fukuda, Jiyoong Kim, Tomomi Ide, Shintaro Kinugawa, Arata Fukushima, Hiroyuki Tsutsui, Akira Ishii, Shin Ito, Hiroshi Asanuma, et al. The impact of creating mathematical formula to predict cardiovascular events in patients with heart failure. *Scientific reports*, 8(1):1–12, 2018.
20. Partho P Sengupta, Hemant Kulkarni, and Jagat Narula. Prediction of abnormal myocardial relaxation from signal processed surface ecg. *Journal of the American College of Cardiology*, 71(15):1650–1660, 2018.
21. S Sharmila and MP Gandhi. Analysis of heart disease prediction using data mining techniques. *International Journal of Advanced Networking and Applications*, 8(5):93–95, 2017.
22. Swati Shilaskar and Ashok Ghatol. Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. *Expert Systems with Applications*, 40(10):4146–4153, 2013.
23. David H Smith, Eric S Johnson, Micah L Thorp, Xiuhai Yang, Amanda Petrik, Robert W Platt, and Kathy Crispell. Predicting poor outcomes in heart failure. *The Permanente Journal*, 15(4):4, 2011.
Stephen F Weng, Jenna Reps, Joe Kai, Jonathan M Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PloS one*, 12(4):e0174944, 2017.

COMPARITIVE ANALYSIS OF DIFFERENT SDN CONTROLLERS: A REVIEW

Shivani[#1], Abhinav Bhandari[#2]
Department of Computer Science and Engineering, Punjabi University Patiala
[1]Shivanir629@gmail.com
[2]bhandarinitj@gmail.com

ABSTRACT- Software defined networking (SDN) is a centralized structure for network that can communicate and command the rest of network. It introduces the concept of abstraction and orchestration in network. Network now is divided into two layers control plane and data plane. These planes interact with each other by various protocols. For communication between two planes of SDN i.e. control plane and data plane Open Flow protocol is used mainly.

Control plane is essential part of SDN paradigm. It is the base of network automation through SDN. So it is mandatory to give attention to the designing and selection of controller. Across the last few years due to rapid development and introduction of SDN controllers in research community bring forth the obstacles to choose a suitable controller and to collate which controller is better. This paper studies and provide feature based comparison of different SDN controller and their deployment in various organizations. This work provide distinction among various open source controller ONOS, Ryu, Open Day light Floodlight etc. by evaluating their features and usage whether in academics or industries.

KEYWORDS - SDN, OpenFlow , SDN Controller , Northbound interface , Southbound interface

I. INTRODUCTION

In recent times internet services are changing swiftly and amount of applications are increasing quickly on web, which needs to be manage appropriately. Managing these issues with underlying traditional architecture of network is quite burdensome. Traditional architecture control is based mainly on routers and switches that forward data packets and taking every possible decision to be applied over them. There is complete association in between the data plane (forward the traffic according to control plane) and control plane (take decision regarding traffic management) makes it hard to manage the network. [1]

Main aim behind software defined networking is centralization of whole network by detaching control plane and data plane on same switch to merely data plane on switch and control plane is programmed separately. All the switches and routers in network are managed by one control unit which is known as controller. The inflexible architecture of traditional networking forbid the programmability feature of devices and cause obstruction in meeting client requirements.[2] SDN is a centralized structure that can communicate and command the rest of network. Legacy network have devices that mostly belongs to individual vendors and have proprietary standard. This makes addition of new network device in existing system quite troublesome. SDN is an advance approach in networking world to overcome existing issues related to network architecture and functionalities. This paper is divided into sections where we discussed about comparison of SDN with traditional networking, architecture and components of SDN, protocol used mainly OpenFlow protocol and our main focus is to comparatively evaluate various controllers on the bases of features and usability.

II. TRADITIONAL VS SDN

In traditional networking every routing device has its own data plane and control plane. There are no predefined routes available for transmission of data packets. Every device decide on its own where to send data and through which way. There is no central device that can command entire network. In SDN we basically separate control plane and data plane. SDN controller is a control plane element and switches resides in data plane as forwarding engine. A single controller can manage more than one switch. This make network more organized and easy to operate. This system will be easily configurable and easy to administrate.If both data plane and control plane resides on same device then it will be problematic to manage the network well. There may be coordination issues to generate global routing table among different control plane of different devices due to proprietary standards. SDN uses automation it means we use software to automate network and manage it well to increase the efficiency.

III. SOFTWARE DEFINED NETWORKING

With rapid advancement in networking world there originate a concept of network abstraction and automation i.e. SDN. Software defined networking is a foundation which involves in separating network control function from its forwarding function. Network comprises of two things mainly. First is data plane whose work is to forward different data packets to various destinations. Every router has its individual routing table that contain information about routing address of various location. These routing tables are managed by intelligent controller in SDN. Second is control plane. In legacy network these tables are managed by individual control plane of every router whereas in SDN single controller can supervise every routing device of network. [4]

Abstraction and automation are the major features of SDN paradigm. Abstraction means to separate forwarding function and control plane. Automation means use of software to automate network and manage it well to increase efficiency of network. Evolution of SDN:

- Development of GSMP (general switch management protocol) in 1996
- IETF developed Forwarding and control element Separation (FORCES) in 2000

- OpenFlow Protocol

SDN Architecture:

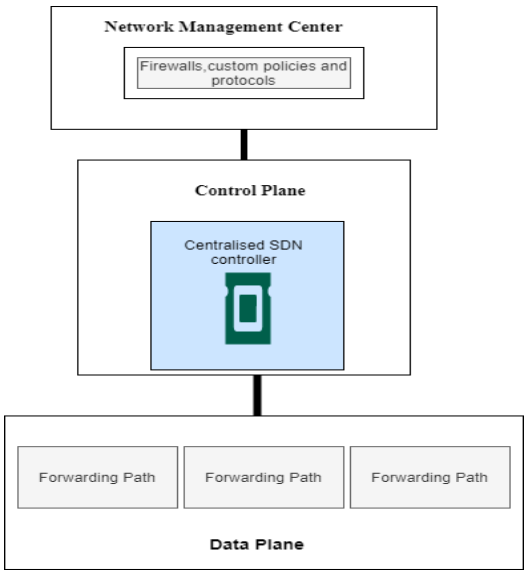


Fig 1 Basic architecture of SDN

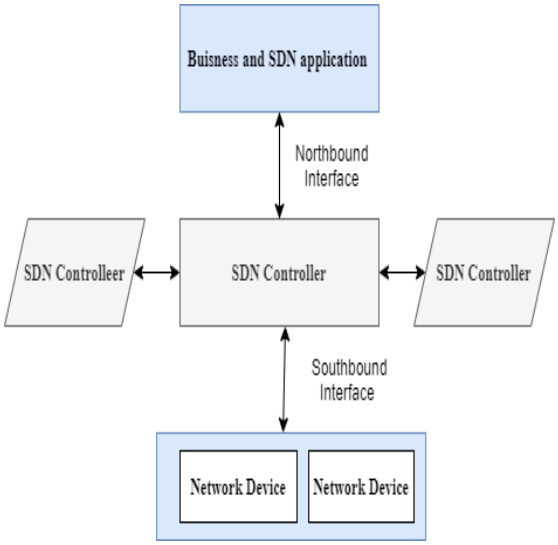


Fig 2 API directionality in SDN

Compared to traditional architecture SDN paradigm involve three layered architecture and these are:

Data Plane, Control Plane, Application layer

1) **Data plane and Southbound interface:** Data plane is forwarding hardware of SDN and whenever controller needs to interact with this plane there are some protocols to be followed, OpenFlow is the most popular southbound protocol.
2) **Control plane/ Controller:** Controller is primarily used to manage the network and apply customize policies and protocols across network devices.
3) **Northbound interface:** Northbound interface represent a programming interface between software module of controller and SDN application running on the network platform.
4) **East-West protocol:** In multi-controller environment East West protocols are used to manage interaction among different controllers.[2]

IV. OpenFlow

Standardization of OpenFlow protocol has lead the foundation of SDN architecture and became first communication protocol for interaction between data and control plane of architecture. ONF (open networking foundation) defined OpenFlow protocol.[5]

OpenFlow switches are of two types mainly: OpenFlow–only, OpenFlow- hybrid. [2]

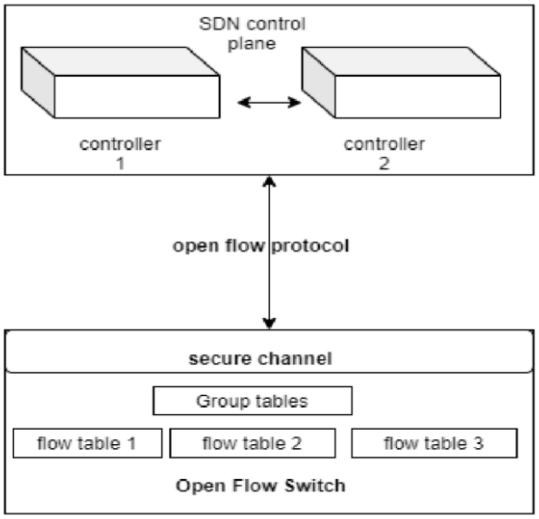


Fig 3 Basic architecture of OpenFlow

OpenFlow is mainly comprises of three components: OpenFlow Switch, OpenFlow Channel and OpenFlow Controller.[6]

OpenFlow switch: These switches are different from switches used in traditional architecture most of them follows the proprietary standard .OpenFlow switches are supervise by OpenFlow controllers over a secure channel using the OpenFlow protocol. Switch contain flow tables used for packet forwarding and lookup. A flow table is composed of a list of flow entries while each entry contains header fields, counters and different actions to be done on packet. Header fields are used to match against packets and information resides in header are VLAN ID, source and destination ports, IP address and so on. Matching process depicted with diagram:

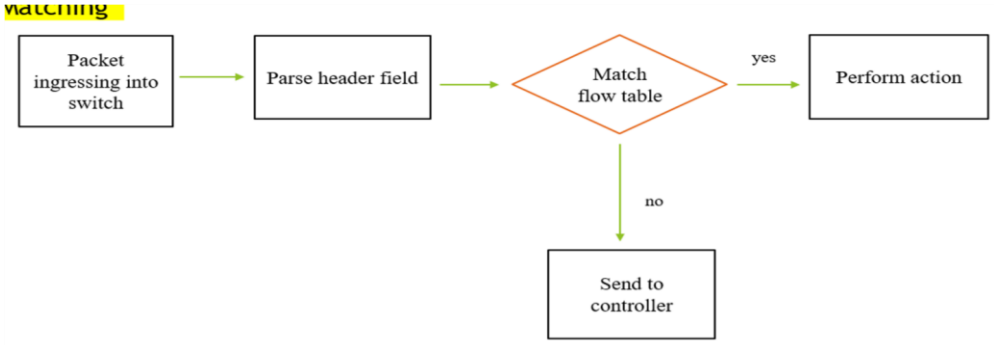


Fig 4 Matching of flow entries

OpenFlow Channel: OpenFlow channel acts as interface between OpenFlow switches and controllers. Three types of communication between switch and controllers are controller-to-switch, Asynchronous and Symmetric. Controller-to-switch messages are those messages only allowed to sent by controller to switch regarding switch state and behavior. Asynchronous messages are sent by the switch to controller for updation regarding various network events and changes occur in state of switch. Symmetric messages are allowed for both switches and controller without any restriction.[6]

OpenFlow controller: This centralized controller is responsible for sustaining, distributing and updating policies and instructions to the network devices. It regulate how to manage packets with invalid flow entries. Controller can also remove or add flow entries based on matching rules.

SDN is capable of programming more than one switch concurrently; but it is still a distributed system and, it has to deal with issues like dropping of invalid packets ,delay of control packets and many more. SDN is becoming easier to implement and deploy with the standardization of OpenFlow protocol in industries. The control plane forms the routing table while the data plane, utilizing the table to determine where the packets should be sent to.[7] Various companies keep on implementing OpenFlow protocols within their data center networks to perform various operations. It helps various organization to maintain their network properly and individually.

Other SDN standard are **IEEE P1520** having defined programmable network interface.[8]

FORCES (forwarding and control element separation) by **IETF** (Internet Engineering Task Force)

V. COMPARITIVE ANALYSIS OF SDN CONTROLLERS

A controller is the crucial element of any SDN- based network, As SDN is being used as an replacement of legacy networks, different controllers have been introduced till now.[9] Controller work as a piece of software for the proper sharing of different resources of network among various application implemented over the network. These application can be security application or load balancing applications.[10] In this paper we provide comparative study of different SDN controllers. With evolution of SDN, vendors have started to develop controllers complying with ONF (open networking foundation) standards. As a result, many SDN controllers were developed, which need to be studied to compare their performance in different environments.[9] Table I represents some of the well-known SDN controllers along with the developers, platform and language supported by them.

Table I Different SDN Controllers

Project	Platform support	Programming language	Developer	Description
NOX	Linux	C++	Nicira	First OpenFlow Controller
POX	Linux Mac OS and window	Python	Nicira	Python version of NOX, used for faster development and prototyping of new network applications
RYU	Linux	Python	NTT,OSRG group	Network controller with APIs for creating applications
Beacon	Mac OS Linux and windows	Java	Stanford	Java based OpenFlow controller
Flood Light	Mac OS Linux and windows	Java	BigSwitch	Derived from Beacon
OpenDay Light	Linux	Java	Linux Foundation	All purposes SDN controller
ONOS	Linux	Java	ONF Linux Foundation	0pen source SDN controller for building next-generation virtualization solutions
IRIS	Linux ,Mac OS and Window	Java	ETRI(electronics and telecommunication research institute)	Open source controller to solve scalability issue

SDN controller features we considered as a base for comparison are cross platform compatibility, Protocols and APIs at Southbound interface and Northbound interface, OpenFlow support , network programmability efficiency(performance ,reliability ,security, scalability).[11] In table II we will do feature based comparison of different controllers by considering previous work done by various researchers. To generate the properties of different controllers various online sources such as journals, conferences, workshops and official websites of controllers were visited ,rectified and consolidated.[12].

Table II Feature Based Comparison

Controllers and features	Programming Language	GUI	Documentation	Modularity	Distributed centralized	Northbound APIS	Southbound APIs	Multithreading support	Application Domain	Virtualization
OOS	Java	Web- based	Good	High	Distributed	Rest API	OpenFlow1.0, 1.3, NETC-ONF	Y	Datacenter,WAN and Transport	Mininet and OVS
OpenDay - light	Java	Web- based	Very good	High	Distributed	Rest API	OpenFlow1.0, 1.3, 1.4, NETCONF/ YANG, OVSDB, BGP/LS, SNMP	Y	Datacenter	Mininet and OVS
NOX	C++	Python +QT4	Poor	Low	Centralized	Rest API	OpenFlow1.0	NOX_MT	Campus	Support
POX	Python	Python +QT4	Poor	Low	Centralized	Rest API	OpenFlow 1.0	N	Campus	Support
RYU	Python	Yes	Fair	Fair	Centralized	Rest API	OpenFlow 1.0, 1.2, 1.3, 1.4, NETCONF, OFCONFIG	Y	Campus	Support
Beacon	Java	Web- based	Fair	Fair	Centralized	Rest API	OpenFlow 1.0	Y	Research	Support
Maestro	Java	-	Poor	Fair	Centralized	Rest API	OpenFlow 1.0	Y	Research	Support
Floodlight	Java	Web/ java based	Good	Fair	Centralized	Rest API	OpenFlow1.0 , 1.3	Y	Campus	Mininet and OVS
IRIS	Java	Web- based	Fair	Fair	Centralized	Rest API	OpenFlow 1.0 ,1.3 OVSDB	Y	Carrier grade	Support
Runos	C++	Web- based	Fair	Fair	Distributed	-	OF 1.3	Y	WAN telecom ,datacenter	-

SDN controllers are probably to be deployed in a data center, it is mandatory that these controllers are to be tested for the amount of traffic encounter at different data center.[13] Most of these SDN Controllers are developed in an educational realm for research motivations some of them have become the interest area of various industries as well such as NOX, Beacon and Floodlight.[14] Currently SDN is deployed at higher organization level . Slowly network industry is shifting from traditional to software based network architecture. Here are some of the well known organization that are using SDN-based technology:

For Datacenter manager: ACI(Application Centric Infrastructure) in Microsoft azure and Amazon

For Network manager: Cisco DNA (Digital Network Architecture) center

SD- Access based solutions: Cisco ISE (Identity Service Engine), Cisco Catalyst 9000 Infrastructure.

Cisco SD- WAN in Google

Cisco SD- Branch: Cisco VNFs (Virtual Network Function), Cisco 5000 series ENCS (enterprise network compute system) [15]

SDN Gateway service – Juniper Network

Vendors, such as Cisco, HP, IBM, VMWare, Lumina Networks, and Juniper adopt SDN technology and jumped into the SDN Controller market with their own contributions. The original HP, Cisco, and IBM Controllers are all developed on the bases of Beacon previously and now have shifted toward OpenDay light. Juniper SDN controller is new part of these organisations.As a challenge to OpenDay light controller ONOS (open networking operating system) developed as open source. Microsoft Azure use azure resource manager to handle all the control plane requests. Organization supporting ONOS include Dell EMC, Intel, Google, NTT, Ciena and Juniper Networks. It is distributed under Apache 2.0 License.[16]

VI. CONCLUSION

SDN means now your system network is under control of programming application. This technology indicates softwarization of network architecture. The general architecture of SDN paradigm composed of SDN controllers and SDN enabled switches (forwarding engine). We need to adopt multi controller environment due to scalability feature of SDN .In this paper we did a comparative feature based analysis of different SDN controllers .We concluded that some of the controllers are merely concepts and for research purposes while others being deployed in organization and being tested by different researchers qualitatively and quantitatively both time to time. ONOS, OpenDay light (ODL) and then Ryu is found to be the mostly deployed SDN controller in IT organizations. Moreover choice of controller is entirely dependent on requirements of users whether needed for research purposes or business deployments.

VII. FUTURE RESEARCH RECOMMENDATION

- Coordination between controllers in multicontroller environment for tackling issues of propagation latency.
- Dynamic load balancing among controllers.
- Development of standardized east west protocols for heterogeneous controller communication.

REFERENCES

[1] H. A. Eissa, K. A. Bozed, and H. Younis, “Software Defined Networking,” *19th Int. Conf. Sci. Tech. Autom. Control Comput. Eng. STA 2019*, pp. 620–625, 2019, doi: 10.1109/STA.2019.8717234.

[2] M. Jammal, T. Singh, A. Shami, and Y. Li, “Software-Defined Networking : State of the Art and Research Challenges,” pp. 1–24.

[3] A. Prajapati, A. Sakadasariya, and J. Patel, “Software defined network: Future of networking,” *Proc. 2nd Int. Conf. Inven. Syst. Control. ICISC 2018*, no. Icisc, pp. 1351–1354, 2018, doi: 10.1109/ICISC.2018.8399028. [4] Y. Aggarwal and U. Kumari, “Software Defined Networking : Basic Architecture &,” no. April, 2019, doi: 10.13140/RG.2.2.29261.69605.

[5] S. Mishra and M. A. R. AlShehri, “Software Defined Networking: Research Issues, Challenges and Opportunities,” *Indian J. Sci. Technol.*, vol. 10, no. 29, pp. 1–9, 2017, doi: 10.17485/ijst/2017/v10i29/112447.

[6] W. Li, W. Meng, and L. F. Kwok, “A survey on OpenFlow-based Software Defined Networks: Security challenges and countermeasures,” *J. Netw. Comput. Appl.*, vol. 68, pp. 126–139, 2016, doi: 10.1016/j.jnca.2016.04.011.

[7] K. Bakshi, “Considerations for Software Defined Networking (SDN): Approaches and use cases,” *IEEE Aerosp. Conf. Proc.*, pp. 1–9, 2013, doi: 10.1109/AERO.2013.6496914.

[8] F. Hu, Q. Hao, and K. Bao, “A survey on software-defined network and OpenFlow: From concept to implementation,” *IEEE Commun. Surv. Tutorials*, vol. 16, no. 4, pp. 2181–2206, 2014, doi: 10.1109/COMST.2014.2326417.

[9] S. M. Mohammad and U. States, “Pr ep rin t n ot pe er re v Pr ep t n pe er,” *J. Emerg. Technol. Innov. Res.*, vol. 11, no. 4, pp. 204–217, 2020, doi: 10.5281/zenodo.4742771.

[10] D. B. Hoang and M. Pham, “On software-defined networking and the design of SDN controllers,” *2015 Int. Conf. Netw. Futur. NOF 2015*, 2015, doi: 10.1109/NOF.2015.7333307.

[11] O. Salman, I. H. Elhajj, A. Kayssi, and A. Chehab, "SDN controllers: A comparative study," *Proc. 18th Mediterr. Electrotech. Conf. Intell. Effic. Technol. Serv. Citizen, MELECON 2016*, no. 978, pp. 18–20, 2016, doi: 10.1109/MELCON.2016.7495430.

[12] L. Mamushiane, A. Lysko, and S. Dlamini, "A comparative evaluation of the performance of popular SDN controllers," *IFIP Wirel. Days*, vol. 2018-April, pp. 54–59, 2018, doi: 10.1109/WD.2018.8361694.

[13] Z. K. Khattak, M. Awais, and A. Iqbal, "07097868," 2014.

[14] D. Erickson, "The Beacon OpenFlow controller," *HotSDN 2013 - Proc. 2013 ACM SIGCOMM Work. Hot Top. Softw. Defin. Netw.*, pp. 13–18, 2013, doi: 10.1145/2491185.2491189.

[15] "Cisco Software Defined Networking." https://www.cisco.com/c/en/us/solutions/software-defined-networking/overview.html (accessed Oct. 11, 2021).

[16] "SDN Controllers platform," https://www.sdxcentral.com/networking/sdn/definitions/sdn-controllers/ (accessed Oct. 10, 2021).

AN EVOLUTIONARY APPROACH TOWARDS VIDEO SURVEILLANCE IN SMART CITIES

Himani Sharma[#1], Navdeep Kanwal[*2]
Department of Computer Science & Engineering, Punjabi University Patiala
[1]himanisharma781@gmail.com
[2]navdeepkanwal@gmail.com

ABSTRACT— The increasing importance of multimedia technology has led to a paradigm shift that happened in favour of multimedia forensics. Over the past few decades, the need of multimedia (i.e. pictures, videos and so on) for smart city applications has also been highlighted. Therefore, the development of sophisticated forgery tools is also closely linked with the increasing efficiency of these technologies. The rapidly growing field makes it possible for the attackers to alter the existing information available in the digital form. The process of modification of digital material into some other form or replacing or modifying the main information contained in videos is referred to as forging. Intruders perform forgeries with the help of smart tools and technologies that are readily accessible in the online environment. The present paper demonstrates a thorough review of various video surveillance methodologies introduced by numerous researchers. It particularly addresses the significance of the methods by categorising them according to functionality. This paper provides an authoritative and comprehensive assessment of existing digital video surveillance methods and also identifies different research gaps within the current approaches. Furthermore, there is a need to develop methods that guarantee the credibility of videos and other multimedia in real time surveillance.

KEYWORDS— Smart Cities, Video Surveillance, multimedia, architecture, dataset

INTRODUCTION

Technology is rapidly advancing across the world, and as a result, it is affecting people, places, and things in the real world. Cities are also becoming smarter every day as a result of the integration of billions of data points. The rest of the world aspires to be intelligent in some way as well. Smart cities incorporate diverse technologies and develop smart economics, smart transportation, smart hospitality, smart government, and smart environment. These are all just a few of the programs that work together to create a smart city. A key technical aspect of a smart city is video surveillance, which is also one of these applications. Video surveillance is important as it provides the secure environment to the citizens. This Video Surveillance Systems have become a critical component of smart city infrastructure for improving public security and combating criminal conduct. People, events and actions have all been monitored using video surveillance equipment for a long time. The three most commonly accessible devices in the visual surveillance sector are CCD cameras, thermal imaging cameras, and night vision technologies. These security systems have been implemented to support the modernised smart city systems, for judicial purposes as well as the ever-expanding smart city applications. In home security applications, there is also a growing need for the evaluation of human behaviours and identify individuals for potential threat analysis and identification. The accessibility of the software tools make it possible to alter the capturing videos and replace it with static view or some other thing with the intention of criminal activity [1], [2]. Thus it increases the necessity to develop the tools to ensure the authenticity of surveillance systems data. Fig. 1 illustrates the layers of generalized architecture in smart Video Surveillance System. Consequently, numerous researchers has developed heterogeneous techniques to detect the forgery present in the surveillance systems. Surveillance Applications [3]-[5] in the Smart City:

- Security sensitive locations
- Personal Identification
- Crowd flux statistics and congestion analysis
- Tampering detection and Localization
- False activity alarm
- Interactive surveillance using multiple cameras.

Motivation

Although video surveillance is a fascinating technology with huge potential, the verdict on its usefulness in the real world is still promising. In a restricted environment, current video analytics systems can be used to fulfil the essential tasks. For illustration, video surveillance might be used to identify security measures or car licence plates, public gatherings at high-traffic highways, airports, and railway stations, particularly in financial-based security destinations like banks and ATMs as well as abandoned belongings in public places. However, it would be ineffective in adverse weather, with minimal cars, or in an overcrowded environment. It's possible that the security camera has been tampered with and is merely displaying the previous behaviour to mislead the public. As a result, several academics have formed ways to check the veracity of surveillance videos.

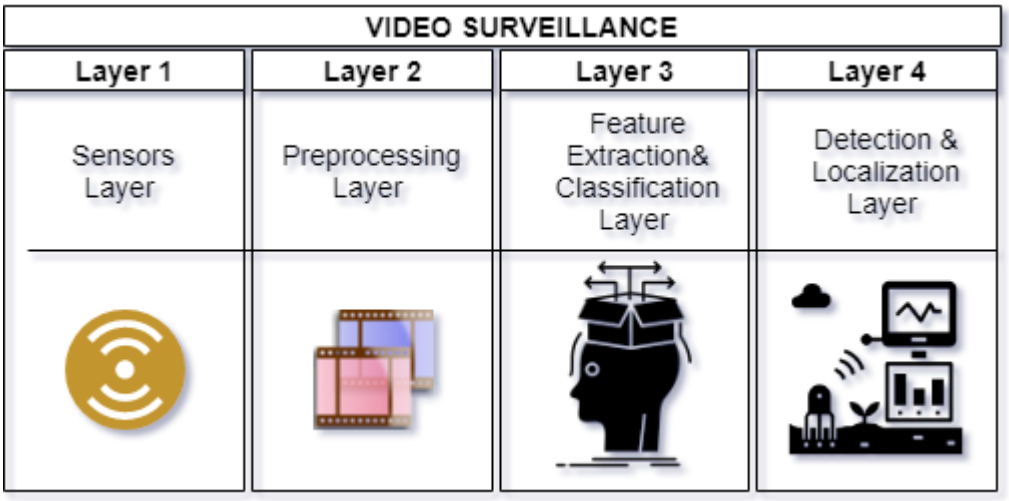


Fig. 1: Generalized Architecture of Video Surveillance System

Paper Organization of the Survey

The current study focuses on the application areas and ongoing difficulties with video surveillance of traditional metropolitan cities, as well as potential solutions. Fig. 2 depicts the organisation of each part, as well as the discussion of all of its subsections.

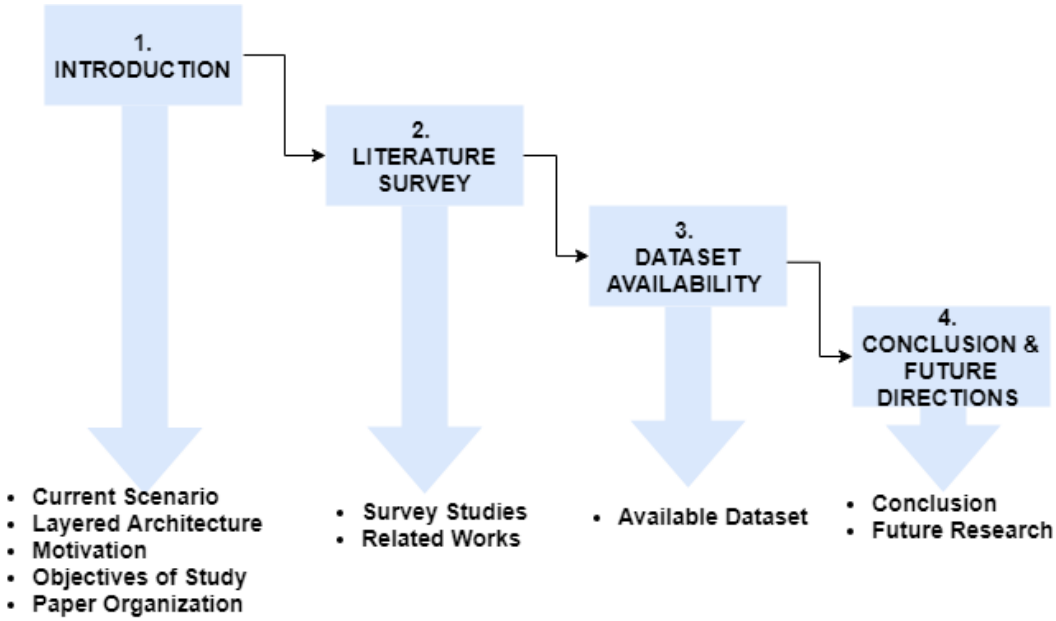


Fig. 2: General Overview of Paper Organization

LITERATURE SURVEY

This chapter presents a comprehensive review of the literature on video surveillance techniques in smart cities, their strengths and weaknesses, and their future possible expansion. An overview of existing survey research is also included in this section.

Related Works

In [6] an author presented an object detection architecture using edge computing for surveillance. By using wireless communication the presented approach achieves an efficient object detection. An innovative secured smart surveillance system that relies on micro services architecture and blockchain technology has also been presented by one of the authors in [7]. Face detection, audio analysis, behavioural analysis, and licence plate recognition are all common surveillance video analysis operations that are frequently run simultaneously and individually. The author in [8] introduced a technique for the identification of target and for this multiple cameras are mounted on the wall. The technique also able to monitor the potentially dangerous activities or targets. The process further introduced for decision making also performs identification that particular is a threat or not. The author [9] described a video surveillance system based on permissioned blockchains

(BCs), edge computing, InterPlanetary File System (IPFS) technology, and convolution neural networks (CNNs). On a broad extent, edge computing is applied to gather and analyse data from sensors that are implemented wirelessly. The huge amount of data stored by utilizing InterPlanetary File System, and to provide real-time surveillance CNN technology is implemented. A data-driven deep learning-based system for smart cities has been developed by the author [10] for sustainable development and via massive video surveillance, rapid action taken to prevent the COVID-19 pandemic. Social distance monitoring is adopted by the author and utilised three object recognition models based on deep learning-based real-time for the detection of individuals presented in the videos which are acquired using monocular camera. The system performance has been evaluated using a real-world video surveillance database for successful execution. An author in [11] suggested a system in which transportation connected with camera nodes is employed as the mobile element of the system. An architecture based on fog computing and wireless visual sensor networks is used to accomplish real-time video surveillance in smart cities. Based on the simulation findings, the system appears to be a viable solution for smart city surveillance applications.

DATASET AVAILABILITY IN VIDEO SURVEILLANCE

Datasets are essential to evaluate the efficiency and verify the results of different techniques developed by the researchers. But only a few video datasets are publicly available. PASCAL Visual Object Classes (VOC) [12] is one of the renowned datasets for object identification research. As part of the visual surveillance study, Oxford University published another dataset from the Oxford Town Center [13]. It includes videos recorded on a city street at a resolution of 1920 x 1080 and captured at 25 frames per second. A further dataset for evaluating videos in a similar area is MS COCO [14]. A VLFD (Video Forensic Library for Frame Duplication) dataset has been introduced by the author in [1] to evaluate the video frame duplication forgery detection techniques.

CONCLUSIONS & FUTURE DIRECTIONS

In a smart city, the main objective of video surveillance is to provide security against illegal events. The primary applications of video surveillance in smart cities, have been examined in this study. The importance of video surveillance in the creation of creative smart cities has also been discussed. Numerous researchers explored different methods to build an effective surveillance system, as shown by the analysis in related work. There are still several places where system functionality may be enhanced for now as well. The system's functionalities will be enhanced in the future in conjunction with pre requisites, such as the use of smart contracts to provide data access control. For the performance assessment and validation of video surveillance methods, the datasets described in the preceding section are adequate. Future researchers may find the study useful in developing new methods for detecting video counterfeiting.

REFERENCES

[1] Sharma, Himani, and Navdeep Kanwal. "Video interframe forgery detection: Classification, technique & new dataset." Journal of Computer Security Preprint (2021): 1-20.

[2] Sharma, Himani, Navdeep Kanwal, and Ranbir Singh Batth. "An ontology of digital video forensics: Classification, research gaps & datasets." 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE). IEEE, 2019.

[3] R. T. Collins, A. J. Lipton, and T. Kanade, “Introduction to the special section on video surveillance,” IEEE Transactions on pattern analysis and machine intelligence, vol. 22, no. 8, pp. 745–746, 2000.

[4] L. J. Fennelly and M. Perry, Physical security: 150 things you should know. Butterworth-Heinemann, 2016.

[5] Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, “Clustered synopsis of surveillance video,” in 2009 Sixth IEEE international conference on advanced video and signal based surveillance. IEEE, 2009, pp. 195–200.

[6] Ren, Ju, et al. "Distributed and efficient object detection in edge computing: Challenges and solutions." IEEE Network 32.6 (2018): 137-143.

[7] Nagothu, Deeraj, et al. "A microservice-enabled architecture for smart surveillance using blockchain technology." 2018 IEEE international smart cities conference (ISC2). IEEE, 2018.

[8] Alshammari, Abdullah, and Danda B. Rawat. "Intelligent multi-camera video surveillance system for smart city applications." 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2019.

[9] Wang, Rong, et al. "A video surveillance system based on permissioned blockchains and edge computing." 2019 IEEE International Conference on Big Data and Smart Computing (BigComp). IEEE, 2019.

[10] Shorfuzzaman, Mohammad, M. Shamim Hossain, and Mohammed F. Alhamid. "Towards the sustainable development of smart cities through mass video surveillance: A response to the COVID-19 pandemic." Sustainable cities and society 64 (2021): 102582.

[11] Mosaif, Afaf, and Said Rakrak. "A New System for Real-time Video Surveillance in Smart Cities Based on Wireless Visual Sensor Networks and Fog Computing." J. Commun. 16.5 (2021): 175-184.

[12] Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." International journal of computer vision 88.2 (2010): 303-338.

[13] Benfold, Ben, and Ian Reid. "Stable multi-target tracking in real-time surveillance video." *CVPR 2011*. IEEE, (2011).

[14] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, (2014).

DEVELOPMENTS IN UNDERWATER IMAGE PROCESSING: ANALYSIS, CHALLENGES AND FUTURE PERSPECTIVE

Sukh Sehaj Singh[1,*], Rohit Sachdeva[2], Rajeev Sharma[3]
[1]*Department of Computer Science & Engineering, Punjabi University, Patiala, India*
[2]*Department of Computer Science, Multani Mal Modi College, Patiala, India*
[3]*Department of Chemistry, Multani Mal Modi College, Patiala, India*
*sukhsehajsingh@yahoo.com

ABSTRACT— Ocean environment is being explored for several purposes. It includes pipelines instalment, coral reef monitoring, habitat preservation, automated vehicles movement etc. However, degraded quality of underwater image necessitates its pre-processing and enhancement before being used in the software-driven environment. Due to dynamic underwater conditions, need of an hour is develop a method which according to the type and level of distortion, gives enhanced image corresponding to raw input image. This paper briefly discusses the emerging trends in the field of underwater image processing. Also, recently proposed techniques related to the field are highlighted with respect to their advantages and limitations. Despite of advances in the method of enhancement, authors are still facing challenges due to non-adaptability of techniques to the changing environment. Based on the thorough analysis and limitations of the existing study, we have mentioned certain future aspects for which research will be established.

KEYWORDS— underwater image, enhancement, restoration, colour correctness, dehazing, histogram equalization, fusion, convolutional neural network, generative adversarial network

INTRODUCTION

Low quality underwater images require processing for real-time applications [1]. Underwater imagery suffers from uneven conditions such as over/under saturation, bright spots, dark spots, non-uniform illumination etc. Fig. 1 represents the emerging fields in the area of underwater image processing.

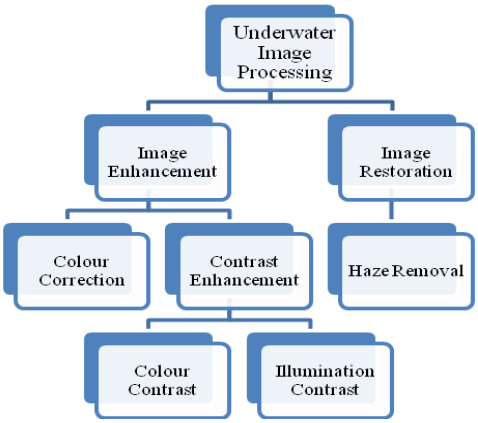


Fig. 6 Underwater Image Processing Fields

Underwater images due to harsh environmental conditions are distorted in nature, thence, image details are diminished [2]. Underwater scenes appear to be bluish or greenish even at smaller depths. Red colour due to long-wavelength attenuates faster with increase in distance (depth) followed by green and blue [3]. Light in the path from object to the camera is scattered by number of suspended particles in the water medium. Scattering causes loss in light intensity and it increases exponentially as we move inside water causing haze like effects [4] [5].

TRENDS IN UNDERWATER IMAGE PROCESSING

In recent years, underwater image processing has arisen as one of the popular fields. Low quality image with significant level of noise and distortion is encouraging researchers to come up with new ideas [6]. Colour as well as contrast correction are most prominent but challenging tasks of underwater image enhancement. Pixel based transformations are done for normal distribution of image pixels to entire intensity range of colour histogram (R, G and B colour space) [7]. Therefore, enhancement operations improve visualisation but structure and naturalness of the image is also supposed to be preserved. Histogram equalisation (HE) based methods for image enhancement suffers from artefacts such as under-enhancement, over-enhancement etc. These techniques enhance the image globally by re-distributing the image pixels and hence, small details like edges, texture are ignored. Rayleigh stretching and fusion based strategies have reduced these anomalies to some extent [8]. Rayleigh distribution restricts the histogram levels. It finds the solution of differential equation to find the min and max ranges of the histogram based on input image pixel representation. Fusion methods superimpose various pre-processed images. For example, multiple scale fusion of global followed by local histogram stretching for enhancement [9]. This single-image enhancement method processes required information form several intermediate steps to give resultant image. Image restoration methods reconstruct underwater image based on some prior

information. Dark channel Prior (DCP) assumes that intensity information about at least one colour channel is nearly lost due to scattering [10]. Restoration techniques find a solution to reverse (or inverse) problem using transmission estimation map. Convolutional neural network based techniques involve designing image restoration model by inputting to the system the raw image and corresponding ground truth. However, to construct a diverse database is difficult in this case due to continuously changing underwater environment. Therefore, generative adversarial (GAN) systems came into picture which generate synthetic underwater database using in-air colour images [11]. Database includes set of both restored image and degraded image with noise, attenuation. This type of processing becomes unrealistic in the sense that sometimes, it generates more than one solution to the problem which increases system complexity. More advantages, limitations for recent developments in underwater image processing have been discussed in the subsequent sections.

LITERATURE REVIEW

Yang et al. (2021) [12] proposed a model for underwater image processing based on hybrid of feature normalisation with recurring neural networks. For performance improvement, multiple activation maps are applied. Size of the various filters (kernels) which are used at the each activation layer for feature transformation is determined based on the difference in attenuation levels of various colour channels. Feature map at each layer is feed-forwarded to the subsequent layer in the network. Parameters deployed for evaluation of the technique are SSIM, PSNR. Datasets used for performance testing are UFO-120 [13] and EUVP [14] underwater imaging datasets. Technique performs well for depth estimation and computer vision related tasks. However, model is still required to be visualised for complex applications like image dehazing.

Liu & Liang (2021) [15] implemented another technique for underwater imaging refinement on the basis of attenuation slope. Prior information is generated based on the statistical inferences from this attenuation slope. Further, transformations are carried out based on the prior. For this, the colour pixel values are represented in RGB colour space followed by their distribution on the attenuation slope based on the varying attenuation values in changing underwater conditions. Colour channel casting is applied for initial processing and to calculate the attenuation difference. The technique uses the reverse mapping to evaluate the ground truth i.e. for computation of the areas in the input image with attenuation distortion. White balancing and guided image filtering is used for final processing. White balancing ensures that white colour in the resultant image renders actual white, not some other colour. Guided filter preserves the image characteristics i.e. edges, texture etc. Evaluation parameters are SSIM, PSNR and LOE. Technique works for noise filtering, colour correction and prevents over-enhancement. Limitation is that technique does not work in non-uniform illumination condition which needs to be addressed in future.

Gao et al. (2021) [16] proposed a technique for underwater image contrast enhancement based on multiple perspective fusion. Firstly, the image is reconstructed using red channel prior. Basically, the idea is to compensate that colour channel with higher levels of distortion. Red colour due to longer wavelength attenuates faster so, the details are nearly lost. Depth estimation maps are generated with respect to each colour channel to estimate the amount of light actually reaching the object without being scattered. Now, transformation (or restoration) is done for the channel containing pixels with lowest intensity levels. Here, two different variants are obtained for the transformed image from previous step. For first, local contrast enhancement is done for refinement of edges and other fine details. Laplacian (contrast) weight is evaluated to assure that edges possess high intensity values. Saliency and gradient weights are added in order to highlight edges and determine the region of interest (ROI). Image pixel based gradient weights are calculated using gradient magnitudes. Another version is obtained by post-processing with unsharp filtering or blurring. Both the variants now combined using multiple scale fusion. Evaluation parameters are UCIQE and UIQM. Degraded underwater images for experimentation are obtained from Ancuti [4] and Fattal [17] datasets. Limitations- technique cannot eliminate speckle (due to light scattering) noise. Also, technique has still to deal with such images that possess blurriness due to motion.

Li et al. (2021) [18] worked on a neural network based technique for the visual improvement of images captured underwater. The methodology applies an auto-encoder which accumulates the details from all colour channels followed by a decoder mechanism. Auto-encoder aims to extract/combine the most persistent characteristics (of different colour channels) required for image restoration. Auto-encoder basically works on the idea of feature selection or data reduction. Being unsupervised, it learns the abstract characterisation of input data to preserve only the task specific features. Decoder system is inspired from the depth map which signifies that how much illumination actually reaches the capturing device without being scattered. System implements the attention mechanism [19] in which decoder based on the abstract (but meaningful) description received form the auto-encoder, generates an output. Process is recurrent i.e. keeps on repeating for multiple units in time. Mechanism helps decoder network to focus on the most relevant sequences in the image for enhanced output. Model shows visual improvement for images distorted due to scattering. The datasets used for experimentation are underwater benchmarking dataset UIEB [20], SQUID [21]. Evaluation parameters are MSE, PSNR, UIQM and UCIQE. Limitation- technique does not perform in low or limited illumination conditions.

Huang et al. (2021) [22] proposed a multi-layer model for underwater image processing based on convolutional network. For colour enhancement, multiple convolution layers are assembled together for data transformation (high-to-low dimension reduction). Data reduction aims to improve system utilisation by avoiding computational complexities. It reduces the feature space for the raw dataset while considering only relevant features required for processing. However, a necessary trade-off has to be maintained between dimension reduction, diversity of dataset and minimum information loss. Residual connections are positioned in order to induce depth in the network. Residual network facilitates image reconstruction by mapping to those areas in the raw image with significant distortions. Intra attention mechanism focuses

on the task-specific homogeneous pixels or regions in the image. It helps correlating to the defected regions in the image for which correction is required. Finally, LBP operator is applied to carry out illumination-invariant transformation to the image. It assigns a binary value to the neighbourhood pixels considering value of centre pixel as threshold. Operations are carried out on image patches with often size of 3 X 3. Both colour corrected image and one obtained by applying LBP are fused together (pixel by pixel) to obtain resultant image. Parameters for evaluation are PSNR, SSIM, UCIQE and UIQM. Technique shows better result for colour and contrast enhancement. Limitations: generated result is hazy and under-enhanced for the image containing far away backgrounds. In certain cases, colour intensity of the resultant image is significantly low.

CONCLUSION AND FUTURE SCOPE

Although, Deep Neural Networks such as CNN and GAN are being deployed for underwater image processing but such approaches multiply system complexity. Moreover, amount of data available for degraded underwater images and corresponding ground truths is not sufficient enough that it could depict the reality of real-world scenario. Despite being feature intrinsic, algorithm should have broad scope and applicability. Also, it should be time and memory efficient. Single image dehazing methods under certain circumstances can be proved more effective provided they should be intelligent so that processing is done according to the specificity and amount of distortion.

REFERENCES

[11] J. Han, M. Shoeiby, T. Malthus, E. Botha, J. Anstee, S. Anwar, R. Wei, L. Petersson, and M. A. Armin, "Single underwater image restoration by contrastive learning," *arXiv preprint arXiv:2103.09697v2*, Apr. 2021.

[12] K. Panetta, C. Gao, and S. Agaian, "Human-Visual-System-Inspired Underwater Image Quality Measures," *IEEE Journal of Oceanic Eng.*, vol. 41, no. 3, pp. 541-551, Oct. 2015.

[13] S. S. Singh, R. Sachdeva, and R. Sharma, "An Integrated Technique For Underwater Image Enhancement: Colour Correction And Dehazing," *Advances in Mathematics: Scientific J.*, vol. 9, no. 6, pp. 3865-3877, Jul. 2020.

[14] C. O. Ancuti, C. Ancuti, C. D. Vleeschouwer, and P. Bekaert, "Color Balance and Fusion for Underwater Image Enhancement," *IEEE Transactions on Image Process.*, vol. 27, no. 1, pp. 379-393, Oct. 2017.

[15] S. S. Singh, R. Sachdeva, and A. Singh, "An Optimized Approach for Underwater Image Dehazing and Colour Correction," in *Proc.* ICICC, 2020.

[16] M. Han, Z. Lyu, T. Qiu, and M. Xu, "A Review on Intelligence Dehazing and Color Restoration for Underwater Images," *IEEE Transactions on Systems, Man, and Cybernetics: Syst.*, vol. 50, no. 5, pp. 1820-1832, Jan. 2018.

[17] K. Iqbal, M. Odetayo, A. James, R. A. Salam, and A. Z. H. Talib, "Enhancing the low quality images using Unsupervised Colour Correction Method," in *Proc. IEEE ICSMC*, 2010, pp. 1703-1709.

[18] A. S. A. Ghani, and N. A. M. Isa, "Underwater image quality enhancement through integrated color model with Rayleigh distribution," *Applied Soft Comput.*, vol. 27, pp. 219-230, Feb. 2015.

[19] A. S. A. Ghani, and N. A. M. Isa, "Enhancement of low quality underwater image through integrated global and local contrast correction," *Applied Soft Comput.*, vol. 37, pp. 332-344, Dec. 2015.

[20] H. Liu, and L. P. Chau, "Underwater image restoration based on contrast enhancement," in *Proc. IEEE ICDSP*, 2016, pp. 584-588.

[21] J. Li, K. A. Skinner, R. M. Eustice, and M. Johnson-Roberson, "WaterGAN: Unsupervised Generative Network to Enable Real-Time Color Correction of Monocular Underwater Images," *IEEE Robotics and Automation Lett.*, vol. 3, no. 1, pp. 387-394, Jul. 2017.

[22] H. H. Yang, K. C. Huang, and W. T. Chen, "Laffnet: A Lightweight Adaptive Feature Fusion Network for Underwater Image Enhancement," *arXiv preprint arXiv:2105.01299v2*, May 2021.

[23] M. J. Islam, P. Luo, and J. Sattar, "Simultaneous Enhancement and Super-Resolution of Underwater Imagery for Improved Visual Perception," *arXiv preprint arXiv:2002.01155*, Feb. 2020.

[24] M. J. Islam, Y. Xia, and J. Sattar, "Fast Underwater Image Enhancement for Improved Visual Perception," *IEEE Robotics and Automation Lett.*, vol. 5, no. 2, pp. 3227-3234, Apr. 2020.

[25] K. Liu, and Y. Liang, "Underwater image enhancement method based on adaptive attenuation-curve prior," *Optics Exp.*, vol. 29, no. 7, pp. 10321-10345, Mar. 2021.

[26] F. Gao, K. Wang, Z. Yang, Y. Wang, and Q. Zhang, "Underwater Image Enhancement Based on Local Contrast Correction and Multi-Scale Fusion," *Journal of Marine Science and Eng.*, vol. 9, no. 2, pp. 1-16, Feb. 2021.

[27] R. Fattal, "Dehazing using color-lines," *ACM Transactions on Graph.*, vol. 34, no. 1, pp. 1-14, Nov. 2014.

[28] C. Li, S. Anwar, J. Hou, R. Cong, C. Guo, and W. Ren, "Underwater Image Enhancement via Medium Transmission-Guided Multi-Color Space Embedding," *IEEE Transactions on Image Process.*, vol. 30, pp. 4985-5000, May 2021.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Proc. NIPS 2017*, 2017, pp. 5998-6008.

[30] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, art. 107038, pp. 1-11, Feb. 2020.

[31] D. Berman, D. Levy, S. Avidan, and T. Treibitz, "Underwater Single Image Color Restoration Using Haze-Lines and a New Quantitative Dataset," *IEEE Transactions on Pattern Analysis and Machine Intell.*, vol. 43, no. 8, pp. 2822-2837, Mar. 2021

[32] Z. Huang, J. Li, and Z. Hua, "Underwater image enhancement via LBP-based attention residual network," *IET Image Process.*, pp. 1-18, Sept. 2021.

ABBREVIATIONS

HE :	Histogram Equalisation
DCP :	Dark channel Prior
GAN :	Generative Adversarial Network
CNN :	Convolutional Neural Network
SSIM :	Structural Similarity Index Measure
PSNR:	Peak Signal-to-Noise Ratio
UCIQE:	Underwater Colour Image Quality Evaluation
UIQM :	Underwater Image Quality Measure
MSE :	Mean-Squared Error
LBP :	Local Binary Patterns

PROPOSED COVID-19 TESTING PROCESS USING MACHINE LEARNING TECHNIQUE

Chirag Bansal[#1], Brahmaleen Sidhu[*2]
[#]*Department of Computer Science and Engineering, Punjabi University Patiala*
[1]chiragbansal254@gmail.com
[2]brahmaleen.sidhu@gmail.com

ABSTRACT— Covid is a transmissible disease which was caused by SARS-CoV-2 Virus. It causes an infection to our nose or upper throat. Early it was breakout only in China but after that in 2020 a new strain of COVID outbreak quickly spread all over the world. This disease was official named by WHO in February 11, 2020. Because of its highly transmissible nature it is very difficult to detect and daily count of death was increasing day by day. So, to detect covid various types of techniques are used but the drawback of all these methods is that they will not give us instant result, if they give those all devices or methods are costly. Covid is test but three different methods 1) Molecular (RT-PCR) Test 2) COVID-19 Antigen Test 3) COVID-19 Antibody Tests but all these tests take time, so if a person is COVID positive it will transmit it till he will not get result. So, we proposed a Machine learning based technique which tell us how we can get the result instantly.

KEYWORDS— COVID, SARS-CoV-2, Machine Learning, Cloud Computing, Breathalyzer, IoT, Sensors, Breath

INTRODUCTION

The novel coronavirus 2019 (COVID-19) pandemic caused by the SARS-CoV-2 continues to form a vital and imperative threat to international health. The happening in early Dec 2019 at intervals the Hubei province of the People's Republic of China has unrolled worldwide. As per August 2021, count of patients confirmed to have the sickness has exceeded 215,000,000[1], in additional than one hundred eighty countries, tho' the amount individuals infected is probably copious higher. over 4,550,000 people have died from COVID-19.

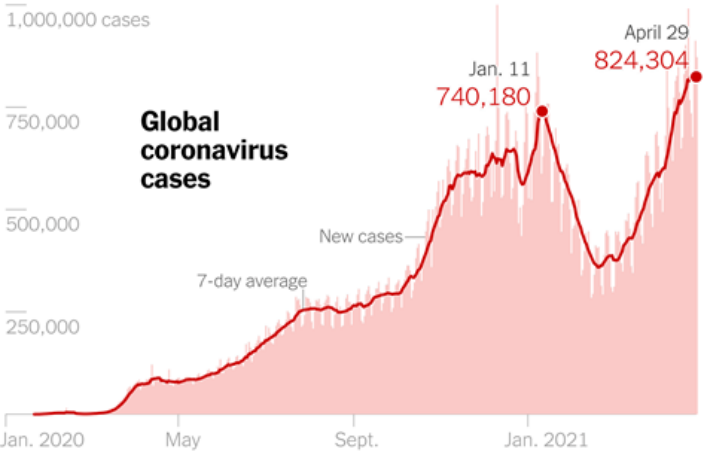


Fig. 1 Global Coronavirus Cases

This pandemic continues to challenge medical systems round the world in many ways, alongside a speedily increasing demand for hospital beds and a big shortage of medical instrumentation as several care staff became infected. the employment of health resources is essential. the foremost valid assay for COVID19, victimization reverse transcription–polymerase chain reaction (RTPCR) has long been in brief offer in developing countries, contributory to hyperbolic infection rates and delaying important preventive measures. Many models were created on the bases of computer tomography (CT) scans, clinical symptoms, laboratory tests but the accuracy of these models is not that good and it take time to take CT scan and all other tests.

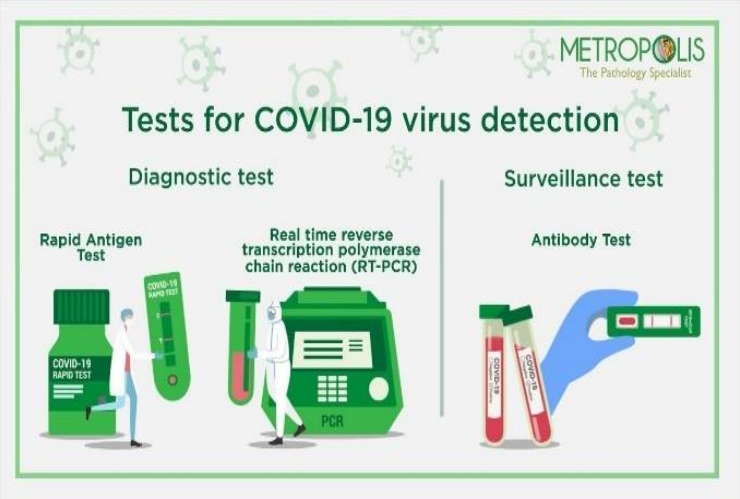


Fig. 2 COVID-19 Currently Testing Techniques

In this paper, we are going to propose a machine-learning approach that predicts a positive SARS-CoV-2 infection based on Breathalyzer concept which work on the exhaled air and some other physical symptoms. The model will be trained on the exhaled air values and other physical symptoms values. Thus, our approach will may be implemented all over the world for effective screening and prioritization of testing for Corona virus in the general population.

LITERATURE REVIEW

There are a lot of innovations has been planned for testing of covid Some are:

Benji Shan et al. (2020) [2] use the IoT based machine learning approach is used to detect the COVID in the patients. They create their own dataset of 49 confirmed cases along with 58 healthy and 33 Lung infection. All over accuracy of that model is 88.75%. They use different sensors to detect COVID and analyse the breath of the patient. Device use VOC which was creating a diver sensing layer that can be swell or shrink. Then VOC is diffused into the sensing layer. It was a multiplexed hybrid sensor array based on clever nanomaterials for detecting and monitoring COVID-19-specific VOC combinations from exhaled breath.

Yazeed Zoabi et al. (2020) [3] use machine learning based prediction of COVID-19 based on symptoms in which they prepare a set of questions on the bases of which the conclusion of COVID will be provided. The dataset they use consist of total 99,232 out of which 90,839 COVID-19 negative and 8,393 are COVID-19 positive. The limitation is this method is that the symptoms that they use may also be associate with other disease.

Biswadev Mitra et al. (2020) [4] They use fever as their key point to detect COVID-19. They do fever screening overseas and meta-analysis the percentage of patients who got admitted in hospital and have positive COVID result. They do a cohort study to all patients who were admitted in the hospital and return positive for covid. As their primary outcome is fever for positive covid test result. Data extract extracted for the outcome variables of body temperature at the time of testing and when repeated, the highest temperature within the next24 h. Age, sex and mode of tempera-true measurement were also extracted. Results were reported using proportions with 95%. Data contain around 65,000 patients.

Winichakoon et al [5] published a letter to the deskman that represent a study of a COVID-19 patient who tested negative for COVID-19 on a nasopharygeal/oropharyngeal RT-PCR swab but positive on RT-PCR of BAL fluid. Another short research looked at 19 instances of individuals who were suspected of having COVID-19. All 19 patients had oropharyngeal RT-PCR swab tests, although only nine of them were positive (47.4 percent). [14]

Xie et al [6], Five patients from China's Hunan region showed hazy gray areas (GGO) on chest computed tomography (CT) that gives signs of COVID-19, but early pharyngeal RT-PCR testing reports were negative. Whereas, Repeat RT-PCR swab testing were found to be positive. Similarly, Fang et al [14] examine fifty-one patients who were eventually confirmed to have COVID19 and who had a chest CT scan and an RTPCR scan through a throat swab (45 patients) or a mucus secretion (six patients) at the time of admission to the hospital were carried out. Of these 51 patients, chest CT scans showed characteristic COVID19 findings in fifty (98%). By comparison, 36 out of 51 (71%) were positive on the initial RTPCR test.

Zou et al,[12] The virus masses in the upper tract of 18 patients varied depending on the subsite, with the infectious agent load in the cavity being 64 times higher than in the tubular cavity. In a study of 213 confirmed COVID19 patients, the authors found that mucosal secretions showed the best positive rate in each severe (88.9%) and mild (82.2%) case, followed by nasal swabs (73.3%, 72.1%) and then throat swab (60.0%, 61.3%).[13]

Further research has additionally proven that preliminary RTPCR assessments may be poor and consequently nice despite habitual assessments. Shanghai Dialect et al. [15] the scientific direction of eighty sufferers from Jiangsu Province who have been in the end recognized with COVID19. nine of those eighty sufferers (11.3%) had 2 RTPCR-poor nasal or mouth

swabs earlier than the 1/3 swab become nice. Additionally, Young et al[16] reported the results of the daily Cavum RTPCR tests performed on eighteen Singapore patients suffered for COVID 19. Interestingly, a few sufferers had wonderful checks, then poor checks, after which wonderful checks again, all inner a comparable hospitalization.

RECENT METHOD

The sensitivity and specificity of the victimization of RTPCR body cavity swabs to identify COVID19 cannot be precisely determined from the information disclosed so far. Specificity. On the contrary, the sensitivity is moderate (perhaps between 63-78%). Throat swabs seem to have intensity sensitivity in many of the diverse varieties of RTPCR activity; Nasal swabs also are barely greater sensitive than throat swabs. RTPCR evaluation of BAL fluid seems to be the maximum accurate manner of virological confirmation, however, BAL fluid is the handiest accrued pretty from the sickest cohort of sufferers. In sufferers with mild to excessive COVID19 symptoms, the feature chest CT experiment findings also are extra touchy than RTPCR tests.

Fig. 3 RTPCR testing method

Given these results, a negative check does not rule out the disease as soon as a patient has a high probability of protesting against COVID19. As a result, guidelines that require high precision for RTPCR testing are dangerous. For example, employers shouldn't use a negative. The test result to help you make a decision about when to return to work. Meanwhile, the perceived desire for accumulated evidence circulated by the preferred media [7] could result in some patients turning to an associated additional erectile dysfunction test solely, putting these people at higher risk for COVID19 could if it hasn't already. Since treatment is not required for sensitive COVID19 cases, patients with mild symptoms do not need to go to the emergency room or get tested; instead, they should be quarantined.

Computed tomography of the chest of COVID19 patients generally shows frosted glass opacities, multifocal uneven consolidation, and / or aperture changes with a peripheral distribution [8]. In one study, the authors found that chest CT scans had a higher sensitivity for the COVID19 designation (88%). as hostile throat swabs from the initial RTPCR victimization (59%) [9]. Another study looked at the ability of radiologists to differentiate COVID19 respiratory disease from non-COVID19 pneumonia. Sensitivities among radiologists were between 73% 93% and specificities between 93% 100%. [10] Currently, most radiological societies do not advocate routine screening for COVID19 with chest CT. [11]

PROPOSED APPROACH

We have an idea to implement but due to lack of data resources we are not able take it in the real world. We want to use IoT base a Machine Learning Technique to test Covid-19 patient. In this approach we will use various sensors to analyse the human being breath all the sensors that we use are human breath sensitive. When a patient blows up the air in the device then all the parameter values will be separate to be feed in the machine learning model. The model that we will create take all the sensors value along with some manual added values which helps us to get better results. The result that we get will be more accurate and we get result in few seconds or minutes. All the testing parameters will be store on a cloud server so that after get a fair number of parameters then the model will be retrained to make it better and efficient.

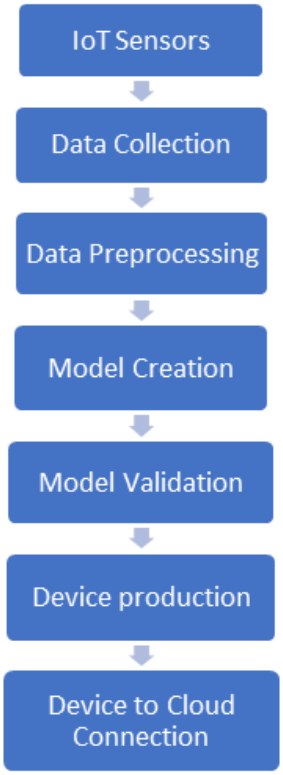


Fig. 4 Detailed Chart of Proposed Techniques

CONCLUSIONS

Due to the increase in the COVID pandemic and rise in the number of cases. A new and fast method of testing will be used. The method that I proposed will be an idea for fast testing. I approach that I want to choose is Machine Learning and in the field of medical science machine learning is a trendy topic and it helps to do various impossible tasks. This approach will help us to reduce covid cases. IoT sensors will also reduce the cost of the device so, it will be available for everyone at cheap cost and anybody will afford it.

Also, we will change the algorithmic approach to get a better result and optimize the approach. We can also add more disease which we can detect with breathalyzer and help to reduce the cost of testing and also reduce the death rate.

REFERENCES

[1] WHO official website for cases count, covid19.who.int/.
[2] B. Shan, Y. Broza, and W. Li contributed equally to the work. “Multiplexed Nanomaterial-Based Sensor Array for Detection of COVID-19 in Exhaled Breath,” ASC Nano, 2020, 14, 9, 12125–12132
[3] Zoabi, Y., Deri-Rozov, S. & Shomron, N. “Machine learning-based prediction of COVID-19 diagnosis based on symptoms.” npj Digit. Med. 4, 3 (2021). https://doi.org/10.1038/s41746-020-00372-6
[4] Mitra, B., Luckhoff, C., Mitchell, R.D., O'Reilly, G.M., Smit, D.V. and Cameron, P.A. (2020), “Temperature screening has negligible value for control of COVID-19.” Emergency Medicine Australasia, 32: 867-869. https://doi.org/10.1111/1742-6723.13578
[5] Winichakoon P, Chaiwarith R, Liwsrisakun C, et al. Negative nasopharyngeal and oropharyngeal swab does not rule out COVID-19. J Clin Microbiol. 2020 In press.
[6] Xie X, Zhong Z, Zhao W, et al. Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing. Radiology. 2020 200343.Siegler K. Many who need testing for COVID-19 fail to get access. [Accessed April 5, 2020]. Available at: https://www.npr.org/2020/04/03/826044608/many-who-need-testing-for-covid-19-fail-to-get-access.
[7] Chung M, Bernheim A, Mei X, et al. CT imaging features of 2019 novel coronavirus (2019-nCoV). Radiology. 2020; 295: 202- 207.
[8] Ai T, Yang Z, Hou H, et al. Correlation of chest CT and RT-PCR testing in coronavirus disease 2019 (COVID-19) in China: a report of 1014 cases. Radiology. 2020;200642. https://doi.org/10.1148/radiol.2020200642
[9] Bai HX, Hsieh B, Xiong Z, et al. Performance of radiologists in differentiating COVID-19 from viral pneumonia on chest CT. Radiology. 2020;200823. https://doi.org/10.1148/radiol.2020200823

[10] Simpson S, Kay FU, Abbara S, et al. Radiological Society of North America Expert Consensus Statement on Reporting Chest CT Findings Related to COVID-19. Endorsed by the Society of Thoracic Radiology, the American College of Radiology, and RSNA. Radiol Cardiothorac Imaging. 2020; 2: e200152.

[11] Zou L, Ruan F, Huang M, et al. SARS-CoV-2 viral load in upper respiratory specimens of infected patients. N Engl J Med. 2020; 382: 1177- 1179.

[12] Y ang Y, Yang M, Shen C, et al. Evaluating the accuracy of different respiratory specimens in the laboratory diagnosis and monitoring the viral shedding of 2019-nCoV infections. medRxiv. 2020. https://doi.org/10.1101/2020.02.11.20021493

[13] Xie C, Jiang L, Huang G, et al. Comparison of different samples for 2019 novel coronavirus detection by nucleic acid amplification tests. Int J Infect Dis. 2020; 93:264–7.

[14] Fang Y, Zhang H, Xie J, et al. Sensitivity of chest CT for COVID-19: comparison to RT-PCR. Radiology. 2020 200432.

[15] Wu J, Liu J, Zhao X, et al. Clinical characteristics of imported cases of COVID-19 in Jiangsu Province: a multicenter descriptive study. *Clin Infect Dis.* 2020 In Press.

[16] Young BE, Ong SWX, Kalimuddin S, et al. Epidemiologic features and clinical course of patients infected with SARS-CoV-2 in Singapore. *JAMA.* 2020 In Press.

NATURAL LANGUAGE PROCESSING – A REVIEW

Navdeep Singh
Computer Science & Engineering
Punjabi University, Patiala, India
navdeepsony@gmail.com

ABSTRACT — Natural language processing belongs to the field of Computer Science, Linguistics and Artificial Intelligence that discusses the interaction that happens between human beings and computers through natural language. It specially deals with how to process huge amounts of natural data so as to understand it, decode it and extract useful information out of it. NLP is not new and is in fact very old but recently there is a huge rise in its applicability. A long range of works can be performed using NLP beginning with machine translation to tagging different parts of a speech apart from parsing the sentences. Present NLP research work shows that the researchers are showing more interest in unsupervised as well as deep learning based techniques. The unsupervised techniques learn from the data itself which is not annotated manually whereas deep learning based techniques need manually annotated data in order to make sense from the natural language. In this paper, various natural language processing fields and corresponding techniques have been discussed.

KEYWORDS — Data Mining, Text Mining, NLP, Natural Language Processing, Tokenization, Lemmatization

I. INTRODUCTION

The main goal of natural language processing is to create a system that can decipher, understand, synthesize and work with the natural languages used by humans. It is the sub-domain of artificial intelligence and is concerned with the creation of meaningful expressions for the evaluation of natural languages. A detailed understanding of the context makes the data more meaningful which subsequently helps in mining and the analysis of the text. Earlier hand written rules were used for the assessment of the learning procedures which are very time consuming and are often not complete. These were replaced by machine learning techniques but recently deep learning techniques have also seen a rapid rise due to their increased accuracy and also due to the increased availability of the computational power and memory. These new developments made a shift in the paradigm towards new data driven techniques from the traditional techniques thus advancing the field of natural language processing. The advantage associated with these techniques is the robustness they possess as they give a fairly accurate output even with the erroneous data. In [1], it has been shown that deep learning techniques outperforms most of the state of the art techniques especially when working with POS tagging, named entity recognition (NER) and semantic role labeling (SRL).

II. TECHNIQUES

Q. A. Text Mining

R. It is the process of extracting relevant and important information from the text that can be further used for the analysis of the data [2]. In it, unstructured textual data is converted into actionable and meaningful data. The motivation for attempting to extract information from such material automatically is compelling even if success is just partial and text mining frequently deals with writings whose function is the conveying of actual information or opinions. It helps the companies make important decisions based on the valuable insights obtained from the data. The main advantage associated with text mining is that the transformed data can be automatically understood by the machines based on the sentiment, text, intent and topic. It is only through text mining that the machines can now process huge amounts of data very effectively and efficiently in comparatively lesser time. Automated text analysis happens only due to the fine combination of machine learning and text mining. The text mining process consists of the following steps.

S. Tokenization – In this step, the word characters are converted into the tokens of variable sizes. A Bag-of-words is obtained using the document term matrix which is constructed from the tokens.If the goal is to group or categorise documents, word frequency alone may not be sufficient [3].

T. Text preprocessing – This step deals with the removal of unwanted words from the data such as punctuation marks, phrases and letters which improves the accuracy of the model.

U. Text Transposition – It deals with the process of text representation and text selection which are needed for the analysis of the data.

V. Attribute Selection – This step is associated mainly with reducing the dimensionality of the data by removing unwanted attributes which helps in reducing the computational complexity of the model.

W. Data Mining – In data mining, various kinds of data categorization is done.

X. Evaluation – The extracted data is deciphered and analyzed in the evaluation step so as to provide meaningful data to the machines.

Y. B. Machine Translation

Z. It is the process of translating the text extracted from the source language into the text of the target language [4]. The process of translation is very challenging due to the fact that there exist large number of languages with each having its own punctuation, grammar and the other reason is that it is very complex for a machine to work with sequences rather than working with the numbers thus making the task even more harder. Machine translation techniques are of three types namely, Statistical machine translation (SMT), Rule-based machine translation (RMT) and Neural machine translation (NMT) [5]. Statistical machine translation uses various statistical models for the translation of the text whereas NMT uses a neural network to perform the same task. Rule-based machine translation depends upon various hand crafted rules for the translation purposes and is thus slower than the other two and requires more effort. Recently, BART, an autoencoder which trains sequence models was proposed in [6] for pretraining sequence-to-sequence models. BART is learned by obfuscating text using an arbitrary noise function and then building a model to recover the original text. When fine-tuned for text production, BART is especially successful, but it also performs well for comprehension tasks. Another model, GPT, that performs machine translation was proposed in [7]. The drawback associated with this model is that it models only leftward context that is troublesome for certain tasks in contrast to BERT which address these limitations as it is a bidirectional encoder. Chronopoulou et al. [8] proposed another model which maps a sequence of tokens to a sequence of missing tokens. Because the encoder and decoder are given separate groups of tokens, it is less effective for discriminative tasks.

AA. C. Morphological Analysis

Morphological analysis deals with the formation and structure of the words present in a statement [9]. The most basic and important unit in this analysis is called 'morphine'. Morphology helps in information retrieval, language modeling and machine translation and is therefore an integral part of natural language processing. The need of morphological analysis arises from the fact that many language dependent applications need the information present in the words such as the information retrieval systems need to know in advance about the stem of the word. In morphological analysis, the information is extracted from each and every word present in the sentence and then encoded to create a meaningful data which is then used by the later layers for processing. There are three main steps to morphological analysis and they are; Inflection, Derivation and Compounding [10]. Inflection deals with the process of conjugation and declination. The main benefit associated with it is that it does not change the part of the speech (POS) while working with the sequence of the words. Derivation refers to the generation of new words but it changes the parts of speech (POS) whereas Composition refers to the process of combining two or more base words in order to create a new word. A model for morphological analysis was proposed in [11] to create a regular expression for each word to automatically extract inflectionals.

BB. D. Natural Language Generation

Natural language generation is the process of creating senseful and meaningful sentences and phrases in the natural language. Narratives are automatically created in it at a very high speed of thousand pages per second to describe the input structured data in human language form. The important aspect associated with natural language generation is that while it can write data, it cannot read [12]. Natural language understanding (NLU) on the other hand performs this function and it converts the unstructured data into the structured data which can be understood by the machines. There are six stages of natural language generation. They are:

Content Determination - It deals with the information decision present in the text.

Document Structuring – It refers to organizing the information in a structured way.

Aggregation – It improves readability by merging the similar sentences with each other.

Lexical Analysis – Creating concepts from the words present in the sentences.

Expression generation – It refers to the creation of the expression that can distinguish between regions and objects.

Realization – It deals with the generation of the final text which is correct in nature and follows all rules of natural language such as morphology, orthography and syntax.

Various advanced NLG models in today's demand are Markov chain model, Recurrent neural networks, Long short-term memory (LSTM) and transformers. UNILM was proposed in [13] which has a special property that it can adjust to different downstream tasks unlike BERT. There are various advantages of using UNILM, the first one being, the unified training method results in a single Transformer LM that employs shared parameters and architecture for many types of LMs, eliminating the need to train and host numerous LM's individually. Second, parameter sharing makes learned text representations more generic since they are jointly tailored for diverse language learning objectives that employ context in different ways, reducing overfitting to a particular LM task. To mislead well-trained sentiment analysis and textual entailment models, a population-based optimization technique used to produce semantically and syntactically comparable adversarial samples was proposed in [14].

CC. E. Optical Character Recognition

It is the process of converting the images of handwritten or typed text from a photo or scanned document into machine encoded text [15]. It is a common technique to produce digital text which can later be searched, edited, stored and used by machines for machine translation, text mining and text to speech tasks. The accuracy of OCR techniques depends on the quality of images from which digital images are to be produced. A two stage model was proposed in [16] for optical

character recognition. In the first phase, we look for rectangular areas in the image that could contain text. In the second phase, we do text recognition, in which a CNN is utilised to recognise and transcribe the word in each of the identified areas. This two-step method offers numerous advantages, including the capability to isolate the training process from model deployment changes, execute word recognition in parallel, and enable text recognition for multiple languages independently. A novel approach for correcting document pictures with different sorts of distortions from a single input image was proposed in [17]. Rather than learning the full image, the method focuses on learning the distortion flow on the input image patches.The patch results are then stitched to a rectified document by processing. Uneven illumination is then corrected to further improve the readability and increase the OCR accuracy. This technique drastically improves the accuracy in both the real and synthetic images.

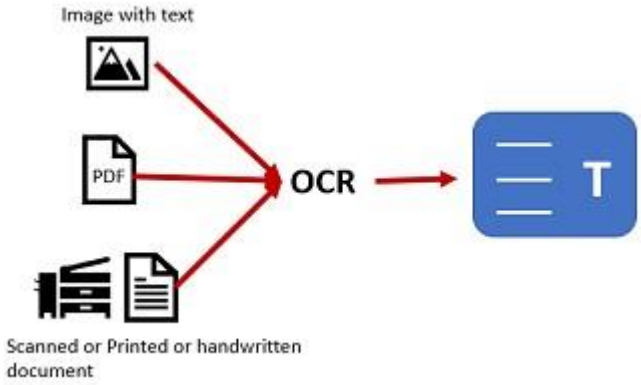


Fig. 1. Optical Character Recognition

Optical recognition techniques are of two types [18]:

Matrix matching: In it a pixel by pixel matching of the stored image and the glyph is done. This is the best technique for handwritten textual images especially with known fonts as its performance degrades when it encounters new fonts. Matrix matching is also referred to by other names such as pattern matching and image correlation.

Feature extraction: In this step, various features such as lines, line intersections and loops are generated from the stored glyphs. The features are then compared with the vectors containing characters which may further reduce to a lesser number of glyphs.

DD. F. Sentiment Analysis

Sentiment analysis deals with the mining of the subjective information that is extracted from the source from which sentiment monitoring can be performed [19]. Sentiment analysis, also called opinion mining, is a systematic process to recognize, segment, and quantify the subjective information present in the statements. The TF-IDR technique, which works by translating words into numbers and is computed using the term frequency-inverse document frequency method, is commonly used in sentiment analysis. The most basic task in sentiment analysis is determining the polarity of the text in the document or the image such as whether the feature is positive or negative which can sometimes be natural [20]. An overview of sentiment analysis is shown in Figure 2 [21].

The primary advantage associated with sentiment analysis is that it can be automated and decisions can be made based on the big data available to the machines rather than making decisions based on the intuition which is not always correct. When implementing sentiment analysis, the use of either a Lexiconbased technique, a Machine Learning method, or a combination of both methods are used. Unsupervised learning is a term used to describe a process based on a lexicon. The Lexicon approach relies just on the dictionary and does not require any training data. There are two types of sentiment analysis [22]:

Lexicon analysis attempts to determine a document's polarity based on the sentiment polarity of its words or phrases. However, solutions based on lexical analysis do not take into account the context of the study.

Machine Learning, includes creating models using a labelled training dataset (text or sentence occurrences) in order to ascertain a document's orientation. Studies utilising these approaches have been conducted on a specific issue. Machine learning is categorised as supervised learning, and it requires training data to be processed.Using machine learning to analyse data is time consuming, as it takes hours in a in training which is necessary to make the model learn about the data. Both the methods provide comparable performance but their combination works well and even improves the management of unstructured data [23].

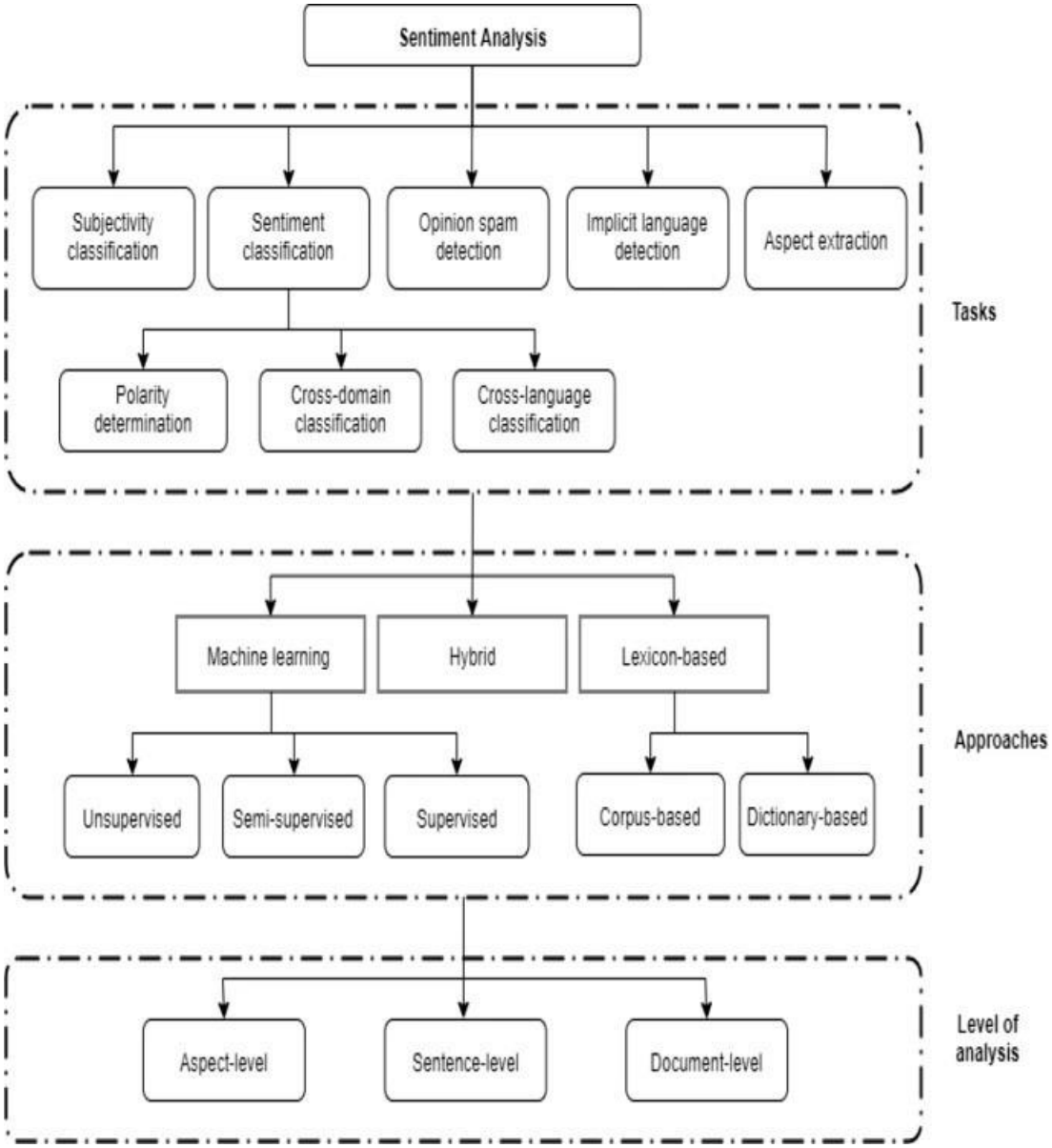


Fig. 2. Semantic Analysis

A comparative analysis of various semantic analysis techniques has been provided in Table 1 [24].

TABLE I
COMPARATIVE ANALYSIS OF SEMANTIC ANALYSIS TECHNIQUES

Technique	Precision	Recall	F1-Score
Logistic Regression + n-gram	0.729	0.779	0.753
SVM + TF-IDF	0.816	0.816	0.816
GBDT + TF-IDF	0.819	0.807	0.813
SVM + BOW	0.791	0.788	0.789
GBDT + BOW	0.800	0.802	0.801

EE. G. Challenges and Future Directions

There are a wide range of challenges which provide us various future directions. The first and foremost is that, not all aspects of language have been covered due to the complexity of real languages, especially when it comes to slang, negation and satire. There is a huge scope of work that can be done on contextual words as same words and phrases can have different meaning and differentiating between them is very challenging. Very less work has been done in the field of lexical, semantic and syntactic ambiguity and these fields present a large scope for future work. Colloquialisms or informal phrases presents a big problem to the NLP models as the models are trained on formal data and as such a lot of work can be done in this direction too.

III. CONCLUSION

Natural language processing is a sub-domain of artificial intelligence and it uses mined data for interaction between humans and machines. NLP is primarily concerned with deciphering and analysing the human language that we normally speak, and then converts it into a language that only the computer understands. Computational linguistics and rule-based human language modeling is coupled with statistical, machine learning, and deep learning models in NLP. Automatic summarization, machine translation, narrative analysis, conference resolution, voice recognition, and other NLP activities are among the most important tasks that are performed using natural language processing.

REFERENCES

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural Language Processing (almost) from Scratch," *J. Mach. Learn. Res.*, Mar. 2011, [Online]. Available: http://arxiv.org/abs/1103.0398.

[2] R. Feldman and J. Sanger, *The Text Mining Handbook*. Cambridge: Cambridge University Press, 2006.

[3] V. B. Kobayashi, S. T. Mol, H. A. Berkers, G. Kismihók, and D. N. Den Hartog, "Text Classification for Organizational Researchers," *Organ. Res. Methods*, vol. 21, no. 3, pp. 766–799, Jul. 2018, doi: 10.1177/1094428117719322.

[4] M. Johnson *et al.*, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 339–351, Dec. 2017, doi: 10.1162/tacl_a_00065.

[5] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. Rush, "OpenNMT: Open-Source Toolkit for Neural Machine Translation," in *Proceedings of ACL 2017, System Demonstrations*, 2017, pp. 67–72, doi: 10.18653/v1/P17-4012.

[6] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," Oct. 2019, doi: 10.18653/v1/2020.acl-main.703.

[7] A. Radfort, K. Narasimhan, T. Salimans, and I. Sutskever, "(OpenAI Transformer): Improving Language Understanding by Generative Pre-Training," *OpenAI*, 2018.

[8] A. Chronopoulou, D. Stojanovski, and A. Fraser, "Improving the Lexical Ability of Pretrained Language Models for Unsupervised Neural Machine Translation," Mar. 2021, doi: 10.18653/v1/2021.naacl-main.16.

[9] T. Ritchey, "Wicked Problems. Modelling social messes with morphological analysis," *Acta Morphol. Gen.*, 2013.

[10] T. Ritcheyy, "General Morphological Analysis: A general method for non-quantified modelling," 1998.

[11] A. Anastasopoulos, C. Cox, G. Neubig, and H. Cruz, "Endangered Languages meet Modern NLP," in *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*, 2020, pp. 39–45, doi: 10.18653/v1/2020.coling-tutorials.7.

[12] R. Perera and P. Nand, "Recent Advances in Natural Language Generation: A Survey and Classification of the Empirical Literature," *Comput. Informatics*, vol. 36, no. 1, pp. 1–32, 2017, doi: 10.4149/cai_2017_1_1.

[13] L. Dong *et al.*, "Unified language model pre-training for natural language understanding and generation," 2019.

[14] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, and K.-W. Chang, "Generating Natural Language Adversarial Examples," *Proc. 2018 Conf. Empir. Methods Nat. Lang. Process. EMNLP 2018*, Apr. 2018, doi: 10.18653/v1/d18-1316.

[15] S. Milyaev, O. Barinova, T. Novikova, P. Kohli, and V. Lempitsky, "Image Binarization for End-to-End Text Understanding in Natural Images," in *2013 12th International Conference on Document Analysis and Recognition*, Aug. 2013, pp. 128–132, doi: 10.1109/ICDAR.2013.33.

[16] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Oct. 2019, doi: 10.1145/3219819.3219861.

[17] X. Li, B. Zhang, J. Liao, and P. V. Sander, "Document Rectification and Illumination Correction using a Patch-based CNN," *ACM Trans. Graph.*, Sep. 2019, doi: 10.1145/3355089.3356563.

[18] M. R. Gupta, N. P. Jacobson, and E. K. Garcia, "OCR binarization and image pre-processing for searching historical documents," *Pattern Recognit.*, vol. 40, no. 2, pp. 389–397, Feb. 2007, doi: 10.1016/j.patcog.2006.04.043.

[19] M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," 2006, doi: 10.1111/j.1467-8640.2006.00276.x.

[20] P. Sasikala and L. Mary Immaculate Sheela, "Sentiment analysis of online product reviews using DLMNN and future prediction of online product using IANFIS," *J. Big Data*, vol. 7, no. 1, p. 33, Dec. 2020, doi: 10.1186/s40537-020-00308-7.

[21] Z. Drus and H. Khalid, "Sentiment Analysis in Social Media and Its Application: Systematic Literature Review," *Procedia Comput. Sci.*, vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[22] I. El Alaoui, Y. Gahi, R. Messoussi, Y. Chaabi, A. Todoskoff, and A. Kobi, "A novel adaptable approach for sentiment analysis on big social data," *J. Big Data*, vol. 5, no. 1, p. 12, Dec. 2018, doi: 10.1186/s40537-018-0120-0.

[23] A. Shelar and C.-Y. Huang, "Sentiment Analysis of Twitter Data," in *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Dec. 2018, pp. 1301–1302, doi: 10.1109/CSCI46756.2018.00252.

[24] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep Learning for Hate Speech Detection in Tweets," *26th Int. World Wide Web Conf. 2017, WWW 2017 Companion*, Jun. 2017, doi: 10.1145/3041021.3054223.

SOFTWARE VULNERABILITY DETECTION USING MACHINE LEARNING – A REVIEW

Jaswant Kaur[#1], Dr. Dhavleesh Rattan[#2], Er. Gurpreet Singh[#3]
[#]*Computer Science and Engineering Department, Punjabi University*
[1]jaswantkaur884@gmail.com
[2]dhavleesh@gmail.com
[3]gurpreet.1887@gmail.com

ABSTRACT— Consistently increasing software vulnerabilities have become one of the major concerns for the software industry. Software Vulnerability Detection is recent years has become one of the major attraction for the researchers related to security experts and they are using different methods and approaches with different results. Machine Learning Techniques integrate well with this to bring more efficient solutions. Recent breakthroughs in machine learning, deep learning usage in detection and mitigation of software vulnerabilities has changed the way vulnerabilities are detected or mitigated in software industry. This change has made programmers and cyber security engineers to use machine and deep learning in large extent in software vulnerability detection and mitigation. In our paper, we have focused on machine learning methods in detection of software vulnerabilities. In all the techniques of vulnerability discovery like: static analysis, symbolic execution and fuzzing, basic fundamentals have been stated first. Then, we reviewed the research area of detection of software vulnerability using machine learning techniques. In the last, we have summarize the benefits and drawbacks of different methods or approaches used in the work. The work includes in-depth review of recent works in the field of software vulnerability using Machine Learning.

This document gives formatting instructions for authors preparing papers for publication in the Proceedings of a conference. The authors must follow the instructions given in the document for the papers to be published. You can use this document as both an instruction set and as a template into which you can type your own text.

KEYWORDS— Machine Learning, Static Analysis, Symbolic Execution, Fuzzing, Software Vulnerability Discovery Include at least 5 keywords or phrases

INTRODUCTION

Vulnerability is a weak factor which can be exploited by hacker to gain unauthorized access, steal data, or can damage the Confidentiality, Integrity, Availability(CIA) of the system. Software Vulnerability is a bug, weakness in a software or Operating System which can be exploited by hackers. Software Vulnerabilities cases are rising rapidly in this internet age. The impact and severity of software vulnerabilities depends on complexity in exploitation and what is the attack surface [1]. Large number of vulnerabilities arises everyday taking a toll out of computing infrastructure, applications, companies, and individuals. For example, there was a server message block (SMB) based vulnerability which was used by WannacryRansomware and attacked millions of people around the world [2]. Software Vulnerability is a bug, problem in a software or OS. Almost all the systems includes the vulnerabilities that grows exponentially. There are three important factors which are used to define software vulnerability:

1) Existence**:** The presence of vulnerability in the application.
2) Accessibility: Probability of hackers or crackers getting access.
3) Exploitable**:** Is the vulnerability exploitable?

A table below presents the top 10 products with most vulnerabilities on the basis of Common Vulnerability Scoring System (CVSS).

TABLE X
TOP 10 SOFTWARE WITH DISTINCT VULNERABILITIES [22]

Software	Vendor	No. of distinct vulnerabilities
Debian Linux	Debian	3067
Android	Google	2563
Linux Kernel	Linux	2357
Mac OS	Apple	2212
Ubuntu Linux	Canonical	2007
Firefox	Mozilla	1873
Chrome	Google	1858
iPhone OS	Apple	1655
Windows Server 2008	Microsoft	1421
Windows 7	Microsoft	1283

There are different vulnerability detection and classification techniques present, some of them are manual, automatic and some are hybrid with integration of manual and automatic techniques. When we take code execution into account, there we can divide the techniques to static, dynamic or hybrid analysis. If we are using open source software then testing can be black-box, white-box or gray-box. This paper focuses on static analysis, symbolic execution and fuzzing techniques.

Machine Learning

Machine learning is a technology where we can make computers act without the need of them programmed explicitly. Table comparing different types of machine learning techniques is defined below:

TABLE II
COMPARING MACHINE LEARNING APPROACHES

Criteria	Supervised	Unsupervised	Reinforcement
What it is ?	In this, machine learns by using labeled data.	With no guidance at all, machine is trained with the use of unlabeled data.	Here agent contacts the environment by executing actions and learning from errors.
Problem Type	Regression and Classification	Association and Clustering	Reward basis
Data Type	Labeled	Unlabeled	Data is not predefined
Training	External Supervision is done.	There is no supervision.	There is no supervision.
Approach	Integrated the labeled inputs to the known outputs.	This approach understands pattern and then discovers the output.	This uses trial and error method.

Machine Learning brings new efficiency and intelligence in classification approaches. Basic fundamentals of vulnerability detection using machine learning are reviewed.

Static Analysis with Machine Learning

1) *Fundamentals of Static Analysis:* This method works by assessing the system on the basis of form, content, documentation and it do not need execution of programs [3]. On the basis of target, it can be marked as either source-code or binary analysis. Although, both the methods have similarities, but binary analysis is more complex [12]. Static analysis method comprises policy matching, data, information and control flow analysis, abstract exposition and model checking [3]. Machine Learning is extensively used with static analysis thesedays with much better results than before.
2) *Recent Research Work Summary:* Walden et al. [4] used programming language just like a native language and performed analysis of source using the text mining methods. Researcher used Bag-of-Words method and software components are used as mixture of terms with integrated frequencies. Researcher performed analysis on 20 apps on Android OS with 5 machine learning algorithms. Algorithms used are Decision Trees, Support Vector Machines (SVM), Random Forest, k-Nearest Neighbor (KNN) and Naïve Bayes. According to the results in the research, Naïve Bayes and Random Forest brings best results. Researcher used a commercial vulnerability analysis software for three experimentation works. First work includes Naïve Bayes and Random Forest techniques for one version of the app, while the other two experiments are performed by creating a prediction analysis model with successive versions and cross-project apps. As per the results obtained, first two results are termed as feasible while the last one is not. Song et al. [5] uses artificial neural networks (ANN) algorithm for binary analysis to intercept the issues related to identify problems. The major challenge is the absence of semantic structure in the binaries because compilers remove it from the source. The main part is the recognition of function in the binary analysis. Work done by researcher includes on-hot encoding method that translates a byte to a vector and add it to the neural network input and then uses bi-directional techniques along RNN hidden values. The researcher experiment proves that RNN is much better in identification of functions as compared with recent methods.

   Smith et al. [6] proposed a technique to detect software vulnerabilities. Researcher integrates the code-metric assessment along with the meta information inside the code repository. To assess the productivity of the proposed method, researcher firstly used the 66 C++ project from the Github that were having a total of 170860 commits comprises of 640 vulnerability related commits. In the proposed work, metadata is created using author, project information, file name and commits and then bring out the code-churn and activities related to developer as codemetrics. Lastly, a Bag-of-Words technique is used for the features discussed before and a classifier is trained data on and before 2010 and then test data is used form year 2011 to 2014 with the precised value of 6 percent.

   Grinblat et al. [7] proposed a technique that used both static and dynamic attributes in order to forecast if the binary program has the ability to have software vulnerability. In the proposed work, researcher used Bag-of-Words along with Word2Vec technique to implement various feature-sets. Random Oversampling was used for class imbalance work [13]. Machine Learning is used for classification with logistic regression and random forest algorithms used. Random Forest turned out to be the better out of both algorithm for classification when trained with dynamic feature-sets.

   Li et al. [8] proposed a method with name Vulnerability Deep Pecker, which uses deep learning algorithms to detect software vulnerabilities. According to the researcher, current vulnerability detection systems mainly rely on the humans and that results in large number of false negatives. Researcher uses RNN algorithm along with Long

Short Term Memory (LSTM) to label vanishing gradient issue. Vulnerability Deep Pecker works in two parts: the learning and detection, where in learning part, system takes out the API methods and the code snippets, modify the code to vector representations and the perform training of Bidirectional LSTM network and in detection part, the system modifies the target in code slices and vectors. According to researcher, the results achieved have much lower false negatives than tradition software vulnerability detection system.

Zhang et al. [9] suggested a method in order to highlight the lack of better quality training data and rely totally on the manual software vulnerability detection features [10] with the use of machine learning algorithms. Researchers tagged 457 functions which are vulnerable and 32530 function which are not vulnerable from around six different projects, and draw out the abstract syntax trees or ASTs from code used with CodeSensor parser from [11]. Author proposed a Bidirectional LSTM based neural network which takes the input. As the work is divided into layers, where the first layer uses Word2vec embedded layer that integrates every element of the sequence to the vector. Second one is the LSTM layer that includes 64 LSTMs in one bidirectional form. Last layer in the network is used for the global pooling layer.

Table III below provides the summary related to the recent research works ion static analysis using machine learning techniques:

TABLE III
RECENT RESEARCH SUMMARY ON STATIC ANALYSIS USING MACHINE LEARNING ALGORITHMS

Author [Reference] (Year)	Approach Used	Benefits	Drawbacks	Future Scope	Dataset	Feature Representations	Detection Granularity
Hanif et al. [23] (2021)	Supervised, Semi-Supervised Learning, Deep Learning, Ensemble Learning	Detects Software Vulnerability with high accuracy	Small Dataset	Work on multi-vulnerability systems and real-world systems and software.	SARD, NVD, CodeChef, WebGoat, Pebble, Vuldeepecker etc.	Static and Dynamic Sequence of Libraries, Code Gadgets, Source Gadgets	Program Level
G. Lin et al. [24] (2020)	Deep Neural Networks	Found different vulnerabilities using neural network techniques.	Minor Test Case	To work on real time software and large dataset.	Vuldeepecker	Text-Based, Sequence Based, Graph based	Program Level
Bilgin et al. [25] (2020)	Machine Learning	Large Public Vulnerability Dataset taken from open-source projects.	5 Predetermined vulnerabilities predicted with ML and Hyper-parameter Optimization.	To work on code similarity and code completion analysis. To improve localization and interpretation for software vulnerability detection.	Draper VDISC	Static and Dynamic Sequence of Libraries, AST	Intra and Inter Procedural
Walden et al. [4] (2014)	BoW + Random Forest + Naïve Bayes	Inside the Project	Cross Project	Expansion for cross project work	NVD	Static and Dynamic Sequence of Libraries	Intra-Procedural. Inter-Procedural
Song et al. [5] (2015)	Bidirectional technique with RNN	It identifies the functions if binary code	It does rely totally on train data	Internal mechanics should be stated well	Self-Generated	Static and Dynamic Sequence of Libraries	Program Level
Smith et al. [6] (2015)	BoW + SVM	Precision & Recall	It relies on manual work analysis before training	To lower the probability	CVE	Static and Dynamic Sequence of Libraries	Program Level
Grinblat et al. [7] (2015)	BoW + W2V + Logistic Regression + Random Forest + Static/Dynamic Features	Accurate	Minor Test Cases	Introduction of CNN	VDiscovery	Static and Dynamic Sequence of Libraries	Program Level
Li et al. [8] (2018)	Code Snippets/Gadget + RNNs Bidirectional LSTM	Low False Negatives along with no relying on manual features	Just contains buffer and resource related errors	Solve the problems or drawbacks	VulDeePecker	Static and Dynamic Sequence of Libraries,	Program Level
Zhang et al. [9] (2018)	CodeSensor + RNNs Bidirectional LSTM	Does not rely in training data or manual features and is much more effective	This work does not relate to vulnerabilities with multiple files	Solve the problems or drawbacks	NVD, CVE	AST, DFT	Program Level

Source of Information

1) ScienceDirect (www.sciencedirect.com).
2) IEEEeXplore(ieeexplore.ieee.org).
3) ACM Digital Library (www.acm.org/dl).
4) Google Scholar (https://scholar.google.com).

TABLE IV
SEARCH PARAMETERS AND STRINGS

Sr. No.	e- source	Search string	Years	Subject	Total	Rejected based on title	Rejected based on abstract	Rejected based on full text	Accepted
1	Science direct	Entire Paper :"Machine Learning" AND " Vulnerability" Title :vulnerability	All	All	33	13	10	9	1
2	Ieee	Abstract :vulnerable* AND software vulnerability AND machine learning	All	All	39	10	11	8	10
3	Acm	Title :software vulnerability AND machine learning	2015-2021	All	901	598	202	95	6
4	Usenix	Entire paper :Malware Detection and Machine Learning	2015-2021	All	202	100	56	45	1
5	Arxiv	Entire paper : Machine Learning Coding Vulnerability	2018-2021	All	9	4	2	2	1

Symbolic Execution with ML Algorithms

1) *Fundamentals of Symbolic Execution:* This [14] is a technique used for reasoning the code which uses symbolic values for inputs and not the actual data values, here the values of the code variable are acting as symbolic expression for input values. When the decision related and jump statement occurred, the technique will add the path constraints of the existing implementation path to the constraint path. Constraint resolver is used to fetch the availability of the path. It the result related to constraint resolves the problem, then the path becomes reachable, otherwise it acts as unreachable.

2) *Recent Research Work Summary:* Li et al. [15] uses machine learning algorithms to direct the main problem in using symbolic execution to production related issues. The issue is related to the resolving of constraints that are closely binded to the optimum problem, i.e. to find the solution to lower the dissatisfaction. Researchers here proposed machine learning based approach that helps in encapsulating the complex tasks as symbolic constraints and shifts the feasibility related issues of path conditions in the optimum issues. Researchers uses a machine learning based approach known as RACOS technique [16] which distinguish between good and bad solutions while achieving a small shrink and error rate. As per the results in their research work, proposed work was feasible and more reliable instruction coverage and efficiency better than recent works. Wen et al. [17] come up with a new technique that integrated machine learning and symbolic execution in order to get the software vulnerabilities. In the very first part, researchers get the vulnerability related functions from CVE. Then in second part, graphs which are related to the software vulnerability function are extracted from the source code and is then used for guiding the symbolic execution engine to get to the target. Results of the research work shows that symbolic execution along with machine learning algorithms get to the vulnerability functions in 36s as compared to 8h of symbolic execution alone.

E. Fuzzing techniques with ML Algorithms

1) *Fundamentals of Fuzzing technique:* This[18] is one of the most popular automated techniques that is used in different invalid datasets like network protocols, API Calls, and many different targets as input in order to make sure the absence of vulnerabilities which can be exploited. This method has three characteristics [19]: first one is generation of random data which can be semi-valid, then second one is that it transmits the data which is generated in first part to the target app, and in last part, it sees if the app is failing while consuming the data. Semi-Valid data is one of the most important part of the fuzzing technique and it directly influence the effectiveness of this technique. Two main methods used for generation of data are data-generation and data-mutation, where data-generation methods usually is based on specifications, on the other hand, data-mutation generates the data by updating the fields of data inputs which are valid. Researchers use ML algorithms in order to resolve issues related to fuzzers.

2) *Recent research work summary:* Pham et al. [20] uses the technique of American Fuzzy Loop or AFL as a proper process based exploration of Markov Chain and takes the likelihood if fuzzing a seed that utilizes program path

acreates a seed that utilizes path **b**with likelihood of **pab.** The workdone by the researchers better the power schedules by associating inversely proportional energy with the immobilized distribution. Peleg et al. [21] initiates neural networks methods to fuzzing and suggested sample fuzz technique that uses learn input likelihood in order to provide proper guidance on where to use fuzz based inputs. Researchers uses character based RNN method for deep learning and adopts the samplespace technique for the sample policy and in the end, researchers used sample fuzz technique to build newer PDF instances.

CONCLUSION

Machine Learning is rapidly getting into almost all the applications which can be healthcare, transportation, network security etc. Software Vulnerability detection is no different. This paper reviews some of recent research work related with software vulnerability detection using machine learning techniques and it categorized in three broad categories i.e static analysis, symbolic execution and fuzzing techniques. For every category, we have provided introduction, summary analysis, which can be helpful in understanding the recent research work in software vulnerability detection or discovery using machine learning techniques. This helps in analyzing the recent work and answering some queries related to usage of machine learning in software vulnerability detection like: What are the benefits of using ML in software vulnerability detection?, Which methods to use of software vulnerability detection?, and What are the recent research work done in ML in software vulnerability detection.

REFERENCES

[1] Nayak, K., Marin, D., Efs, P., D, T.: Some vulnerabilities are different than others. In: Stav, A., B, H., Portokal, G. (eds.) RAID 2014. LNCS, vol. 8688, pp. 426–446. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11379-1_21

[2] Chen, Q.: B, R.: Automated behavioral analysis of malware: a case study of WannaCryRansomware. In: the 16th IEEE International Conference On Machine Learning And Applications, pp. 454–460, Cancun, Mexico (2017). https://dblp.uni-trier.de/pers/hd/c/Chen:Qian

[3] Liu, B., S, L., C, Z., Li, M.: Software vulnerability discovery techniques: a survey. In: the 4th International Conference on Multimedia Information Networking and Security, Nanjing, China (2012)

[4] Scan, R., Walden, J., Hovs, A., Josen, W.: Predicting vulnerable software components via text mining. IEEE Trans. Softw. Eng. 40(10), 993–1006 (2014). https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=32

[5] Shin, E., Song, D., Moazzezi, R.: Recognizing functions in binaries with neural network. In: the 24th USENIX Security Symposium, Washington, D.C., USA (2015)

[6] Perl, H., De, S., Smith, M.: VCCFinder: finding potential vulnerabilities in opensource projects to assist code audits. In: Proceeding of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 426–437, Denver, Colorado, USA (2015)

[7] Grieco, G., Grinb, G., U, L., Rawat, S., F, J., Moun, L.: Toward large-scale vulnerability discovery using machine learning. In: Proceedings of the 6th ACM Conference on Data and Application Security and Privacy, pp. 85–96, San Antonio, TX, USA (2015)

[8] Li, Z.: VulDeePecker: a deep learning-based system for vulnerability detection. In: the 25th Annual Network and Distributed System Security Symposium, NDSS, California, USA (2018)

[9] Lin, G., Zhang, J.: Crossproject transfer representation learning for vulnerable function discovery. IEEE Trans. Ind. Inf. 14, 3289–3297 (2018).

[10] Chen, L., Yang, C., Liu, F., Gong, D., Ding, S.: Automatic mining of security sensitive functions from source code. CMC: Computing. Mater. Cont. 56(2), 199–210 (2018)

[11] Y, F., Lotman, M., Rick, K.: Generalized vulnerability extrapolation using abstract syntax trees. In: Proceedings of the 28th Annual Computer Security Applications Conference, pp. 359–368 (2012) 316

[12] Ghaffar, S., Shahriari, H.: Software vulnerability analysis and discovery using machine learning and data-mining techniques: a survey. ACM Computing. Survey. 50(4) (2017)

[13] He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions. Knowledge. Data Eng. 21(9) (2009)

[14] Chu, D.H., Jaffar, J., Murli, V.: Lazy symbolic execution for enhanced learning. In: the 5th International Conference on Runtime Verification, pp. 323–339, Toronto, ON, Canada (2014).

[15] Li, X.: Symbolic execution of complex program driven by machine learning based constraint solving. In: Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, pp. 554–559, Singapore, Singapore (2016)

[16] Yu, Y., Qian, H., Hu, Y.Q.: Derivative-free optimization via classification. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp. 2286–2292 (2016)

[17] Meng, Q., Wen, S., Zhang, B., Tang, C.: Automatically discover vulnerability through similar functions. In: 2016 Progress in Electromagnetic Research Symposium (PIERS), Shanghai, China (2016).

[18] Oehler, P.: Violating assumptions with fuzzing. IEEE Secur. Priv. 3(2), 58–62 (2005)

[19] Liu, B., Shi, L., Cai, Z., Li, M.: Software vulnerability discovery techniques: a survey. In: Fourth International Conference on Multimedia Information Networking and Security (2012)

[20] Böhme, M., Pham, V.T., Roy, A.: Coverage based greybox fuzzing as Markov Chain. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, NY, USA (2016)

[21] Godefroid, P., Pele, H., Singh, R.: Learn&Fuzz: machine learning for input fuzzing. In: Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering, pp. 50–59. Urbana-Champaign, IL, USA (2017)

[22] Top 50 Products by Total number of Distinct Vulnerabilities (2020), https://www.cvedetails.com/top-50-product-cvssscore-distribution.php.

[23] Hazim H, Mohd H Nasir, Mohd R, Ahmad F, Nor A, The rise of software vulnerability: Taxonomy of software vulnerabilities detection and machine learning approaches, Journal of Network and Computer Applications, Volume 179, 2021, 103009, ISSN 1084-8045

[24] G. Lin, S. Wen, Q. -L. Han, J. Zhang and Y. Xiang, "Software Vulnerability Detection Using Deep Neural Networks: A Survey," in Proceedings of the IEEE, vol. 108, no. 10, pp. 1825-1848, Oct. 2020.

SPORTS ANALYTICS WEB API USING DEEP LEARNING APPROACH

CHINU SINGLA[1,*], RAMAN MAINI[1], MUNISH KUMAR[2],
[1]*Department of Computer Science and Engineering, Punjabi University Patiala, Punjab, India*
[2]*Department of Computational Sciences, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab*
[1]cheenusingla10@gmail.com
[2]research_raman@yahoo.com
[3]munishcse@gmail.com

ABSTRACT: This paper is intended for sportspersons in general, but specifically it targets basketball players to help them understand where their efforts should be targeted to improve or add some new techniques to it so as to increase the probability of scoring the basket. In order to construct the shooting prediction model of basketball free throw, we analyze movies of basketball free throw motions with a full hi-vision video camera. Human body pose detection based on deep learning can be of considerable importance in the surveillance industry. This innovation can be used at sports venues, airports, train stations, and other crowded places to enhance security. Human pose estimation, coupled with other data science algorithms for motion recognition and analysis, and deep neural networks (DNNs) to tackle the constraints involved in this, is the perfect combination to help in preventing and restraining violent situations.

KEYWORDS: Sports analytics, deep learning, API, R-CNN, Machine learning, OpenPose.

I. Introduction

Sports analytics has become more popular in these days [1]. It provides specialized methodologies for gathering and analyzing sports data in order to make decisions for successful planning and implementation of latest strategies. Sports analytics is described as the procedure of data management, predictive model implementation, and the use of information systems for decision making to gain a competitive benefit on the field of play [2]. Sports analytics is the concept of analyzing sports data through analytic methodologies that aids in making valuable conclusions [3]. The shooting prediction model is a binary prediction as to whether to enter a basket or not. As major binary prediction models are logistic regression and Support vector machine (SVM). The SVM using the kernel method is a nonlinear model which may make high accuracy but cannot calculate the shooting probability. The data of basketball free throw in this experiment were taken from one side only by a web camera, so it was suitable to analyze with 2-dimensional data provided by OpenPose. However, analysis of general sports motion requires 3-dimensional data like a tennis or ballet dance, so it is necessary to use 3 dimensional OpenPose or expand 2D data generated by 2D OpenPose to 3D data. A technology such as positioning generates a high amount of data, and there are various areas such as business data, image data and industrial process data [4].

The output will serve as a system for basketball sports analytics. Physically the system does not require any attached device. The product does require an input file image/video which is used for the analytics which points out the required. For example, the user has the access to upload a file up to 5Mb and the necessary things will be predicted as per output. Firstly, the user will have to register him/her against the normal authentication terminal and once the terminal identifies the card user by matching the cryptographic key generated by the software with the one generated by the card, the welcome message is displayed to the user. Next the software inside the terminal will check whether the user is registered or not. If the user has been denied access, then the software will send the message "Access Denied" onto the screen. Otherwise, an "Access Granted" message will be displayed and the student can then enter the lab and access the resources. Then the user will have to upload the file against the dashboard terminal as shown in Figure 1 and when the user clicks the analytics button it generates the analytics of the system with the prediction of basket, the probability message is displayed to the user on the screen.

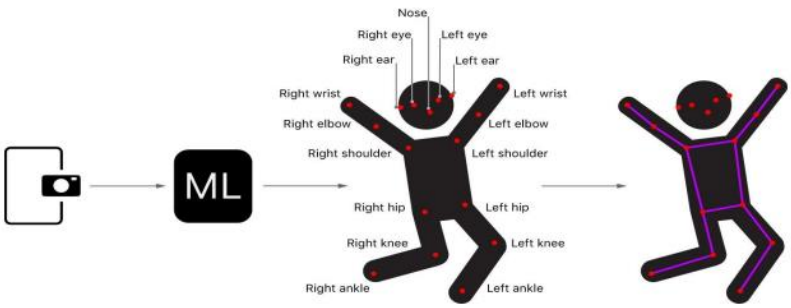


Fig. 1 Media fragmentation for the analytics

II. Related Work

This section describes machine learning based sport data analysis. Constantinou et al. [5] created probabilistic models based on possession rates and other historical figures of different teams to predict the result of matches. Podgorelec et al. [6] developed a latest image dataset of four similar sports (American football, soccer, rugby and field hockey) and built a technique to group those images using transfer learning of CNN with Hyper Parameter Optimization. Their proposed work

was then compared to conventional-CNN and a CNN with transfer learning. Kapadia et al. [7] used machine learning techniques to mitigate the similar issue but for the cricket world in the Indian Premier League (IPL). Kerr [8] showed three experiments in his thesis. In the first experiment, three models were built using various characteristics to analyze which team won a given game, without any prior knowledge of goals. Several classifiers were used to predict which team produced the sequence of ball-events occurred during a game in the second experiment. And in the last one, he predicted which particular team attempted a given set of passes. Jayalath [9] considered the famous logistic regression model to study the importance of one-day international (ODI) cricket predictors. Brooks et al. [10] aimed on investigating features of passing in soccer and described two techniques for obtaining insights from that. Ghimire et al. [11] used Adjusted Plus-Minus (APM) techniques to monitor the contribution of player in basketball and hockey. Knobbe et al. [12] used linear modeling and subgroup discovery to choose key features and generate interpretable models for sport data analytics in professional speed skating. Sidle and Tran [13] considered multi-class classification methods to predict baseball live pitch types. Whereas, Chu and Swartz [14] introduced a Bayesian inference system using parametric models to analyze fouling time distributions. Karetnikov [15] developed a principally new complex performance prediction framework for cycling where Maximum Mean Power (MMPs) and the race position are the performance metrics.

III. System Methodology

The system should require deep learning, batch, CSS, HTML, Python and flask knowledge for maintenance. If any problem acquires in server side and deep learning methods, it requires code knowledge and deep learning background to solve. Client-side problems should be fixed with an update and it also requires code knowledge and network knowledge. Figure 2 describes the flow of data between user and the final product for basketball analytics system.

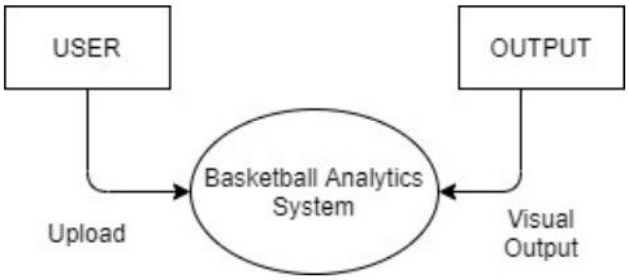


Fig. 2 Data Flow Diagram

A. Pose Estimation

OpenPose is a pose estimation library which estimates pose using VGG19 neural network with part affinity vector. This is trained on COCO dataset and has been originally implemented in caffe. We convert it to tensor flow 2.0 so that keras deep learning models can be used with it. The OpenPose network first extracts feature from an image using the first few layers (VGG-19 in the above flowchart). The features are then fed into two parallel branches of convolutional layers. The first branch predicts a set of 18 confidence maps, with each map representing a particular part of the human pose skeleton. The second branch predicts a set of 38 Part Affinity Fields (PAFs) which represents the degree of association between parts. Successive stages are used to refine the predictions made by each branch. Using the part confidence maps, bipartite graphs are formed between pairs of parts (as shown in the above image). Using the PAF values, weaker links in the bipartite graphs are pruned. Through the above steps, human pose skeletons can be estimated and assigned to every person in the image. We currently use the 17 key point-based system. OpenPose system architecture is being represented in Figure 3.

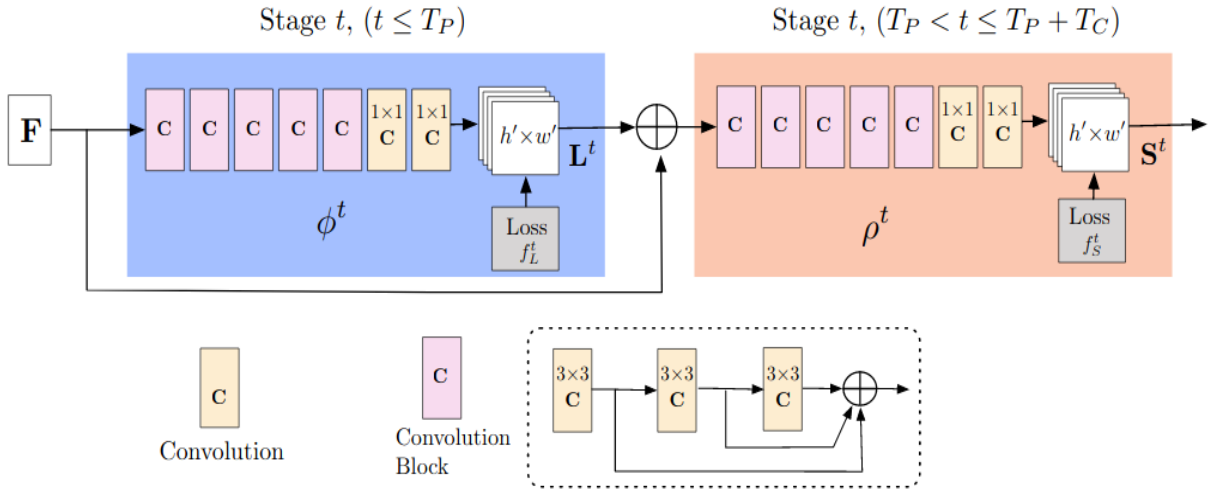


Fig.3 OpenPose system architecture

B. Assumptions and Dependencies

- The software has to be integrated onto the terminal that in turn has a limited capability for API request.
- There are no memory requirements.
- The product must have a user-friendly interface that is simple enough for the users.
- Response time for loading the software and for processing an output should be no longer than seven seconds.
- A general knowledge of basic computer skills and of basic working of navigating and uploading is required to use the product.
- The image/video should be at least 360p so to get better output.

C. Product functions

The product should be able to perform the following operations:

- It must be able to authenticate the user and manage the access against the values stored in the database.
- It must be able to access the video status by the database and the videos uploaded.
- The software must be able to update the video access privileges onto a particular user's credentials and the database where the privileges themselves will be modifiable only by the system administrators (or some authorized staff members).
- The software must be able to determine whether the shooting can result in a basket or not a basket for a particular image/video.
- Should also successfully analyze the pose of the player and track the basketball

D. Objectives

The goal is to design the software for a selected group of individuals or team managers and to develop software that should be easy to use for all types of users. While designing the software one can assume that user type has the following characteristics:

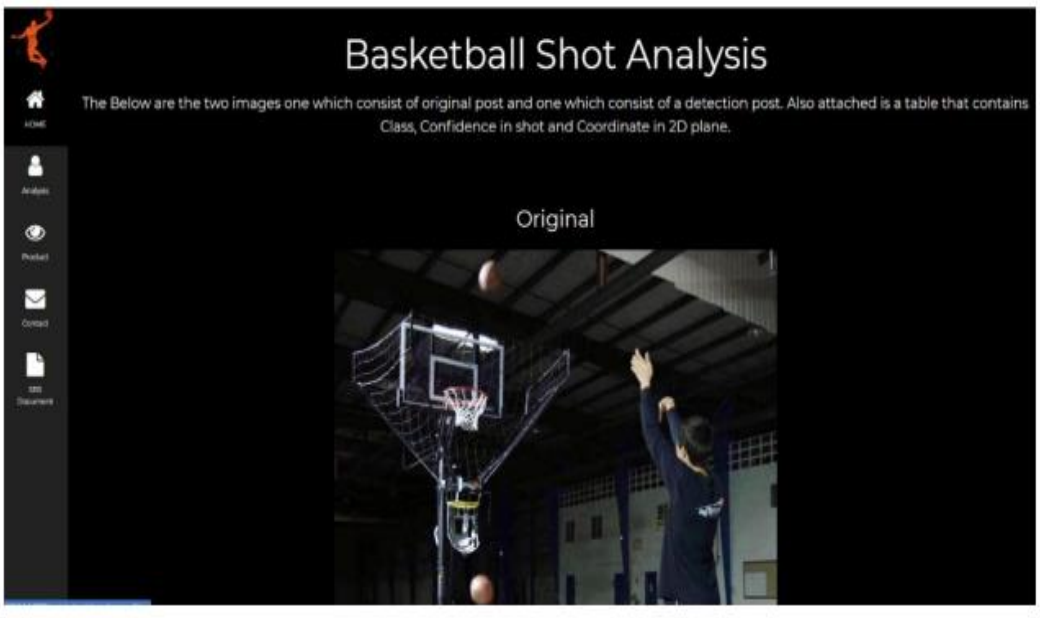
- The user is computer-literate and has little or no difficulty in using a smart card to access information such as room status.
- In order to use the system, it is not required that a user beware of the internal working of a smart card but he/she is expected to know how to access the results in the analytics column.

E. Test levels

- Making a deep learning model for object detection.
- Converting the famous OpenPose library which was only compatible with Caffe with Tensor Flow.
- Analyzing the resulting variables using a main logic and feeding the result to the API backend.
- Binding the backend to support rendering functions for the media and feed the results to the front end.
- Designing a user-friendly Web API for uploading the media.

Fig 4 a, b shows the Graphical user interface for Basketball analysis.

a)



b)

Fig. 4 (a, b) GUIs for Basketball analytics

IV. Conclusion and Future Work

The contribution of this paper is to provide a comprehensive sketch of software product, its parameters, and its purposes. However, there are still many open questions and challenges in the area. This section presents several possible directions for future work.

- Data privacy is a main concern (buying fan data and types of data and how data is analyzed). Therefore, refining the data so that it is ready for fan consumption is a tedious task.
- Data analytics in sport is hugely important. As established, clear data is needed to improve stadium services.
- A sport-specific platform that brings together rights stakeholders, sponsors can be created.
- Personalized sport agility training systems can be created using player activities, sports nutrition, exercise drills, tactics, and techniques using big data analytics' capabilities.

REFERENCES

[1] Rajitha Minusha Silva. "Sports analytics". PhD thesis. Science: Statistics and Actuarial Science, 2016.

[2] Benjamin Alamar and Vijay Mehrotra. "Beyond 'Moneyball': The rapidly evolving world of sports analytics, Part I". In: Analytics Magazine, 2011.

[3] Thomas A Severini. Analytic methods in sports: Using mathematics and statistics to understand data from baseball, football, basketball, and other sports. Crc Press, 2020.

[4] Thomas A Runkler. Data Analytics. Springer, 2020.

[5] Anthony C Constantinou, Norman E Fenton, and Martin Neil. "pifootball: A Bayesian network model for forecasting Association Football match outcomes". In: Knowledge-Based Systems 36, 2012, pp. 322–339.

[6] Vili Podgorelec, Spela Peˇcnik, and Grega Vrbanˇciˇc. "Classification of Similar Sports Images Using Convolutional Neural Network with HyperParameter Optimization". In: Applied Sciences 10.23, 2020, p. 8494.

[7] Kumash Kapadia et al. "Sport analytics for cricket game results using machine learning: An experimental study". In: Applied Computing and Informatics, 2020.

[8] Matthew George Soeryadjaya Kerr. "Applying machine learning to event data in soccer". PhD thesis. Massachusetts Institute of Technology, 2015.

[9] Kalanka P Jayalath. "A machine learning approach to analyze ODI cricket predictors". In: Journal of Sports Analytics 4.1, 2018, pp. 73–84.

[10] Joel Brooks, Matthew Kerr, and John Guttag. "Using machine learning to draw inferences from pass location data in soccer". In: Statistical Analysis and Data Mining: The ASA Data Science Journal 9.5, 2016, pp. 338–349.

[11] Shankar Ghimire, Justin A Ehrlich, and Shane D Sanders. "Measuring individual worker output in a complementary team setting: Does regularized adjusted plus minus isolate individual NBA player contributions?" In: PloS one 15.8 (2020), e0237920.

[12] Arno Knobbe et al. "Sports analytics for professional speed skating". In: Data Mining and Knowledge Discovery 31.6 (2017), pp. 1872–1902.

[13] Glenn Sidle and Hien Tran. "Using multi-class classification methods to predict baseball pitch types". In: Journal of Sports Analytics 4.1,2018, pp. 85–93.

[14] Dani Chu and Tim B Swartz. "Foul accumulation in the NBA". In: Journal of Quantitative Analysis in Sports 1.ahead-of-print, 2020.

[15] Aleksei Karetnikov. "Application of Data-Driven Analytics on Sport Data from a Professional Bicycle Racing Team". Eindhoven University of Technology, The Netherlands, 2019

AUTOMATED CANDIDATE SELECTION SYSTEM FOR RECRUITMENT USING NLP

Iqra Maryam Imran, Dr. Jayalakshmi D. S.
Dept. of Computer Science and Engineering,
M S Ramaiah Institute of Technology, Bangalore
[1]iqraimran630@gmail.com
[2]jayalakshmids@msrit.edu

ABSTRACT - In the recent times there is a very high demand for automation in recruitment process. Recruiting people that suit a certain job description is a critical activity for the majority of businesses. Traditional recruiting practices are becoming ineffective as online recruitment grows in popularity. In the current generation, companies receive a lot of resumes and reviewing each of the resumes in a very time-consuming process. A lot of research is being carried out using Machine Learning for automating the process of recruitment. In this paper, we design and develop an application that leverages Natural Language Processing and Machine Learning techniques to find a perfect candidate for recruitment in all domains by extracting valuable information from the resume to find the best suitable role for the candidate and ranking it according to the preference and requirement of the company.

KEYWORDS- Candidate Selection, Automation, NLP, Machine Learning techniques, Linear SVC

I. INTRODUCTION

The process of recruitment is a critical role for every organization's HR division and the first step toward building strength. In the recent times there is a very high demand for automation in the recruitment process. Recruiters are not focused in getting majority of their repetitive work and tasks like calling candidates for interview, processing various resumes, coordinating with them candidates for scheduling interviews. It takes a large amount of time for work like resume screening, employee management, compliance, and more. Short-listing the resumes to obtain the best possible candidate manually is another time consuming process[1]. For all these reasons we have seen an exponential growth in terms of automation in the field of recruitment. AI is forced to be recognized in HR theme related automation. Artificial intelligence for recruitment is a booming category requirement in HR. The technology is designed to very specifically reduce time consuming activities like manual screening and resume interpretation. AI for recruiting refers to the application of AI and machine learning to recruitment processes in order to streamline some aspects of the recruiting workflow, particularly repetitive high-volume tasks[2]. It is a type of technology that allows businesses to automate recruiting tasks and workflows, so that they can increase recruiter productivity, shorten time-to-fill, lower cost-per-hire, and improve the enterprise's overall diversity and productivity. Many recruiters may question the need for automation when traditional methods are adequate.

As internet connectivity continues to rise, all big companies have moved their recruitment procedure. Recruiters can attract many people for their opportunities by use of online job posts on numerous employment portals and websites. Business apps and enterprise systems acquired popularity and recognition as web technologies advanced and the internet became more widely available[3]. While e-recruitment has made recruiters and applicants pleasant and affordable, there are new challenges. Big companies and recruitment agencies receive tens of thousands of CVs each day. This scenario is made worse because of the higher mobility of workers and the economic troubles in which many people seek employment. It is impossible for recruiters to go through each and every CV for this small number of positions, with less than 5% of the people can choose from these submissions. Another challenge facing the organizations is that these applicants are not using an uniform summary format. People have various professional backgrounds and come from many different sectors. Each has various kinds of education, has worked on different projects and has a unique style of presenting credentials in the resume. Summary materials are unstructured and are not written in standard formats or templates in various file formats. That does not make reading resumes simple, and hence recruiters spend a great amount of time selecting the right candidates through the manual process [4]. This difficulty of processing non- organized and various resumes has been eased by many work portals and external websites. These demand candidates to fill in all of their curriculum vitae information online in an organized manner and so provide metadata for candidates. The problem with this technique is that candidates have to make duplicated attempts and often fail to finish the information in these templates. These websites use a general format that is not customized to the domain and is therefore not appropriate for all applications. These templates are then used by employer to look for candidates with keywords. This keyword-based search feature is not enough to match the job description of candidates. This is such that it relies only on particular keywords and has different limits to extract from them such as avoiding natural language semantics like synonyms and words combinations, and contextual significance of the contents included in the abstract. These search strategies hence often produce irrelevant results and may not choose the candidates that really deserve being shortlisted.

To achieve better outcomes for resume listing, more efficient ways to applicant matching and job description need to be investigated. Our proposed method will choose the best possible candidates for a particular job opening by looking for the main features of the profile of the applicants with the requirements defined in the job description. We have seen that there are hundreds, often thousands, of resumes that are analyzed by the recruiter for a single job. It is little wonder that hired managers often spend hours at the top of the funnel screening. NLP saves a lot of time for evaluation of bound curricula and candidates for screening. NLP can assist erase subconscious discrimination and increase the diversity of candidates in addition to saving time. The requirement is to propose a system that streamlines the recruitment process across all sectors

by extracting valuable information from the CV and ranking it according to the company's preferences and needs. This system makes it easier for companies, candidates looking for jobs and the client firm that hires the candidates. Our algorithm will work to optimize the present result with Machine Learning by way of the previous results and the previous ranking constraints. This guarantees that the candidate for this particular vacancy is hired. Our system enables the firm, according to its specific standards and limitations, to draw the finest possible list of candidates. This kind of approach will improve and enhance our recruiting industry as the person concerned gets a simple solution. It reduces the manual effort required in screening the resumes and finding the best match. Thus our proposed system plays a very important role in the future of recruitment.

This paper is organized in the following sections: Section 2 describes the related work which has been done in this field. Section 3 provides an insight into the entire system we have developed and the detailed methodology and the theoretical concepts involved with our solution. Section 4 represents the results of the experiment performed using our system and Section 5 concludes our work followed by the future scope in Section 6.

II. RELATED WORK

Resume analysis is an excellent technique for "weeding out" the incorrect applicant and finding the perfect expert, allowing us to interview the candidate with the most relevant model of behavior and a sense of his personal traits already in place.The E-recruitment process has been found to involve processes such as searching the job folder, evaluating candidates, short-listing candidates, and making final decisions by recruiting managers. Furthermore, the process is afflicted by a variety of challenges, including an increase in the number of unsuitable job seekers, complaints about discrimination and diversity, and synchronization issues. Although there is a vast literature on the benefits and drawbacks of the E-recruitment process, there is a scarcity of information on the clear technique of E-recruitment and proper practices, suggesting the need for more research in this area[5].

A. Natural language processing techniques

Natural Language Processing (NLP) is a technique that allows computers to understand human languages. Words are typically used as the fundamental unit in deep-level grammatical and semantic analysis, and word segmentation is typically the core job of NLP. According to a research 'Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language' by Dongyang Wang, Junli Su and Hongbin Yu[6], Feature extraction solves the problem of determining the most compact and informative feature collection. It is the key technology for processing high-dimensional data, and it finds use in industrial detection and diagnostic systems, speech recognition, biotechnology, targeted marketing, and a range of other application domains. The key to experimental study is extracting useful characteristics from text while avoiding unnecessary data processing.

'Sentiment Analysis using Feature Extraction and Dictionary-Based Approaches' by D. Deepa, Raaji and A. Tamilarasi[7] uses the feature extraction techniques and the dictionary-based methods to determine the polarity of words in tweets. Using a machine learning classifier, the method is compared to feature engineering approaches such as CountVectorizer, TF-IDF, and Word2Vec, and each word is scored using SentiWordNet and VADER dictionaries. The results prove that feature extraction technique is a better approach than the dictionary-based approach.

Along with Information extraction, a paper by Veena G, Hemanth R and Jithin Hareesh named Relation Extraction in Clinical Text using NLP Based Regular Expressions introduces relation extraction. It is similar to information extraction but the only difference being, even the relationship between the entities is extracted in relation extraction along with information extraction[8]. Even these types of extraction are done with natural language processing. This helps in making the processing task easier.

Various approaches can be used to evaluate the datasets using natural language processing. A research on the various approaches is performed by Abhishek Kumbhar, Mayuresh Savargaonkar, Aayush Nalwaya published as Keyword Extraction Performance Analysis[9]. The datasets are analysed using five Natural Language Processing approaches: TF-IDF, RAKE, TextRank, LDA, and Shallow Neural Network. They employ a ten-fold cross-validation technique and use recall, precision, and F-score to assess the effectiveness of the aforementioned approaches. Their study and findings give guidance on the best techniques to follow for various sorts of datasets.

B. Machine learning techniques

Text classification is an important component of NLP, since it tries to predict the categories for input texts in a certain classification system. There are a number of ways to this like feature selection and classification models. A research 'An Exploration on Text Classification with Classical Machine Learning Algorithm' by Yuhan Zheng[10] assess the code in several categorization models, which means that the computers will still have to be trained extensively using a huge number of training sets, which will take a long time. In addition, the selection of a classification model varies depending on the situation, and no classification model is acceptable for every situation.

Machine Learning (ML) today has a large number of various algorithms. There are various methods for classification models, including Logistic Regression, Nave Bayes Classifier, K-Nearest Neighbors, Decision Tree, and Random Forest Classifiers. A work proposed by Vedant Bahel; Sofia Pillai and Manit Malhotra called 'A Comparative Study on Various Binary Classification Algorithms and their Improved Variant for Optimal Performance' present a comparative study of various binary classifier and have implemented various boosting algorithms[11]. The articles presented a broad comparison

of the described models against common criteria such as speed, computing power, transparency, and so on. This, in turn, indicated the overall performance of each model. The Random Forest Classifier outperformed all other algorithms tested throughout the investigation, followed by the Nave Bayes Classifier.

Another work 'Emotional Text Analysis Based on Ensemble Learning of Three Different Classification Algorithms' by WenShuo Bian; ChunZhi Wang; ZhiWei Ye and Lingyu Yan developed an integrated learning model that incorporates three distinct classification algorithms: logistic regression, support vector machine, and K-Neighborhood method This method outperforms a single classification algorithm in terms of accuracy. The experimental findings demonstrate that the model performs well in terms of applicability and performance[12].

C. Existing e-recruitment systems

This section summarizes some of the literary work performed in this domain of e-recruitment systems. The proposed solutions use various approaches with the aim of achieving automated screening of candidates.

A study was conducted on the impact of Artificial Intelligence technologies on the process of recruitment, which included a study on key Artificial Intelligence capabilities, its potential outcomes and importance in the process of recruitment. It is recommended that firms learn to collaborate with AI technologies so they may teach AI technologies to be extensions of their teams rather than replacements[13]. Recruiters stated that using AI technology in the recruiting process may speed up the process while also being cost effective. Using AI technology in the recruiting process can improve the quality of the process by increasing accuracy and decreasing human bias. The use of AI in the recruiting process can make it easier for recruiters to find the appropriate applicant with the correct skill set for the right job. Overall, the use of AI technology in the recruiting process can minimize recruiters' effort while improving applicant experience. Firms should learn to work with AI technologies; AI technologies should be taught to be extensions of their teams rather than substitutes.

The work presented as The Architecture of Distant Competencies Analyzing System for IT Recruitment[14] says that there are two issues to be addressed, which are calculating the technical knowledge competency of students and analyzing their personal attributes. Recruiters are unable to automatically choose students for further work unless these issues are addressed. Another work is proposed as Automated Resume Evaluation System using NLP (Rohini Nimbekar , Yogesh Patil and Rahul Prabhu)[15] that proposed a model that takes the relevant data from a resume and separates it depending on its values It then rates resumes based on the essential criteria, and companies examine just the top few applicants. It is recommended that firms learn to collaborate with AI technologies. They may teach AI technologies to be extensions of their teams rather than replacements. The main disadvantage was that the domain was limited to engineering student resumes, and the amount of sample data versus the test data was very modest. The domain might be expanded to include various other fields such as telecommunications, healthcare, government positions and e-commerce.

Another research paper named Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming[16], presents a data analytics approach that may be used by recruiters in real-world settings as a decision support tool to improve candidate hiring decisions for certain roles or vocations. The proposed approach consists of two components: a local prediction model for recruiting success based on candidate and job type, and a global optimization model for the recruitment process. The first section of this paper is all about the interpretability of ML modeling, which gives useful insights into possible recruitments based on the candidate's background characteristics as well as the anticipated work placement. The probability of successful recruiting per employee and job are the outcome of these models. The second half of this research is based on an organizational-level mathematical modeling formulation that takes into account multi-objective considerations and enhances the recruiting process across multiple candidates and roles by leveraging the success probability outputs of the ML models.

With all the research, we have found a various classification algorithms performed for predicting the personality of the user. According to Automated Personality Classification Using Data Mining Techniques by Bhavna Singh and Swasti Singhal[18], the Naïve Bayes Algorithm gave better accuracy among the two methods tested. The average accuracy obtained by Naïve Bayes is around 60%. The Support Vector Machine method performance was lower than Naïve Bayes as it was not quite accurate due to the difficulties of separating a class of a word as dataset.

The most recent paper found, An Automated Interview Grading System in Talent Recruitment using SVM, by Muhammad Yusuf and Kemas M Lhaksmana[19], states that the algorithm SVM has the highest average accuracy which is found to be 84%. The next best result is given by naive Bayes which is 81% and then KNN giving 79%. The average f1-score of SVM is greater than that of other techniques, confirming its excellent accuracy. In this work, SVM consistently outperforms naive Bayes and KNN when applied to any dataset and in a variety of schemes. Another benefit of SVM is its flexibility, which results in strong performance when applied to a dataset with an unbalanced class distribution. Pre-processing or feature extraction strategies must be further determined via more study. These stages must be tailored to the features of the language as well as the scope of the subject matter. Furthermore, classification algorithms must be used with other machine learning approaches.

Table 1 summarizes and depicts a comparative study on the most recent papers on the existing smart recruitment systems.

TABLE XI
LITERATURE SURVEY

SR. No.	Title	Objective	Result	Drawback/ Future Work
1.	A Comparative Study on Various Binary Classification Algorithms and their Improved Variant for Optimal Performance[11]	The objective is to give a comparative examination of several binary classifiers and algorithms, as well as to summarize the associated justifications for optimal performance of the presented classification models.	The results prove that when compared to the other algorithms investigated throughout the research, the Random Forest Classifier performed best, followed by the Nave Bayes Classifier.	Limited numbers of classification algorithms are compared. The future work would be to evaluate more algorithms like SVM etc.
2.	An Automated Candidate Selection System Using Bangla Language Processing[17]	The objective is to make the process of recruitment easier and quicker by developing an automated candidate selection system for the process of recruitment.	The smart system built employed 50 CVs of technical background testing of the system in the system performance assessment and discovered that the system works effectively to return the best prospects by matching the specified requirements with the candidate's resumes.	The future work on this system is to make the system more dynamic by taking the job description and the resumes directly. This system could be further used by expanding it to languages other than Bengali.
3.	An Exploration on Text Classification with Classical Machine Learning Algorithm[10]	The objective is to put code into action in order to perform functions and see the effects of feature selection and classification models.	The paper concludes that to some extent, the choice of n-gram can influence the outcome of text categorization. When the texts are diverse, picking the proper model to accomplish the categorization enhances the outcomes.	The future work involves more consideration should be given to the selection of categorization models as well as machine training. A larger number of training sets must be used to teach the machines.
4.	Smart Talents Recruiter – Resume Ranking and Recommendation System[20]	To find the best applicants for a particular job need using an effective candidate rating policy. The Smart Applicant Ranker (SAR) system will be used by all categories of users: job seekers who post their resumes and job recruiters who search for the best applicants.	An enhanced candidate recommendation system with is created improved performance and accuracy. The accuracy obtained is 83.33%	A higher accuracy and precision could be obtained. The future work involves more research along the same lines to attain better accuracy.
5.	An Automated Interview Grading System in Talent Recruitment using SVM[19]	To investigate machine learning approaches for exploring the characteristics of the job applicants through verbatim interview analysis.	SVM has the highest average accuracy with the measure of 85%. After SVM is naive Bayes which gives an accuracy of 81% followed by KNN with an accuracy of 79%.	Determine preprocessing or feature extraction techniques
6.	Resume Parser with Natural Language Processing[21]	This paper proposes a model that can parse information from unstructured resumes and transform it to a structured format and select the extracted resumes based on the job description.	The proposed model successfully converted different formats of resumes to text and was able to parse relevant information from them to obtain the best suitable candidate for each requirement.	The future work includes extracting the information of the candidates from social networking sites like Twitter etc so the suitability to the role can be evaluated more accurately.
7.	Automated Resume Evaluation System using NLP[15]	The study presents a methodology for extracting relevant information from resumes and rating based on the company's preferences and requirements.	The proposed model successfully extracts the necessary data from a resume and segments them based on their values.	The domain is restricted to the field of computer science only. The future work can be extended further to other domains as well.

III. METHOD

A. Flowchart

The entire process of this project is shown in Fig. 1. The process is carried out by considering various conditions that are combined when preprocessing, feature extraction and model selection between Naive Bayes, Logistic Regression and SVM.

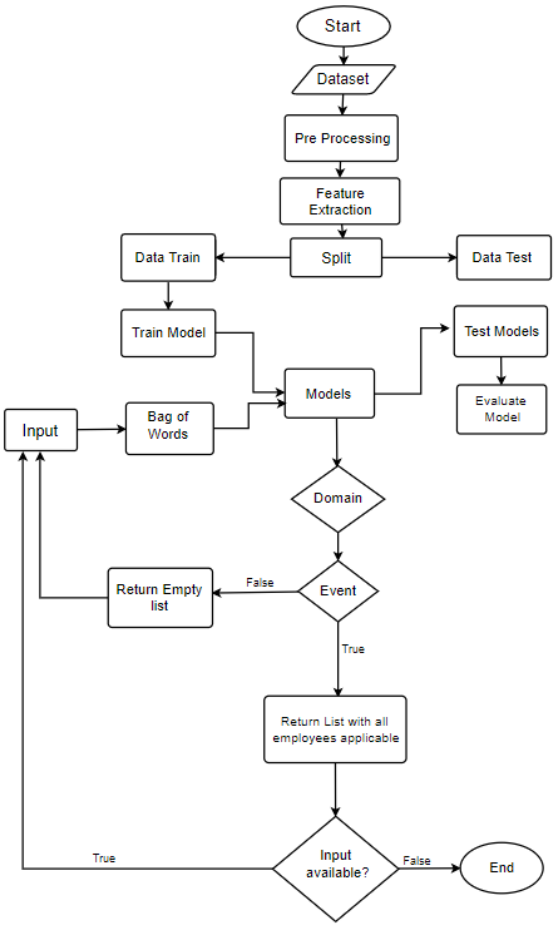


Fig. 1 Process

B. System architecture

System architecture is a conceptual model that specifies a system's structure, behavior, and additional perspectives. It defines the framework of the project to be carried out. In our machine learning project, it defines the various layers of machine learning cycle that is involved. It includes all the steps of transforming the raw CV's into the training data that helps us to enable the decision making for the results. Fig. 2 depicts the entire system architecture starting from the applicant uploading the resume to providing that applicant with the best possible role depending on his skills and work experience.

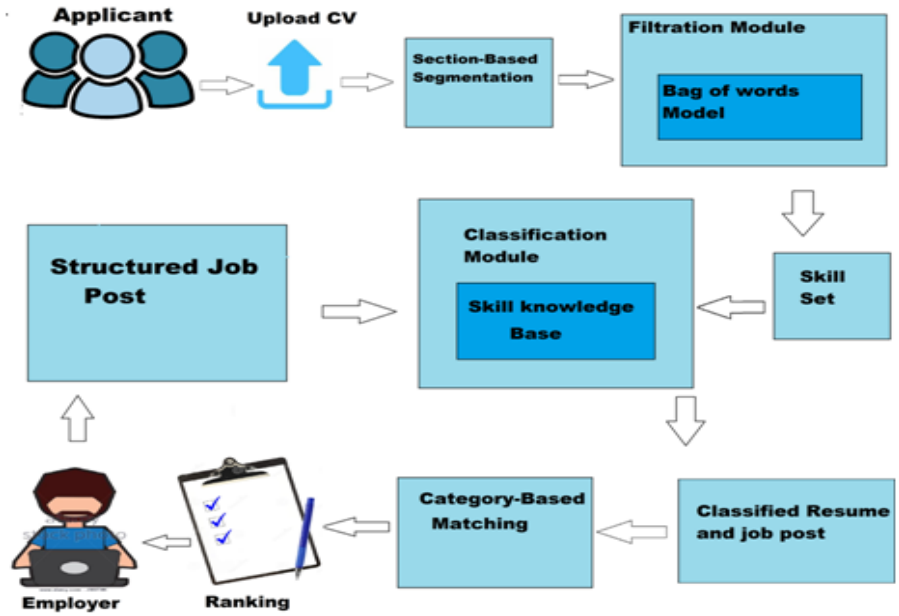


Fig. 2 System Architecture

C. Data preprocessing

Data preparation is a crucial stage in Machine Learning that helps enhance data quality and facilitates the extraction of meaningful insights from data. Text categorization is greatly influenced by preprocessing. This phase tries to eliminate any unnecessary words or symbols. Although preprocessing has an effect on text classification, it is not as powerful as feature extraction, feature selection, and classification. Preprocessing consists of many processes that must be completed. Few of the pre processing techniques implemented in our paper include:

1) *Stopwords:* Stopwords are the words in the text, in our case, the resumes, which does not add much meaning to a sentence and thus are not very useful. Stopwords are being removed from this project since they create noise without providing any information value in modelling. We extracted a list of English stopwords from the nltk package and removed them from the resumes.
2) *Tokenization:* To process text, we must first divide it into smaller parts. Tokenization in Python refers to the process of breaking up a huge body of text into smaller lines, words, or even inventing terms for a language other than English. The different tokenization routines are provided within the nltk module and can be utilized in applications. In this project, we have used nltk.word_tokenize that returns a tokenized copy of the text.
3) *Lemmatization:* Lemmatization is the process of reducing words to their basic word, which results in linguistically accurate lemmas. Otherwise, it becomes difficult to identify the same word in a different format[23]. It uses vocabulary and morphological analysis to alter root words. In most cases, lemmatization is more complex than stemming. Stemmer operates on a single word without knowing the context. Considering an example, lemmatization will exactly understand the base form of 'leaving' to 'leave', whereas, stemming will cut off the 'ing' part and convert it to leav. We utilized WordNet's Lemmatizer to transform each word into its base term. WorldNet is a huge, freely and publicly accessible English lexical database that aims to build organized semantic links between words. It is also capable of lemmatization and was one of the earliest and most commonly used lemmatizers. To lemmatize, make an instance of the WordNetLemmatizer() and use the lemmatize() method on a single word.

D. Feature extraction

Text feature extraction is the process of extracting a list of words from text input and converting them into a feature set that a classifier may use. Feature extraction aims to reduce the number of features in a dataset by producing new ones from existing ones. In our methodology, the Bag of words strategy is utilized to extract characteristics from resumes.

Bag of words is one of the most common and convenient feature extraction techniques is the bag of words method. It generates a feature set from all of the words in an instance. The approach is known as a "bag" of words since it is unconcerned with how many times a word occurs or the sequence of the words, all that counts is that the word exists in a list of words. The traits can be used to model using machine learning approaches. This method is extremely flexible and simple. It is often implemented in a variety of ways to extract features from text data. A text data representation is a bag of words. It is very helpful in our project as it returns the frequency of words in the resumes. It contains both the lexicon of specific words and the frequency with which those recognized words appear in resumes. The only issue faced with the bag of words approach is determining how to judge the existence of recognized phrases as well as how to create the lexicon of familiar words.

E. Machine Learning Algorithms Used

The proposed system uses 3 machine learning algorithms.

i. Logistic Regression
ii. Multinomial Naive Bayes
iii. Linear SVC(Support Vector Classifier)

1) Logistic Regression

Logistic regression is a very effective algorithm for classification problems, more accurately in cases of binary classification. This algorithm uses a logistic function to determine the relationship between the output and input variables. This logistic function is known as the sigmoid function which is the basic principle of logistic regression. It takes any real-valued input and obtains output by translating it to a value between 0 and 1.
Considering an observation with input x, which is represented as a vector of features [x1, x2,..., xn]. The classifier output y can be 1 which means the resume matches to the requirements or 0 which means the resume does not match to the requirements. The objective is to calculate the probability $P(y = 1|x)$ that this resume is a member of the class flagged which indicates if the resume matches the requirement or not. So perhaps the decision is "positive sentiment" versus "negative sentiment". $P(y = 1|x)$ is the probability that the resume has positive sentiment, and $P(y = 0|x)$ is the probability that the resume has negative sentiment.

Logistic regression solves our task of finding if the requirements match with the resumes by learning, from our dataset, after converting them into a vector of weights and a bias term. This algorithm outputs predictions about test data points on a binary scale, zero or one. Thus it clearly classifies if the resume matches the requirement or not. Each weight wi is a real number, and is associated with one of the input features xi. The weight wi represents how important that particular feature like the education, experience etc is to the classification decision, and can be positive (flagged which means the requirements match) or negative (meaning the requirements do not match and belong to negative class). The bias term is

also added to the weighted features. Thus we can derive the equation for z, which is the weighted sum to know the class it belongs. The derived equation as follows,

$$z = \left(\sum_{i=1}^{n} w_i x_i\right) + b$$

In our project, since weights are real-valued, the output might even be negative representing the resumes that do not match the requirement which lets us have the result z ranging from $-\infty$ to ∞. To create a probability, we'll pass z through the sigmoid function, $\sigma(z)$. The sigmoid function is also called the logistic function. The logistic function is represented graphically in Figure 5.2 and the sigmoid function as follows,

$$y = \sigma(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + \exp(-z)}$$

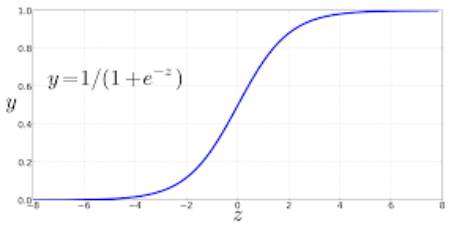


Fig. 3 Logistic Regression

Fig. 3 depicts the graph on which the data points, the resumes would be potted depending on how close the resume is to the requirement. The line divides the resumes that match or do not match the requirements. The distance from the line helps determine the accuracy of matching the resume to the requirement.

2) Multinomial Naive Bayes

One of the first machine learning algorithms was the Naive Bayes classifier. It can anticipate and forecast data based on prior results and is suitable for binary and multiclass classification. This category has more algorithms than the Naive Bayes classifier. Multinomial Naive Bayes, for example, is frequently used for document categorization based on the frequency of certain phrases in the document. Multinomial Nave Bayes use the phrase frequency, or the number of times a specific word appears in a text. Following normalization, term frequency may be used to create maximum likelihood estimates based on training data to forecast conditional probability.

To understand how the classification mechanism works in detail, let us consider a resume r, class c, then the probability is computed as,

$$P(c|r) \propto P(c) \prod_{1 \le k \le n_r} P(t_k|c)$$

We interpret $P(t_k|c)$ as a measure of how much evidence tk contributes that the class c is the correct class. P(c) is the prior probability of the resume occurring in class c. If the result does not give unambiguous proof supporting one class against another, we select the one with the larger prior probability. Thus we can classify with accuracy if the resume falls in the class indicating it matches the requirement or not.

3) Linear SVC (Support Vector Classifier)

Another approach used for categorization and, in some situation regression, is support vector machines. SVM is fantastic because it provides extremely accurate results while consuming relatively few processing resources. They work by drawing a line between various groupings of data points in order to classify them. Points on one side of the line will be assigned to one class, while points on the other will be assigned to a different class.

The SVM's goal is to find a hyperplane in an N-dimensional space that distributes the data points clearly. The number of features determines the size of the hyperplane. In Linear SVC, hyperplane is graphically represented as a line that separates one class from another. Data points on opposite sides of the hyperplane are classified differently. When there are just two input characteristics, the hyperplane is merely a line. When the number of input characteristics is increased to three, the hyperplane transforms into a two-dimensional plane. As we have two classes in our project, we classify with a line, which is known as Linear SVC. On either side of the hyperplane that separates the data, two parallel hyperplanes are built. Bigger the margin or distance between these parallel hyperplanes, the better the classifier's accuracy.

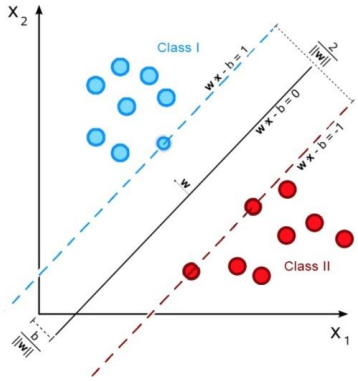


Fig. 4 Linear SVC

Fig. 4 depicts the various data points on the graph which represents the various resumes in the system. The line divides the resumes into 2 classes, Class 1 containing the resumes that match the requirement and class 2 representing the resumes that do not match the requirements. The accuracy of the resume matching the requirement or not is derived from the distance from the separating line.

IV. RESULT AND ANALYSIS

We have designed and developed an application that leverages Natural Language Processing and Machine Learning techniques to extract the valuable information from the resumes and help the recruiters to make the process of recruitment easier. The dataset used in the system contains resumes of all domains.

The various test cases of the system are as follows.

A. Test Case 1 – User Registration

The users, the candidates that want to upload their resumes into the system first need to create an account for registration. Figure 6 is a snapshot of the registration page.

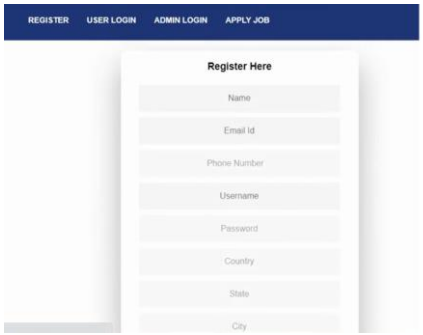


Fig. 5 User Registration

Once the candidate is registered to the system, he can upload his resume as shown in Fig. 6, the snapshot from the system.

Fig. 6 Resume Upload

B. Test Case 2 – User login

Once the user has registered into the system, uploaded his resume and all other details, he can login with his credentials to check for any other updates on his profile. This ensures the safety and privacy to ones data. Fig. 7 is a snapshot of the user login page.

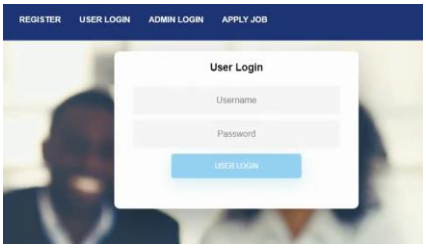


Fig. 7 User Login

C. Test Case 3 – Admin login

The admin can view all the resumes in the pool and use the trained model to find the suitable candidates depending on the requirements. Fig. 8 is a snapshot of the login page for the admin.

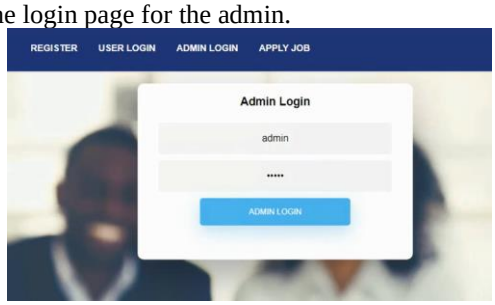


Fig. 8 Admin Login

D. Test Case 4 – No matches found

The admin can find in the system the best match as per the requirements. Fig. 9 is a snapshot of the system where we can search for resumes from the pool depending on the department and the major skill that we have been looking for.

Fig. 9 No resume

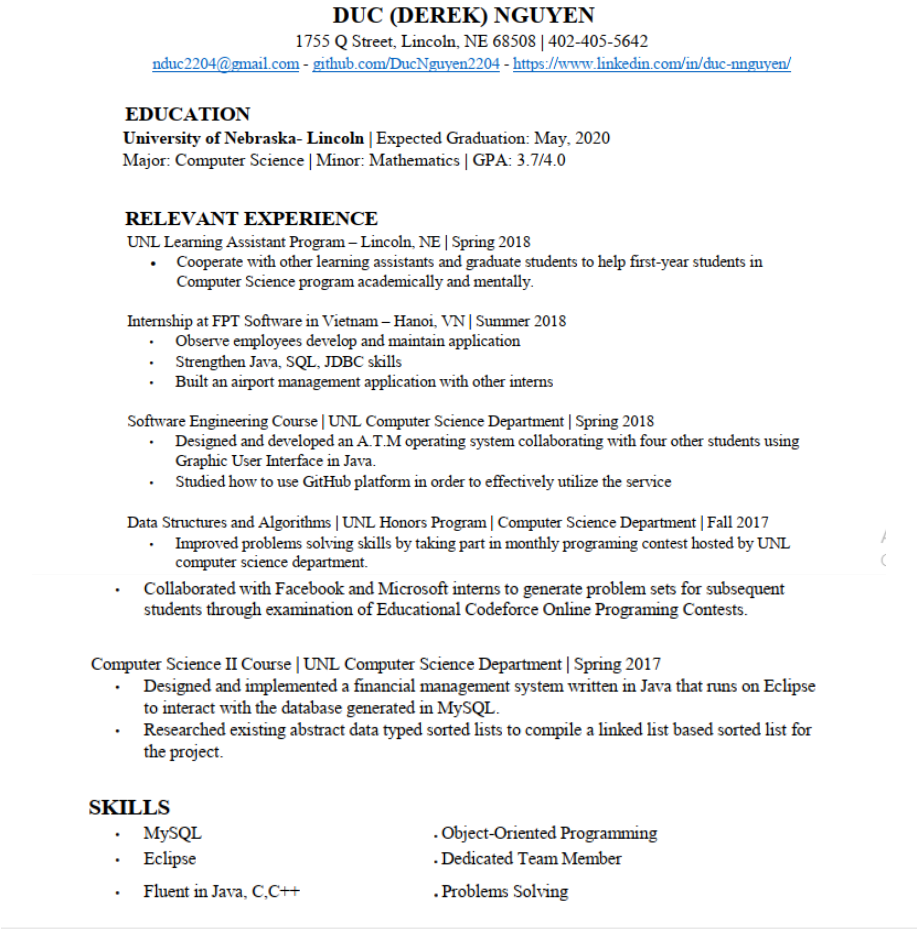
E. Test Case 5 – Match found

Once all the details are entered, Fig. 10 depicts how the system suggests with the best possible resume from the existing resumes in the pool. It then gives the option of downloading the resume, thus making the process of recruitment easier for the recruiters.

Fig. 10 Resume Found

In order to find the major skill in the resumes, the model has been trained with the machine learning algorithms. The system provides the most suitable job for the candidates with accuracy using the trained models. This process of automation saves the time and effort needed to manually go through each and every resume. Following are some of the examples.

F. Example 1 - Resume of a candidate named Nguyen.



DUC (DEREK) NGUYEN

1755 Q Street, Lincoln, NE 68508 | 402-405-5642

nduc2204@gmail.com - github.com/DucNguyen2204 - https://www.linkedin.com/in/duc-nnguyen/

EDUCATION

University of Nebraska- Lincoln | Expected Graduation: May, 2020

Major: Computer Science | Minor: Mathematics | GPA: 3.7/4.0

RELEVANT EXPERIENCE

UNL Learning Assistant Program – Lincoln, NE | Spring 2018

- Cooperate with other learning assistants and graduate students to help first-year students in Computer Science program academically and mentally.

Internship at FPT Software in Vietnam – Hanoi, VN | Summer 2018

- Observe employees develop and maintain application
- Strengthen Java, SQL, JDBC skills
- Built an airport management application with other interns

Software Engineering Course | UNL Computer Science Department | Spring 2018

- Designed and developed an A.T.M operating system collaborating with four other students using Graphic User Interface in Java.
- Studied how to use GitHub platform in order to effectively utilize the service

Data Structures and Algorithms | UNL Honors Program | Computer Science Department | Fall 2017

- Improved problems solving skills by taking part in monthly programing contest hosted by UNL computer science department.
- Collaborated with Facebook and Microsoft interns to generate problem sets for subsequent students through examination of Educational Codeforce Online Programing Contests.

Computer Science II Course | UNL Computer Science Department | Spring 2017

- Designed and implemented a financial management system written in Java that runs on Eclipse to interact with the database generated in MySQL.
- Researched existing abstract data typed sorted lists to compile a linked list based sorted list for the project.

SKILLS

- MySQL
- Eclipse
- Fluent in Java, C,C++
- Object-Oriented Programming
- Dedicated Team Member
- Problems Solving

Fig. 11 Nguyen's resume

As the resume says, the candidate looks suitable for job like Java developer as his past experience and skills are with Java. The prediction made by our system says the candidate is right for the post of java developer. Fig. 12 shows the results obtained by our system.

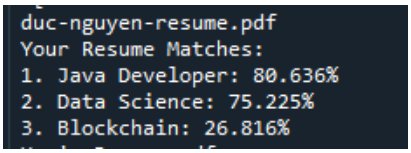


Fig. 12 Prediction for Nguyen

As the result says, the candidate is suitable as a Java developer with the accuracy of 81%. The candidate could also be taken for the role of Data Science as he has some past experience in the field.

G. Example 2 - Huy's resume.

Fig. 13 is the output obtained for Huy's resume. The output says that he is most suitable for the role of Data Science. The accuracy of the role of Data Science is 83.6%. The system also suggests the candidate could be taken as a python developer or DevOps Engineer but the accuracy being very less. This way our system finds the 3 best possible options.

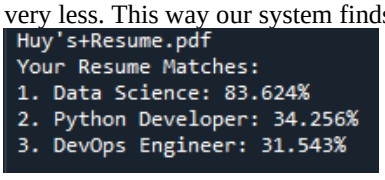
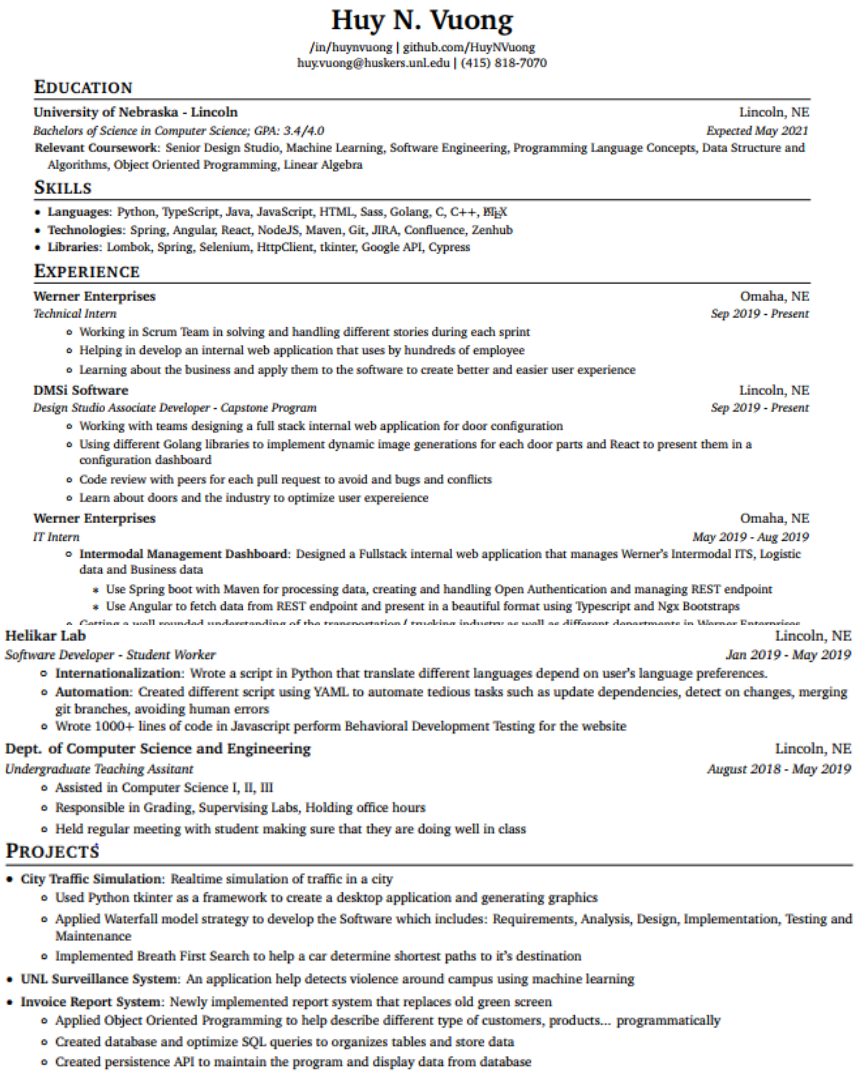


Fig. 13 Prediction for Huy

On reading Huy's resume, we understand that the results obtained by the system are correct if the process is done manually. The resume is shown below in fig. 14. The system proves to find the right results and thus reduces the work of recruiters.



Huy N. Vuong

/in/huynvuong | github.com/HuyNVuong
huy.vuong@huskers.unl.edu | (415) 818-7070

EDUCATION

University of Nebraska - Lincoln — Lincoln, NE
Bachelors of Science in Computer Science; GPA: 3.4/4.0 — *Expected May 2021*
Relevant Coursework: Senior Design Studio, Machine Learning, Software Engineering, Programming Language Concepts, Data Structure and Algorithms, Object Oriented Programming, Linear Algebra

SKILLS

- **Languages**: Python, TypeScript, Java, JavaScript, HTML, Sass, Golang, C, C++, LaTeX
- **Technologies**: Spring, Angular, React, NodeJS, Maven, Git, JIRA, Confluence, Zenhub
- **Libraries**: Lombok, Spring, Selenium, HttpClient, tkinter, Google API, Cypress

EXPERIENCE

Werner Enterprises — Omaha, NE
Technical Intern — *Sep 2019 - Present*
- Working in Scrum Team in solving and handling different stories during each sprint
- Helping in develop an internal web application that uses by hundreds of employee
- Learning about the business and apply them to the software to create better and easier user experience

DMSi Software — Lincoln, NE
Design Studio Associate Developer - Capstone Program — *Sep 2019 - Present*
- Working with teams designing a full stack internal web application for door configuration
- Using different Golang libraries to implement dynamic image generations for each door parts and React to present them in a configuration dashboard
- Code review with peers for each pull request to avoid and bugs and conflicts
- Learn about doors and the industry to optimize user expereience

Werner Enterprises — Omaha, NE
IT Intern — *May 2019 - Aug 2019*
- **Intermodal Management Dashboard**: Designed a Fullstack internal web application that manages Werner's Intermodal ITS, Logistic data and Business data
  - Use Spring boot with Maven for processing data, creating and handling Open Authentication and managing REST endpoint
  - Use Angular to fetch data from REST endpoint and present in a beautiful format using Typescript and Ngx Bootstraps
- Getting a well rounded understanding of the transportation / trucking industry as well as different departments in Werner Enterprises

Helikar Lab — Lincoln, NE
Software Developer - Student Worker — *Jan 2019 - May 2019*
- **Internationalization**: Wrote a script in Python that translate different languages depend on user's language preferences.
- **Automation**: Created different script using YAML to automate tedious tasks such as update dependencies, detect on changes, merging git branches, avoiding human errors
- Wrote 1000+ lines of code in Javascript perform Behavioral Development Testing for the website

Dept. of Computer Science and Engineering — Lincoln, NE
Undergraduate Teaching Assitant — *August 2018 - May 2019*
- Assisted in Computer Science I, II, III
- Responsible in Grading, Supervising Labs, Holding office hours
- Held regular meeting with student making sure that they are doing well in class

PROJECTS

- **City Traffic Simulation**: Realtime simulation of traffic in a city
  - Used Python tkinter as a framework to create a desktop application and generating graphics
  - Applied Waterfall model strategy to develop the Software which includes: Requirements, Analysis, Design, Implementation, Testing and Maintenance
  - Implemented Breath First Search to help a car determine shortest paths to it's destination
- **UNL Surveillance System**: An application help detects violence around campus using machine learning
- **Invoice Report System**: Newly implemented report system that replaces old green screen
  - Applied Object Oriented Programming to help describe different type of customers, products... programmatically
  - Created database and optimize SQL queries to organizes tables and store data
  - Created persistence API to maintain the program and display data from database

Fig. 14 Huy's resume

In this project, we have ensured that the system will work for all the domains. Fig. 15 depicts the wide range of domains or skill set that is included in our dataset. This overcomes the drawback from the previous papers of the system not being operated across domain.

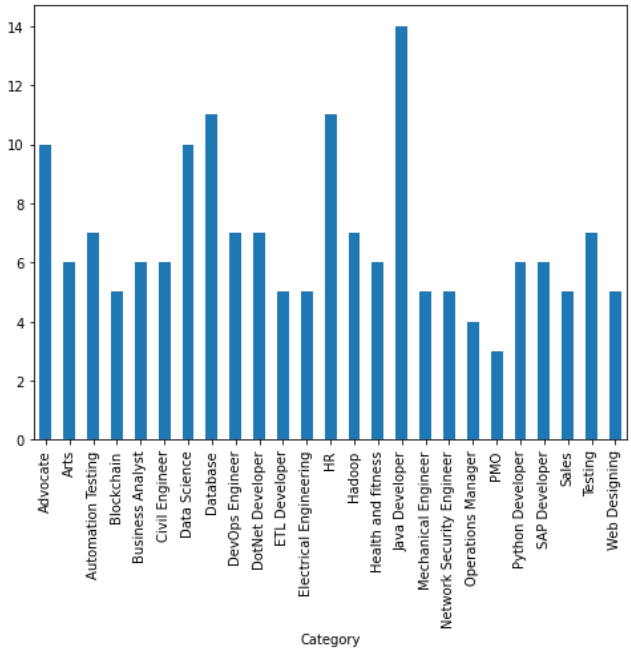


Fig. 15 Graph showing various domains

Another important feature is that this project successfully obtains the highest accuracy in comparison to the other latest papers using the Linear SVC algorithm. The Accuracy is the number of correctly classified resumes over the total number of resumes. A snapshot of the results obtained using the various algorithms are depicted in Fig. 16. The accuracy obtained using the Logistic Regression algorithm is 51.2% which is the lowest among the best 3 chosen algorithms in this paper. The Multinomial Naïve Bayes algorithm gave an accuracy of 76.7% which worked out to give better results than Logistic Regression. The Linear SVC algorithm gives the highest accuracy of 88.3%. This accuracy obtained is higher than the accuracies obtained in the previous latest papers.

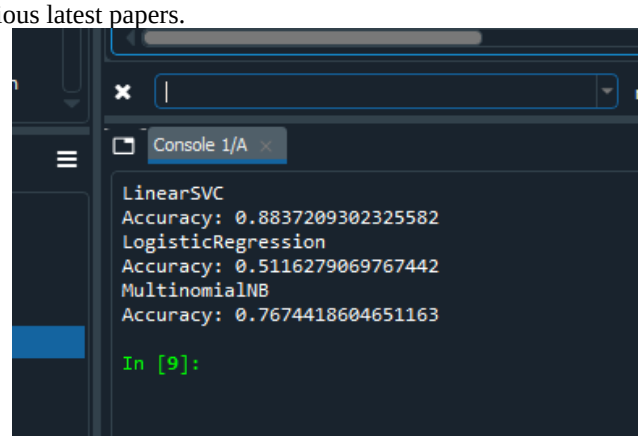


Fig. 16 Accuracy

V. CONCLUSION

The proposed model extracts the necessary information the resumes and finds the best possible match for the job recruitment. The system has the ability to find the best suitable role of a candidate. This proposed system works well for all domains and fields of recruitment. Based on the results and analysis of this research, SVM proves to be the best classification algorithm to be used giving an accuracy of 88%.

VI. FUTURE WORK

Further research is needed to use classification methods with other machine learning techniques. There are a wide number of classification methods that exist which have been tried in the making of this system. The observations and results conclude SVM to be the best approach but using deep learning techniques could further improve the accuracy and performance. The future work would be to use the deep learning techniques to obtain better accuracy and an overall better system.

REFERENCES

[1] Pradeep Kumar Roy, Sarabjeet Singh Chowdhary, Rocky Bhatia, A Machine Learning approach for automation of Resume Recommendation system, Procedia Computer Science, Volume 167, 2020, Pages 2318-2327, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.284.

[2] K. Maji and A. B. Bera, "Developing a Suitability Index Algorithm for Recruitment," 2020 National Conference on Emerging Trends on Sustainable Technology and Engineering Applications (NCETSTEA), 2020, pp. 1-5, doi: 10.1109/NCETSTEA48365.2020.9119944.

[3] V. Yadav, U. Gewali, S. Khatri, S. R. Rauniyar and A. Shakya, "Smart Job Recruitment Automation: Bridging Industry and University," 2019 Artificial Intelligence for Transforming Business and Society (AITB), 2019, pp. 1-6, doi: 10.1109/AITB48515.2019.8947445.

[4] M. V. Belova and A. B. Zhernakov, "Modern Methods of Resume Processing in Recruiting Information Systems," 2018 XVII Russian Scientific and Practical Conference on Planning and Teaching Engineering Staff for the Industrial and Economic Complex of the Region (PTES), 2018, pp. 19-22, doi: 10.1109/PTES.2018.8604263.

[5] Aljuaid, Abdulrahman & Abbod, Maysam. (2020). Artificial Intelligence-Based E-Recruitments System. 144-147. 10.1109/IS48319.2020.9199979.

[6] D. Wang, J. Su and H. Yu, "Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language," in IEEE Access, vol. 8, pp. 46335-46345, 2020, doi: 10.1109/ACCESS.2020.2974101.

[7] D. Deepa, Raaji and A. Tamilarasi, "Sentiment Analysis using Feature Extraction and Dictionary-Based Approaches," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2019, pp. 786-790, doi: 10.1109/I-SMAC47947.2019.9032456.

[8] V. G, H. R and J. Hareesh, "Relation Extraction in Clinical Text using NLP Based Regular Expressions," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019, pp. 1278-1282, doi: 10.1109/ICICICT46008.2019.8993274.

[9] A. Kumbhar, M. Savargaonkar, A. Nalwaya, C. Bian and M. Abouelenien, "Keyword Extraction Performance Analysis," 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2019, pp. 550-553, doi: 10.1109/MIPR.2019.00111.

[10] Y. Zheng, "An Exploration on Text Classification with Classical Machine Learning Algorithm," 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), 2019, pp. 81-85, doi: 10.1109/MLBDBI48998.2019.00023.

[11] V. Bahel, S. Pillai and M. Malhotra, "A Comparative Study on Various Binary Classification Algorithms and their Improved Variant for Optimal Performance," 2020 IEEE Region 10 Symposium (TENSYMP), 2020, pp. 495-498, doi: 10.1109/TENSYMP50017.2020.9230877.

[12] W. Bian, C. Wang, Z. Ye and L. Yan, "Emotional Text Analysis Based on Ensemble Learning of Three Different Classification Algorithms," 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2019, pp. 938-941, doi: 10.1109/IDAACS.2019.8924413.

[13] M. Yusuf and K. M. Lhaksmana, "An Automated Interview Grading System in Talent Recruitment using SVM," 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020, pp. 34-38, doi: 10.1109/ICOIACT50329.2020.9332109.

[14] Rzheuskyi, Antonii & Kutyuk, Orest & Vysotska, Victoria & Burov, Yevhen & Lytvyn, Vasyl & Chyrun, Lyubomyr. (2019). The Architecture of Distant Competencies Analyzing System for IT Recruitment. 254-261. 10.1109/STC-CSIT.2019.8929762.

[15] R. Nimbekar, Y. Patil, R. Prabhu and S. Mulla, "Automated Resume Evaluation System using NLP," 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), 2019, pp. 1-4, doi: 10.1109/ICAC347590.2019.9036842.

[16] Dana Pessach, Gonen Singer, Dan Avrahami, Hila Chalutz Ben-Gal, Erez Shmueli, Irad Ben-Gal, Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming, Decision Support Systems, Volume 134, 2020, 113290, ISSN 0167-9236

[17] Islam, Moinul & Yasmin, Farzana & Arefin, Mohammad & Ayon, Zaber & Ripon, Rony. (2021). An Automated Candidate Selection System Using Bangla Language Processing. 10.1007/978-3-030-68154-8_90.

[18] Ombhase, Manasi & Gogate, Prajakata & Patil, Tejas & Nair, Karan & Hegde, Prof. (2019). Automated Personality Classification Using Data Mining Techniques. 10.13140/RG.2.2.35949.59363.

[19] M. Yusuf and K. M. Lhaksmana, "An Automated Interview Grading System in Talent Recruitment using SVM," 2020 3rd International Conference on Information and Communications Technology (ICOIACT), 2020, pp. 34-38, doi: 10.1109/ICOIACT50329.2020.9332109.

[20] A. Mohamed, W. Bagawathinathan, U. Iqbal, S. Shamrath and A. Jayakody, "Smart Talents Recruiter - Resume Ranking and Recommendation System," 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS), 2018, pp. 1-5, doi: 10.1109/ICIAFS.2018.8913392.

[21] Sanyal, Satyaki & Hazra, Souvik & Ghosh, Neelanjan & Adhikary, Soumyashree. (2017). Resume Parser with Natural Language Processing. 10.13140/RG.2.2.11709.05607.

[22] S. Ahn et al., "A Fuzzy Logic Based Machine Learning Tool for Supporting Big Data Business Analytics in Complex Artificial Intelligence Environments," 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), 2019, pp. 1-6, doi: 10.1109/FUZZ-IEEE.2019.8858791.

[23] T. D. Jayasiriwardene and G. U. Ganegoda, "Keyword extraction from Tweets using NLP tools for collecting relevant news," 2020 International Research Conference on Smart Computing and Systems Engineering (SCSE), 2020, pp. 129-135, doi: 10.1109/SCSE49731.2020.9313024.

[24] D. F. Mujtaba and N. R. Mahapatra, "Ethical Considerations in AI-Based Recruitment," 2019 IEEE International Symposium on Technology and Society (ISTAS), 2019, pp. 1-7, doi: 10.1109/ISTAS48451.2019.8937920.

[25] A. N. Ray and J. Sanyal, "Machine Learning Based Attrition Prediction," 2019 Global Conference for Advancement in Technology (GCAT), 2019, pp. 1-4, doi: 10.1109/GCAT47503.2019.8978285.

[26] A. Hemalatha, P. B. Kumari, N. Nawaz and V. Gajenderan, "Impact of Artificial Intelligence on Recruitment and Selection of Information Technology Companies," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 60-66, doi: 10.1109/ICAIS50930.2021.9396036.

[27] A. Khanfor, A. Hamrouni, H. Ghazzai, Y. Yang and Y. Massoud, "A Trustworthy Recruitment Process for Spatial Mobile Crowdsourcing in Large-scale Social IoT," 2020 IEEE Technology & Engineering Management Conference (TEMSCON), 2020, pp. 1-6, doi: 10.1109/TEMSCON47658.2020.9140085.

[28] G. Gao, J. Wu, Z. Yan, M. Xiao and G. Chen, "Unknown Worker Recruitment with Budget and Covering Constraints for Mobile Crowdsensing," 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), 2019, pp. 539-547, doi: 10.1109/ICPADS47876.2019.00083.

[29] Daryani, Chirag & Chhabra, Gurneet & Patel, Harsh & Chhabra, Indrajeet & Patel, Ruchi. (2020). AN AUTOMATED RESUME SCREENING SYSTEM USING NATURAL LANGUAGE PROCESSING AND SIMILARITY. 99-103. 10.26480/etit.02.2020.99.103.

[30] Islam, Moinul & Yasmin, Farzana & Arefin, Mohammad & Ayon, Zaber & Ripon, Rony. (2021). An Automated Candidate Selection System Using Bangla Language Processing. 10.1007/978-3-030-68154-8_90.

COMPARATIVE ANALYSIS OF VARIOUS APPROACHES FOR SENTIMENT ANALYSIS

Swati Kashyap[#1], Williamjeet Singh[#2]
[#]*Department of Computer Science and Engineering, Punjabi University Patiala*
[1]Swatikashyap510@gmail.com
[2]williamjeet@gmail.com

ABSTRACT— Sentiment analysis (SA), is the process of obtaining and evaluating people's thoughts, feelings, attitudes, and perceptions about various topics, products, and services. Sentiment analysis poses as a powerful tool for businesses, governments, and researchers to extract and analyze public mood and views, gain business insight, and make better decisions. This paper provides a comprehensive examination of sentiment analysis methodologies, problems, and trends in order to provide academics with a global overview of sentiment analysis and related topics. The paper discusses the various uses of sentiment analysis as well as the general procedure for performing this assignment. The report then examines, compares, and investigates the various approaches in order to gain a comprehensive understanding of their benefits and downsides. Following that, the difficulties of sentiment analysis are examined in order to elucidate future directions.

KEYWORDS— Sentiment Analysis, Approaches for SA, Applications of SA, NLP (Natural Language Processing), SA Level

I. INTRODUCTION

Sentiment analysis is a method used to extract, transform and interpret the opinion of text into a positive, negative or natural feeling using the Natural Language Processing (NLP)[1].Positive, negative and neutral models focus on emotions and feelings (scary, joyful, sad, etc.), emergency (e.g., urgent and not urgent) as well as intents (interested v. not interested)[2]. Using sentimental analysis, you may assess how customers feel about various company sectors without reading a wide range of customer comments at once. Today, not only researchers, but also corporations, governments and organizations have become well-known for sentiment analysis [3].

The increasing Internet use has created the Web the world's most significant and universal information source. Millions of individuals share their ideas and feelings in forums, blogs, wikis and social networks. The increasing Internet use has created the Web the world's most significant and universal information source. Millions speak in forums, blogs, wiki, social networks and other digital resources and express their opinion and feelings [4]. These views and feelings are extremely significant to our everyday lives and hence the user produced data need to be analyzed for the automatic monitoring of Public opinion and decision-making support. That's why in the past decades and a half research communities have been showing increasing interest in the subject of sentiment analyses. Sentiment analysis, which has lately increased massively in publications on sentimental analysis and Opinion mining, has been the fastest growing and most active study topic since 2004 [5].This paper gives brief introduction to sentiment analysis , compared the approaches used for sentiment analysis , review to application areas and challenges for Sentiment Analysis.

A. Level of Sentiment Analysis

The task of sentiment analysis has been investigated at several levels. However, sentiments and opinions can be detected mainly at the document level, sentence level, or the aspect level. There are various stages examined in the task of sentiment analysis. But mainly at document level, sentence level or aspect level, feelings and views can be detected.

1)Aspect Level: This level takes out a finely grained analysis since it seeks to develop feelings about the precise characteristics of entities. For example, the phrase, "The iPhone 11 camera is great'." the review is about "camera" which is an aspect of the iPhone 11' , and the review is favorable i:e positive. Therefore the job at this level helps to determine just what individuals like or don't like. According to [6] Aspect extraction, which might be implicit or explicit, is the primary job for sentiment analysis.

2)Sentence Level: The focus is on the sentence at this level. The main goal is to figure out whether the sentence is positive, negative, or neutral. However, in order to accomplish this, the phrase must be categorized as objective, presenting facts, or subjective, expressing feelings and ideas. This level of analysis was approached in a number of ways. It is closely related to the classification of subjectivity that categorizes phrases called impartial phrases representing truthful facts from phrases, i.e. subjective phrases expressing subjective views and opinions[7] .

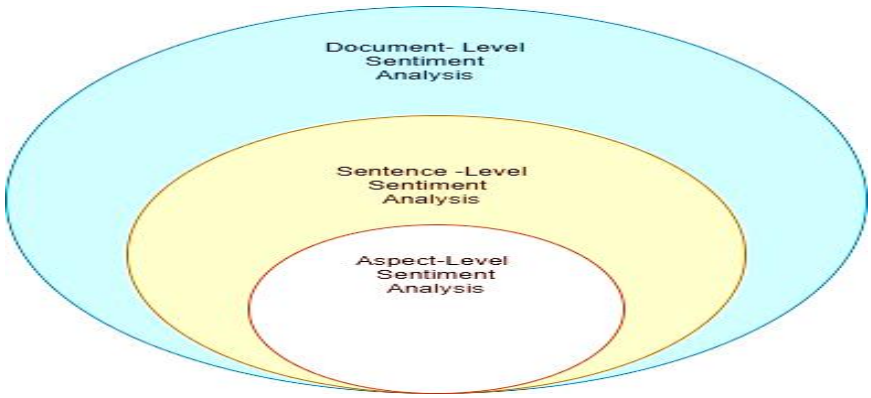


Fig. 3 Levels of Sentiment Analysis

3)Document Level: The procedure at this level tries to classify whether a document as a whole reflects a negative or positive attitude or viewpoint [8]. e:g Framework determines if a text file comprising reviews of only one product reflects an overall favorable or negative opinion about the product by calculating whether the entire review expresses an overall positive or negative opinion about the product.

B. Sentiment Analysis Process:

Every day, the amount of textual material available on the internet expands. Textual material is getting increasingly difficult to harvest and search. Text is the most used data type on the internet since it is simple to create and distribute. However, collecting relevant information from a large ocean of stuff is more challenging.

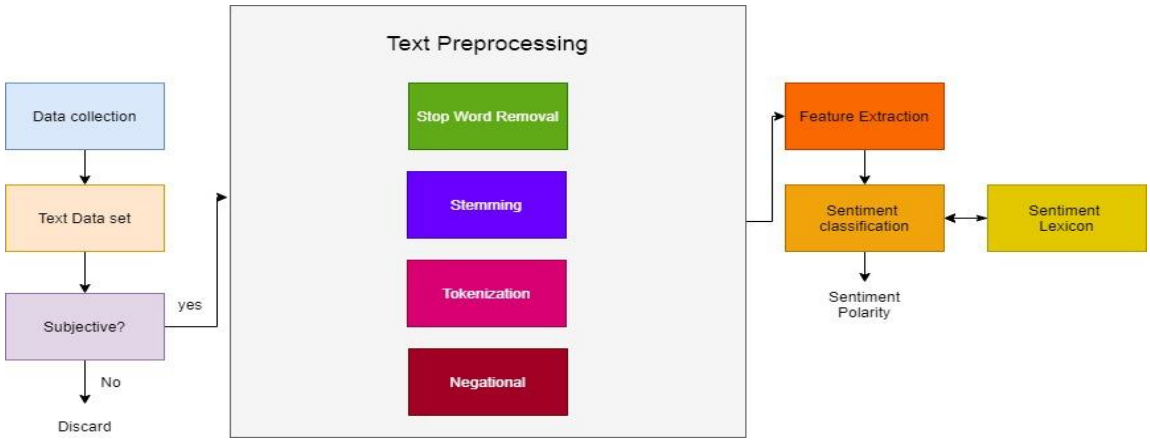


Fig. 4 Process of Sentiment Analysis

C. Our Contribution:

Because of the following factors, this survey makes a substantial contribution.

- In this article, we provide a thorough examination of the many approaches to sentiment analysis that are currently in use.
- This study summarizes the quantity of research done in various fields, the resources used, and the prospective applications of sentiment analysis in order to gain a better understanding of the topic.
- In addition to the current research in the field, this work demonstrates potential problems and expansions.

II. RESEARCH METHOD

This section contains the review paper's taxonomy as well as other research questions relevant to this study, the answers to which are given in section 6 (discussion).

A. Taxonomy of Research:

This part briefly reviews the many components of sentiment analysis, such as the level and procedures of sentiment analysis, sentiment analysis methodologies, and applications, which are covered in greater depth in the next sections of this work. The taxonomy is depicted in Fig 3, which comprises categories such as level, approaches, sentiment analysis applications, and their subcategories.

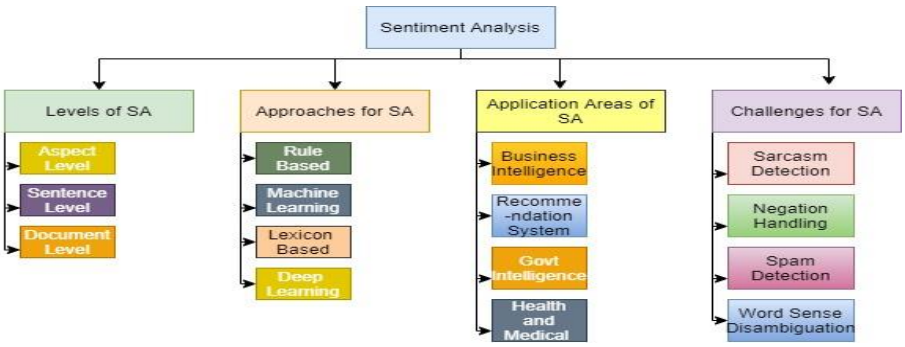


Fig. 5 Organization of the Paper

B .Research Question:

RQ1. Which Approach of sentiment Analysis done feature engineering on its own?
RQ2.Why we prefer deep Learning over Machine Learning?
RQ3. Which Application of Sentiment Analysis is used to analyze customers' impressions of products or services?
RQ4. What are the different Sentiment Analysis challenges?

III. SENTIMENT ANALYSIS APPROACHES

Sentiment Analysis have numerous approaches and some of them are discussed in this section with their advantages and limitations.

A. Rule-Based Approach:

Rule-based classification refers to any scheme that creates classification based on IF and THEN rules. As a result, to accomplish sentiment classification, the classifiers in this technique rely on a set of rules. A rule can be expressed as LHS RHS, where the LHS represents the rule's antecedent or a collection of conditions on the feature set expressed in DNF (Disjunctive Normal Form), and the RHS represents the rule's conclusion or consequence (class label) if the LHS is satisfied [9]. A linear classifier is a vector of numerical input features of real values [10].To determine the sentiment of a sentence, the basic technique is rules-based and uses a dictionary of words labeled by sentiment. Sentiment ratings should frequently be combined with other rules to remove negations, sarcasm, or dependent clauses from sentences.

Because the sequential merger of words is not considered in rule-based systems, they are extremely basic. Superior processing methods can be used, and the most recent rules can be applied to support newer vocabularies and forms of expression.

The addition of new rules, on the other hand, can have an impact on previously achieved outcomes and make the entire system highly convoluted. Rule-based systems necessitate continuous fine-tuning and maintenance, which will necessitate funding at regular periods.

B. Machine Learning Approach:

Machine learning strategies rely on machine learning algorithms rather than human designed rules. Machine learning is a technique for parsing data, learning from it, and making intelligent judgments based on what it has learnt. A sentiment analysis task is typically modeled as a classification problem, in which the classifier receives text data and assigns it to one of three classes: positive, negative, or neutral.

On the basis of test cases used in the training process, our model learns to compare a particular input data to the associated output data in the training process shown in Figure 4(a). The textual input is transformed into a features vector by the feature extractor. • In the prediction process represented in Figure 4(b), the feature extractor translates concealed textual inputs into feature vectors, which are then fed into the algorithm to generate a model.

The model is then fed these vectors, which generate prediction tags for the relevant vectors. Machine learning have different algorithms which are discussed in Table1

TABLE XII
DEPICTS THE PERFORMANCE OF DIFFERENT MACHINE LEARNING ALGORITHMS WITH THEIR ACCURACY

Study	Year	Title	Dataset	Sentiment analysis method	Results
[11]	2013	Sentiment Analysis in Twitter using Machine Learning Techniques	1200 Twitter posts about electronic product	SVM, Naive Bayes, Maximum Entropy	90%, 89.5%, 90%
[12]	2015	Twitter Sentiment Classification Using Naïve Bayes Based on Trainer Perception	Twitter tweets	Naïve Bayes	90%
[13]	2016	SemEval-2016 Task 4: Sentiment Analysis in Twitter	Twitter Dataset	SVM	84.5%
[14]	2016	A Topic-based Approach for Sentiment Analysis on Twitter Data	Twitter Dataset	SVM	74.09%

[15]	2017	Polarity Shift Detection Approaches in Sentiment Analysis: A survey	Product Review	Lexicon-based and Supervised Machine Learning-based	84.6%
[16]	2017	A Sentiment Analysis Method of Short Texts in Microblog	COAE2014(BBC DataSet)	Language Technology Platform (LTP) for dependency syntax analysis	86.5%
[17]	2017	Document Level Sentiment Analysis from News	BBC News Dataset	Machine Learning Approaches	57.7%
[18]	2017	A Feature Based Approach for Sentiment Analysis using SVM and co-reference	Resolution Training Dataset of Product Review	SVM & co-reference Resolution	73.6%
[19]	2018	Aspect-Level Sentiment Analysis on E-Commerce Data	Amazon Customer Review Data	Naïve Bayes , SVM	90.423 %, 83.43%
[20]	2018	Sentiment Analysis of Twitter Corpus Related to Artificial Intelligence Assistants	Reviews of Electronic product	Valence Aware Dictionary and Sentiment Reasoner (VADER)	87.4%
[21]	2020	Twitter Sentiments Analysis Using Machine Learning Methods	Twitter Dataset	Naïve Bayes classifier, SVM ,Maximum Entropy method	86% 74.6% 82.6%

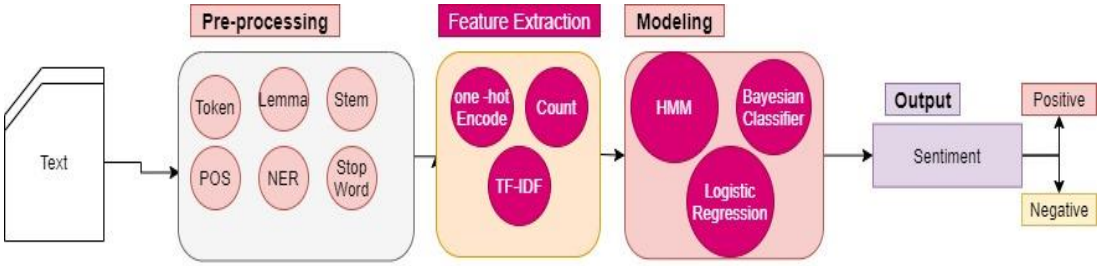


Fig. 6 Machine Learning Approach

C. Lexicon Based Approach

One of the two main approaches used for sentiment analysis is the lexicon-based (also known as knowledge-based) approach, which requires a lexical resource called an opinion lexicon (a predefined list of words) that associates words to their semantic orientation as negative or positive words using scores[22]. For example, a score could be a basic polarity value like +1, 1 or 0 for positive, negative, or neutral terms, or a value expressing sentiment strength or intensity. Calculating the semantic orientation values of the words that make up a document yields its final orientation. The sentiment values from the lexicon are assigned to each element when the document is tokenized into single words or micro phrases. A formula or algorithm (e.g., sum and average) can be used to determine the overall sentiment of a document. At the sentence and feature level sentiment analysis, the lexicon-based technique is highly useful. It can be called an unsupervised strategy because it does not require any training data. The fundamental problem with this technique, on the other hand, is domain dependency; because words can have various meanings and senses, therefore a positive word in one domain may not be positive in another. Given the word ‘‘small" and the sentences ‘‘The TV screen is too small". and ‘‘This camera is very small"., the word ‘‘small" in the first sentence is negative, because people generally prefer wide screens, whereas it is positive in the second sentence, because if the camera is small, it will be easy to carry. The design of a domain-specific sentiment lexicon or the use of a lexicon adaption strategy can be used to avoid this difficulty.

D. Deep Learning Approach

Deep learning creates an "artificial neural network" that can learn and make intelligent judgments on its own by layering algorithms. The use of ANNs-based deep learning (DL) for sentiment analysis has lately gained popularity. DL is a new branch of machine learning that includes supervised and unsupervised methods for learning feature representation. . Deep learning refers to neural networks with numerous layers of perceptron inspired by our brain. As a result, this architecture allows for the training of more complicated models on a considerably bigger dataset, resulting in state-of-the-art outcomes in a variety of application domains, ranging from computer vision and audio recognition to natural language processing [23]. CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), and DBN (Deep Belief Networks) are only a few of the neural network models included in DL [24]some of them discussed with their performances in Table 2. These models do not require pre-defined features chosen by an engineer; instead, they can learn sophisticated features from the dataset on their own. On the other hand, they are complex and computationally expensive. Deep learning techniques for sentiment analysis have been the subject of several studies [25] & [23].

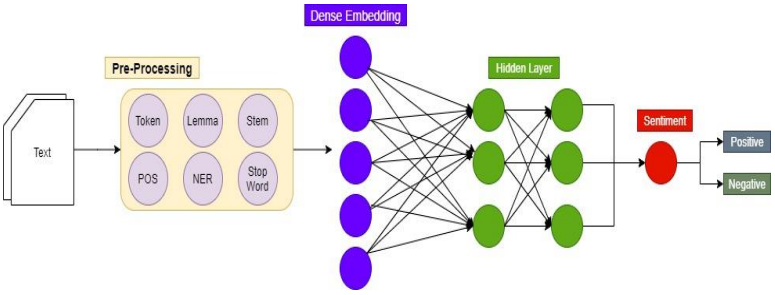


Fig. 7 Deep Learning Process without Feature Engineering

Much of the present research in this subject is centered on machine learning-based approaches. Deep learning is one of the most rapidly expanding fields of machine learning. As a result, deep learning is being studied as a subject because it offers potential advantages over other approaches due to the following:

- Deep learning has a number of advantages over traditional machine learning algorithms, including the ability to perform feature engineering on its own .A algorithm examines the data for features that correlate and then combines them to enhance faster learning without being expressly prompted to do so.
- When the data amount is high, Deep Learning outperforms conventional techniques. Traditional Machine Learning techniques, on the other hand, are preferable when dealing with tiny amounts of data.
- When it comes to complicated issues like picture classification, natural language processing, and speech recognition, deep learning really shines.

Table II
DEPICTS THE DEEP LEARNING ALGORITHMS ASSOCIATED WITH THEIR ACCURACY

Study	Year	Title	Dataset	Sentiment analysis method	Results
[26]	2018	LSTM with sentence representations for Document-level Sentiment Classification	Yelp 2014 dataset, Yelp 2015 dataset , IMDB	LSTM	Yelp 2014: 63.9 % Yelp 2015: 63.8% IMDB: 44.3 %
[27]	2018	Sentiment Analysis using Neural Networks: A New Approach	Product Data Review Twitter Data	Convolutional Neural Network	74.15%
[28]	2017	Sentiment classification: Feature selection based approaches versus deep learning	IMDB Dataset, Sentiment140,Amazon Multi-domain	CNN, LSTM, CNN + LSTM	IMDB: 89.1% Sentiment140: 71.5% Multi-domain dataset: 85%
[29]	2017	Weakly-supervised Deep Embedding for Product Review Sentiment Analysis	Review from Amazon on digital cameras, cell phones and lap	CNN + LSTM	CNN: 87.7 LSTM: 87.9
[30]	2020	Sentiment Analysis using Machine learning and Deep Learning	Twitter data	Machine learning and deep learning	81 % to 90%

TABLE III
ADVANTAGES AND LIMITATIONS OF ALL THE APPROACHES

Approaches	Advantages	Limitations
Rule Based Approach	Data for training isn't necessary. Extreme accuracy. It's a wonderful technique to collect data since you can set up the system with rules and then let data stream in spontaneously as people use it.	Lower recall. It is difficult and time-consuming to list all of the rules.
Machine Learning Approach	There is no need for a dictionary. Show good classification accuracy.	Many times, a classifier trained on textual input in a single field does not work with different fields.
Lexicon Based Approach	It is not necessary to name the information and the learning method.	Requires immense semantic assets that are normally unavailable
Deep Learning Approach	High Accuracy The reliability of natural data variances is automatically learned.	To do more than previous techniques, it takes a significant volume of data

IV. APPLICATIONS OF SENTIMENT ANALYSIS:

Sentiment analysis is particularly helpful in a wide range of application areas from consumer opinion [31] identification to mental health monitoring based on social media posts for patient. Moreover, new technologies like Big

Data, Cloud Computing, and Block chain [32] have expanded the field of applications providing endless possibilities for the analysis of feelings in virtually every area.

A. Business Intelligence

People nowadays use a variety of online forums for social interaction, as well as social media alternatives like Twitter and Facebook. This real-time interaction between media consumers takes place. These communications provide a wide range of business intelligence opportunities. A business process is a collection of interconnected, ordered activities or processes that result in a specific service or product that fulfils a specified goal for a specific customer or customers. The application of sentiment analysis in the field of business intelligence has numerous benefits. For example, organisations can use sentiment analysis data to enhance products, examine client feedback, or implement a new marketing plan [33]. The most typical use of sentiment analysis in the field of business intelligence is analyzing customers' impressions of products or services. Social media and business intelligence have become inextricably linked. Both social networks and business intelligence make it possible to reach a big number of customers/users at the same time, resulting in a diversity of data that may be used to uncover hidden knowledge [34]. These studies, however, are not limited to product producers; customers may use them to compare items and make more informed decisions.

B. Recommendation system

One of the most common and well-understood applications of big data is recommender systems. A recommender system, as a specialized information filtering system, tries to make predictions based on user preferences and interests. A recommender system is an algorithm that seeks to recommend relevant goods to consumers (movies, music, or purchases) [35]. For some industries, a good recommender system can bring in a lot of money. As a result, the use of sentiment analysis in such systems can help them produce better *recommendations* [36] &[37] . Their use has been widespread, with interesting use-cases ranging from product recommendations to movies, music, books, research articles, search queries, social tags, experts, people, jokes, restaurants, financial services, and even twitters followers[38] .

C. Government intelligence

People make comments on a variety of topics, including politics, religion, and social issues, in addition to items and services. As in the work of Georgiadou et al.[39], who used sentiment analysis of Twitter posts to investigate and aggregate public sentiment toward Brexit outcomes, using sentiment analysis to identify opinions on government policies or other similar issues is very helpful for monitoring possible public reaction on implementation of certain policies. The Sentiment Political Compass (SPC), a data-driven paradigm that cfharacterizes newspapers' attitudes toward political parties, was used by Falck et al.[40] to measure proximity between newspapers and political parties.

D. Healthcare and medical domain

The use of sentiment analysis in the medical field has recently sparked a lot of attention. This application helps healthcare providers to collect and evaluate data on diseases, adverse drug reactions, epidemics, and patient moods [2019 health] in order to improve healthcare services. Anu J Nair et.al [41]. More than 25000 tweets with different hashtags related to COVID -19 were included in the dataset. Initially, they began by gathering information on how a wide number of people felt about the pandemic situation. They used different algorithms on the same dataset to see how accurate each approach was. For training and testing reasons, the data set was divided into 75 and 25%. This equates to 18,750 hours of instruction and 6,250 hours of assessment. First, they used sentiment analysis with logistic regression to determine the dataset's polarity score. They take 75 and 25% because it is beneficial to provide a larger dataset for training because it will result in more exact results when testing with the testing dataset. They discovered that out of 25,000 data points, 15,340 had favorable feelings or sentiments concerning COVID time, 9,620 had negative feelings or sentiments, and 40 had neutral feelings or sentiments. Twitter data can be used as a source of health-care data analysis by public health professionals, clinical researchers, and others [42]. A lot of effort has gone into the tweets, which have yielded valuable information. The common opinions and relationships among diabetes, diet, exercise, and obesity-related tweets were retrieved and analyzed using language analysis.

V. SENTIMENT ANALYSIS CHALLENGES:

A. Sarcasm detection

Sarcasm is described as "the behavior of saying or writing the reverse of what someone intends, or of speaking in a way meant to make someone else feel stupid or show them that he is angry," according to the Macmillan English dictionary[43]. The difficulty of sarcasm in sentiment analysis occurs when someone writes something nice but really means something negative, or vice versa, making sentiment analysis more difficult. In our daily lives, we employ a lot of sarcastic terms. As a result, sarcasm detection is gaining popularity as a way to solve the problem of obtaining deceptive attitudes by automatically detecting sarcastic expressions in a text. Sarcasm recognition is a difficult NLP task due to the intricacy and ambiguity of sarcasm [44]. Many methods for detecting sarcasm have been proposed[45]. The voice and text communities have made the most significant contributions. As a special case of sentiment analysis, sarcasm detection is dependent on the performance of the sentiment analysis task. Figure 6 shows an example of a caustic tweet that was misunderstood by the internet bot and hence inappropriately responded to.

Figure 8: Misinterpreting Tweet

Jain et al. [46] a mash-up of English and Indian native language, deep learning was utilized to identify sarcasm in real time (Hinglish). A bidirectional LSTM with a softmax attention layer and a convolutional neural network are combined in their suggested model. The semantic context vector for English features was learned from Glove word representation and forwarded to CNN using the softmax attention layer. HindiSenti (Hindi SentiWordNet) feature vector and supplementary punctuation-based features were merged with the CNN model. With a classification accuracy of 92.71 percent, our model surpasses baseline deep learning algorithms.

B. Negation Handling

Negations are crucial in linguistics because they influence the polarity of other words. Words like no, not, and shouldn't are examples of negatives. When a negation appears in a sentence, it's critical to figure out which words are affected by this term. The scope of negation can be confined to the next word after the negation, or it can be extended to include further words after the negation. For example, the scope of negative is solely confined to the following word after negation in the statement "this mobile is not nice yet it works well." In another sentence, the scope of negation is extended until the end of the sentence, as in ”the battery does not work for a long time.” These examples demonstrate how the scope of negation varies depending on linguistic elements like as conjunctions, punctuation marks, and the part of speech (POS) of negation, among others. Lazib et al. [47] suggested a syntactic path-based hybrid neural network for negation scope detection. In both constituency and dependency parse trees, the CNN model was utilized to capture relevant syntactic properties between the token and the cue along the shortest syntactic path, while the Bi-LSTM learned the context representation throughout the sentence in both forward and backward directions. Their model received an F-score of 90.82 percent.

C. Spam Detection

Customers' online reviews on product quality have a significant impact on electronic buying. Some imposters see this as a chance to write spam reviews in order to improve or deteriorate a product's reputation. As a result, detecting those reviews is critical for safeguarding consumers' interests [48]. In the subject of sentiment analysis, spam identification is critical. Spam and phone reviews can harm brands' reputations and artificially affect consumers' perceptions of products, services, corporations, and other entities since online opinions influence consumer buying decisions [49]. Because there is no obvious difference between reviews, developing spam detection system that can recognize fraudulent reviews among a large number of reviews is a difficult undertaking. Among the systems proposed to execute the work of spam identification, Saumya and Singh[48] devised a system that effectively employs three features: sentiment of review and its comments, content based factor, and rating deviation. This method uses the comment data to determine whether the review is spam or not. The authors combined the labeled data with a machine learning model to categorize the remaining unlabeled data, as well as two over-sampling strategies to make the class similar, as the quantity of spam reviews is typically significantly lower than the number of genuine reviews. Their system received a 91 percent F-score.

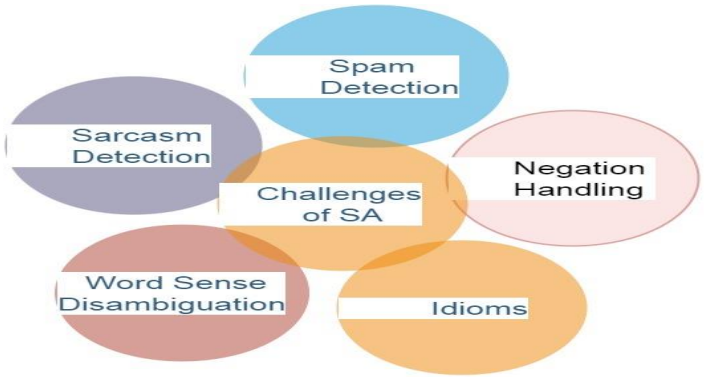


Figure 9: Challenges for sentiment analysis

D. Word sense disambiguation (WSD)

The goal of Word Sense Disambiguation (WSD) is to figure out what a word means in context [50]. A word can have multiple meanings, and the meaning of a term can vary depending on the context and area in which it is used. The goal of word sense disambiguation is to figure out which meaning of a word has been utilized in a phrase. When used with a television, for example, the word "curved" has a positive connotation, but when used with a mobile phone, it might have a negative connotation. As a result, determining a word sense from a sentence is extremely difficult. Wang et al. [51] suggested a knowledge-based solution that relies on the well-known lexicon WordNet to solve this difficult problem. Using Latent Semantic Analysis (LSA) and PageRank, this method represents the WSD problem with semantic space and semantic path concealed behind a given language. The experimental results show that this strategy is effective, as it has achieved good outcomes. Similarly, word polarity disambiguation (WPD) is a difficult problem to solve. The goal of WPD is to determine the polarity of sentiment-ambiguous words in a given context. Xia et al. [52] used a Bayesian model and opinion-level features to solve this problem. They looked at the context of the level by establishing intra- and inter-opinion aspects. The Bayesian model was utilized to improve the effectiveness of the opinion-level features and to resolve polarity in a probabilistic manner.

E. Idioms

The primary principle behind automatically interpreting an idiom's figurative meaning is to interpret its dictionary definition's literal meaning instead [53]. A figure of speech is not always understood by machine learning techniques. For example, because the algorithm understands things in the literal sense, an expression like "not my cup of tea" will perplex it. As a result, when an idiom is used in a remark or a review, the algorithm may misinterpret the sentence or even overlook it. To solve this issue, a sentiment analysis platform must be trained to recognize idioms. When dealing with numerous languages, the challenge multiplies. Only if the neural networks in an emotion mining API are trained to recognize and interpret idioms can this problem be handled with sentiment analysis accuracy. Idioms are mapped to nouns that represent emotions such as wrath, joy, determination, success, and so on, and the models are then trained accordingly. To put it another way, only then can a sentiment analysis technology provide reliable insights from such content.

VI. DISCUSSION

In this article, we presented a brief review on sentiment Analysis, comparison of the approaches used in sentiment analysis, different applications area of sentiment analysis and different challenges to sentiment analysis. We discussed rule based is a practical method for analyzing text that does not require any training or the use of machine learning models. This method yields a set of principles based on which the text is classified as positive, negative, or neutral. For some areas, lexicon-based techniques produce higher outcomes. They can't, however, recognize new entities that aren't in the dictionary. Entity detection is performed by rule-based systems, which generate rules manually or automatically. A chunk of text communication is typically represented as a bag of words in lexicon-based techniques. Following this representation of the message, all positive and negative words or phrases inside the message are assigned sentiment values from the lexicon. Machine learning and deep learning are terms that are sometimes used interchangeably. Deep learning is a type of machine learning that is more advanced than traditional machine learning. When basic machine learning makes a mistake, it requires human intervention to remedy the problem - to adjust the output and “force” the model to learn. However, thanks to its complex algorithm chain, a neural network may learn to correct itself in deep learning. **RQ1**Deep Learning approaches have proven to be quite effective in sentiment analysis. It has superior performance than standard feature-based approaches due to automatic feature extraction, rich representation capabilities, and automatic feature extraction. RQ3: The most significant distinction between deep learning and regular machine learning is how well it performs when data scales up. Deep learning techniques do not perform well when the data is small. This is due to the fact that deep learning algorithms require a vast amount of data to fully comprehend it. Traditional machine learning algorithms, on the other hand, with their handmade rules, win in this circumstance. In next section we have discussed some of the application areas of Sentiment analysis like business intelligence, recommendation system, Government intelligence and health domain. RQ4The most typical use of sentiment analysis in the field of business intelligence is analyzing customers' impressions of products or services. A recommender system, as a specialized information filtering system, tries to make predictions based on user preferences and interests and seeks to recommend relevant goods to consumers (movies, music, or purchases). Government intelligence includes variety of topics, including politics, religion, and social issues, in addition to items and services and gather review from there. Health and Medical Domain helps healthcare providers to collect and evaluate data on diseases, adverse drug reactions, epidemics, and patient moods in order to improve healthcare services. In this pandemic of covid 19 number of paper has been published some of them take review how different peoples feels about the situations, some talking about government decisions to cope with covid 19 and many more.

Then we discussed about some of major challenges for sentiment analysis. People talk about everything and anything under the sun, and it's nearly hard for a computer to understand their sentiments and ideas on particular issues.RQ2 Sarcasm detection, negation handling, spam detection, word sense disambiguation and idioms are some of the major challenges that occur in sentiment analysis. Numbers of researchers proposed different models to tackle with all challenges which we were discussed in section 5.

VII. CONCLUSION and FUTURE WORK

The many applications, methodologies, and challenges for sentiment analysis were discussed in this study. We've gone over the benefits and drawbacks of many methodologies for sentiment analysis, including rule-based, lexicon-based, machine learning, and deep learning. We've also spoken about how sentiment analysis can be used in a variety of ways. Our focus for future study will be on the evolving prospects and potential enhancements in multimodal sentiment analysis. Future attention is also required for the use of hybrid classification algorithms

ACKNOWLEDGEMENT

Dr. Williamjeet Singh is working as an Assistant Professor in Department of Computer Science and Engineering at Punjabi University, Patiala, Punjab, India. He achieved his BTech from BBSBEC, Fategarh Sahib under Punjab Technical University in 2005.He Completed his MTech (CSE) degree from Punjabi University, Patiala in the year 2007. He was awarded PhD degree in the year 2015 in the faculty of Engineering and Technology from Punjabi University.. Sign Language, Speech Recognition, Cellular Networks, Algorithms, Speech Technology, Data Mining, and Sentiment Analysis are among his research interests. He has over 11 years of expertise in teaching and research. He has numerous research articles published in prestigious national and international conferences and journals.

Swati Kashyap is a student of MTech (CSE) in Punjabi University Patiala, Punjab, India. She has earned her BTech in 2019 from Punjab Technical University Jalandhar, Punjab, India. Her area of interest is Sentiment Analysis, Machine Learning and Indian Sign Language.

REFERENCES

[1] Z. Drus and H. Khalid, “Sentiment analysis in social media and its application: Systematic literature review,” *Procedia Comput. Sci.*, vol. 161, pp. 707–714, 2019, doi: 10.1016/j.procs.2019.11.174.

[2] P. Sudhir and V. D. Suresh, “Comparative Study of Various Approaches, Applications and Classifiers for Sentiment Analysis,” *Glob. Transitions Proc.*, 2021, doi: 10.1016/j.gltp.2021.08.004.

[3] J. F. Sánchez-Rada and C. A. Iglesias, “Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison,” *Inf. Fusion*, vol. 52, no. December 2018, pp. 344–356, 2019, doi: 10.1016/j.inffus.2019.05.003.

[4] F. Javier *et al.*, “A Brief Review on the U se of Sentiment Analysis Approaches in Social Networks.”

[5] M. V Mäntylä, D. Graziotin, and M. Kuutila, “The evolution of sentiment analysis — A review of research topics , venues , and top cited papers,” *Comput. Sci. Rev.*, vol. 27, pp. 16–32, 2018, doi: 10.1016/j.cosrev.2017.10.002.

[6] M. Tubishat, N. Idris, and M. A. M. Abushariah, “Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges,” *Inf. Process. Manag.*, vol. 54, no. 4, pp. 545–563, 2018, doi: 10.1016/j.ipm.2018.03.008.

[7] C. Du, M. Tsai, and C. Wang, “BEYOND WORD-LEVEL TO SENTENCE-LEVEL SENTIMENT ANALYSIS FOR FINANCIAL REPORTS Research Center for Information Technology Innovation , Academia Sinica , Taiwan Department of Computer Science , National Chengchi University , Taiwan MOST Joint Research Center,” pp. 1562–1566, 2019.

[8] O. Alqaryouti, N. Siyam, A. A. Monem, and K. Shaalan, “Aspect-based sentiment analysis using smart government review data,” *Appl. Comput.Informatics*, 2019, doi: 10.1016/j.aci.2019.11.003.

[9] H. Sankar and V. Subramaniyaswamy, “Investigating sentiment analysis using machine learning approach,” *Proc. Int. Conf. Intell. Sustain. Syst. ICISS 2017*, no. Iciss, pp. 87–92, 2018, doi: 10.1109/ISS1.2017.8389293.

[10] A. Abdullah, Q. Aqlan, B. Manjula, and R. L. Naik, *A Study of Sentiment Analysis : Concepts , Techniques , and Challenges*. Springer Singapore.

[11] M. S. Neethu and R. Rajasree, “Sentiment analysis in twitter using machine learning techniques,” *2013 4th Int. Conf. Comput. Commun. Netw. Technol. ICCCNT 2013*, 2013, doi: 10.1109/ICCCNT.2013.6726818.

[12] M. N. M. Ibrahim and M. Z. M. Yusoff, “Twitter sentiment classification using Naive Bayes based on trainer perception,” *2015 IEEE Conf. e-Learning, e-Management e-Services, IC3e 2015*, pp. 187–189, 2016, doi: 10.1109/IC3e.2015.7403510.

[13] P. Nakov, A. Ritter, S. Rosenthal, and F. Sebastiani, “SemEval-2016 Task 4 : Sentiment Analysis in Twitter,” pp. 1–18, 2016.

[14] P. FICAMOS and Y. LIU, “A Topic based Approach for Sentiment Analysis on Twitter Data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 12, pp. 201–205, 2016, doi: 10.14569/ijacsa.2016.071226.

[15] S. Zirpe, “Polarity Shift Detection Approaches in Sentiment,” *Int. Conf. Inven. Syst. Control Polarity*, pp. 1–5, 2017, doi: 10.1109/ICISC.2017.8068737.

[16] J. Li and L. Qiu, “A Sentiment Analysis Method of Short Texts in Microblog,” 2017, doi: 10.1109/CSE-EUC.2017.153.

[17] V. S. Shirsat, “Document Level Sentiment Analysis from News Articles,” *2017 Int. Conf. Comput. Commun. Control Autom.*, pp. 1–4, 2017.

[18] M. H. Krishna, K. Rahamathulla, and A. Akbar, “A feature based approach for sentiment analysis using SVM and coreference resolution,” *Proc. Int. Conf. Inven. Commun. Comput. Technol. ICICCT 2017*, no. Icicct, pp. 397–399, 2017, doi: 10.1109/ICICCT.2017.7975227.

[19] S. Vanaja and M. Belwal, "Aspect-Level Sentiment Analysis on E-Commerce Data," in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, Jul. 2018, no. Icirca, pp. 1275–1279, doi: 10.1109/ICIRCA.2018.8597286.

[20] C. W. Park and D. R. Seo, "Sentiment Analysis of Twitter Corpus Related to Artificial Intelligence Assistants," *2018 5th Int. Conf. Ind. Eng. Appl.*, pp. 495–498, 2018.

[21] L. Mandloi, "Twitter Sentiments Analysis Using Machine Learninig Methods," pp. 1–5, 2020.

[22] A. Jurek, M. D. Mulvenna, and Y. Bi, "Improved lexicon-based sentiment analysis for social media analytics," *Secur. Inform.*, vol. 4, no. 1, 2015, doi: 10.1186/s13388-015-0024-x.

[23] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, pp. 1–25, 2018, doi: 10.1002/widm.1253.

[24] W. Li, F. Qi, M. Tang, and Z. Yu, "Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification," *Neurocomputing*, vol. 387, pp. 63–77, 2020, doi: 10.1016/j.neucom.2020.01.006.

[25] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: a review," *Artif. Intell. Rev.*, vol. 53, no. 6, pp. 4335–4385, 2020, doi: 10.1007/s10462-019-09794-5.

[26] G. Rao, W. Huang, Z. Feng, and Q. Cong, "LSTM with sentence representations for document-level sentiment classification," *Neurocomputing*, vol. 308, pp. 49–57, 2018, doi: 10.1016/j.neucom.2018.04.045.

[27] S. Dhar, "Sentiment Analysis using Neural Networks : A New Approach," *2018 Second Int. Conf. Inven. Commun. Comput. Technol.*, no. Icicct, pp. 1220–1224, 2018.

[28] A. K. Uysal and Y. L. Murphey, "Sentiment Classification: Feature Selection Based Approaches Versus Deep Learning," *IEEE CIT 2017 - 17th IEEE Int. Conf. Comput. Inf. Technol.*, pp. 23–30, 2017, doi: 10.1109/CIT.2017.53.

[29] W. Zhao *et al.*, "Weakly-supervised deep embedding for product review sentiment analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 185–197, 2018, doi: 10.1109/TKDE.2017.2756658.

[30] Y. Chandra and A. Jana, "Sentiment Analysis using Machine Learning and Deep Learning," *2 7th Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 5–8, 2020.

[31] S. Kumar, M. Yadava, and P. P. Roy, "Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction," *Inf. Fusion*, vol. 52, no. October 2018, pp. 41–52, 2019, doi: 10.1016/j.inffus.2018.11.001.

[32] J. Frizzo-Barker, P. A. Chow-White, P. R. Adams, J. Mentanko, D. Ha, and S. Green, "Blockchain as a disruptive technology for business: A systematic review," *Int. J. Inf. Manage.*, vol. 51, no. October, pp. 0–1, 2020, doi: 10.1016/j.ijinfomgt.2019.10.014.

[33] R. Cobos, F. Jurado, and A. Blazquez-Herranz, "A Content Analysis System That Supports Sentiment Analysis for Subjectivity and Polarity Detection in Online Courses," *Rev. Iberoam. Tecnol. del Aprendiz.*, vol. 14, no. 4, pp. 177–187, 2019, doi: 10.1109/RITA.2019.2952298.

[34] S. Alcabnani, M. Oubezza, and J. Elkafi, "A business intelligence model to analyze consumer opinions on social networks using machine learning techniques," *2020 IEEE 2nd Int. Conf. Electron. Control. Optim. Comput. Sci. ICECOCS 2020*, 2020, doi: 10.1109/ICECOCS50124.2020.9314548.

[35] Z. Y. Khan, Z. Niu, S. Sandiwarno, and R. Prince, *Deep learning techniques for rating prediction: a survey of the state-of-the-art*, vol. 54, no. 1. Springer Netherlands, 2021.

[36] J. Serrano-Guerrero, J. A. Olivas, and F. P. Romero, "A T1OWA and aspect-based model for customizing recommendations on eCommerce," *Appl. Soft Comput. J.*, vol. 97, p. 106768, 2020, doi: 10.1016/j.asoc.2020.106768.

[37] X. Fu, T. Ouyang, Z. Yang, and S. Liu, "A product ranking method combining the features–opinion pairs mining and interval-valued Pythagorean fuzzy sets," *Appl. Soft Comput. J.*, vol. 97, p. 106803, 2020, doi: 10.1016/j.asoc.2020.106803.

[38] S. Hu, A. Kumar, F. Al-Turjman, S. Gupta, S. Seth, and Shubham, "Reviewer Credibility and Sentiment Analysis Based User Profile Modelling for Online Product Recommendation," *IEEE Access*, vol. 8, pp. 26172–26189, 2020, doi: 10.1109/ACCESS.2020.2971087.

[39] E. Georgiadou, S. Angelopoulos, and H. Drake, "Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes," *Int. J. Inf. Manage.*, vol. 51, no. November, p. 102048, 2020, doi: 10.1016/j.ijinfomgt.2019.102048.

[40] F. Falck *et al.*, "Measuring Proximity Between Newspapers and Political Parties: The Sentiment Political Compass," *Policy and Internet*, vol. 12, no. 3, pp. 367–399, 2020, doi: 10.1002/poi3.222.

[41] A. J. Nair, G. Veena, and A. Vinayak, "Comparative study of Twitter Sentiment on COVID - 19 Tweets," *Proc. - 5th Int. Conf. Comput. Methodol. Commun. ICCMC 2021*, no. Iccmc, pp. 1773–1778, 2021, doi: 10.1109/ICCMC51019.2021.9418320.

[42] R. Meena and V. T. Bai, "Study on Machine learning based Social Media and Sentiment analysis for medical data applications," *Proc. 3rd Int. Conf. I-SMAC IoT Soc. Mobile, Anal. Cloud, I-SMAC 2019*, pp. 603–607, 2019, doi: 10.1109/I-SMAC47947.2019.9032580.

[43] M. Birjali, A. Beni-Hssane, and M. Erritali, "A method proposed for estimating depressed feeling tendencies of social media users utilizing their data," *Adv. Intell. Syst. Comput.*, vol. 552, no. His, pp. 413–420, 2017, doi: 10.1007/978-3-319-52941-7_41.

[44] D. Jain, A. Kumar, and G. Garg, "Sarcasm detection in mash-up language using soft-attention based bi-directional LSTM and feature-rich CNN," *Appl. Soft Comput. J.*, vol. 91, p. 106198, 2020, doi: 10.1016/j.asoc.2020.106198.

[45] Y. Ren, D. Ji, and H. Ren, "Context-augmented convolutional neural networks for twitter sarcasm detection," *Neurocomputing*, vol. 308, pp. 1–7, 2018, doi: 10.1016/j.neucom.2018.03.047.

[46] L. Ren, B. Xu, H. Lin, X. Liu, and L. Yang, "Sarcasm Detection with Sentiment Semantics Enhanced Multi-level Memory Network," *Neurocomputing*, vol. 401, pp. 320–326, 2020, doi: 10.1016/j.neucom.2020.03.081.

[47] L. Lazib, B. Qin, Y. Zhao, W. Zhang, and T. Liu, "A syntactic path-based hybrid neural network for negation scope detection," *Front. Comput. Sci.*, vol. 14, no. 1, pp. 84–94, 2020, doi: 10.1007/s11704-018-7368-6.

[48] S. Saumya and J. P. Singh, "Detection of spam reviews: a sentiment analysis approach," *CSI Trans. ICT*, vol. 6, no. 2, pp. 137–148, 2018, doi: 10.1007/s40012-018-0193-0.

[49] L. Ronquillo, V. Zamudio, D. Gutierrez-Hernandez, C. Lino, J. Navarro, and F. Doctor, "Towards an automatic recommendation system to well-being for elderly based on augmented reality," *Proc. 2020 16th Int. Conf. Intell. Environ. IE 2020*, pp. 126–131, 2020, doi: 10.1109/IE49459.2020.9155010.

[50] J. M. Duarte, S. Sousa, E. Milios, and L. Berton, "Deep analysis of word sense disambiguation via semi-supervised learning and neural word representations," *Inf. Sci. (Ny).*, vol. 570, pp. 278–297, 2021, doi: 10.1016/j.ins.2021.04.006.

[51] Y. Wang, M. Wang, and H. Fujita, "Word Sense Disambiguation: A comprehensive knowledge exploitation framework," *Knowledge-Based Syst.*, vol. 190, no. xxxx, p. 105030, 2020, doi: 10.1016/j.knosys.2019.105030.

[52] Y. Xia, E. Cambria, A. Hussain, and H. Zhao, "Word Polarity Disambiguation Using Bayesian Model and Opinion-Level Features," *Cognit. Comput.*, vol. 7, no. 3, pp. 369–380, 2015, doi: 10.1007/s12559-014-9298-4.

[53] I. Spasic, L. Williams, and A. Buerki, "Idiom-based features in sentiment analysis: Cutting the gordian knot," *IEEE Trans. Affect. Comput.*, vol. 11, no. 2, pp. 189–199, 2020, doi: 10.1109/TAFFC.2017.2777842.

A REVIEW OF VARIOUS TRUST BASED ROUTING MODELS IN WIRELESS SENSOR NETWORKS AND IOT

Satpal Singh[1], Dr. Subhash Chander[2]
[1]Research Scholar, Punjabi University Patiala
[2]Assistant Professor, University College Jaitu

ABSTRACT— A trust-based network routing is one of the most efficient schemes to ensure the routing and the security of wireless sensor networks. A trust and reputation-based models are used widely nowadays with blockchain method to ensure better routing and security in WSN. The various existing schemes use different consensus mechanisms to authenticate any transaction. But using blockchain in WSN consume more energy as all the transactions need to wait for the blockchain network signal to commit or drop transaction. The security is one of the major factors ensured by using blockchain method, as it is a non-tempered network. The WSN security attacks are not possible in such network. In this paper various existing blockchain based schemes for routing in WSNs and IoT were discussed with performance analysis.

KEYWORDS— Blockchain; WSN; IoT, Trust; Reputation

INTRODUCTION

The wireless sensor networks had gained a lot of popularity in last decade due to its usage in wide variety of applications. In such network's nodes are allowed to move freely and are connected to a wireless base station. The data is transmitted among the nodes where each node acts as router for another. The nodes can freely join the network by fulfilling an authentication procedure. Due to this WSNs are prone to variety of attacks. Also, WSNs nodes had limited energy, as the network progresses the nodes energy starts to deplete in each round. The goal is to enhance the network lifetime. There is a tradeoff between nodes energy and security in WSNs. Energy efficient routing may lead in security lack and vice versa (Yang et al., 2019). The need is to design an optimal model ensuring efficient routing and security. The blockchain is one of the popular methods used nowadays which is integrated with WSNs. In figure 1 various blockchain applications in networking are shown:

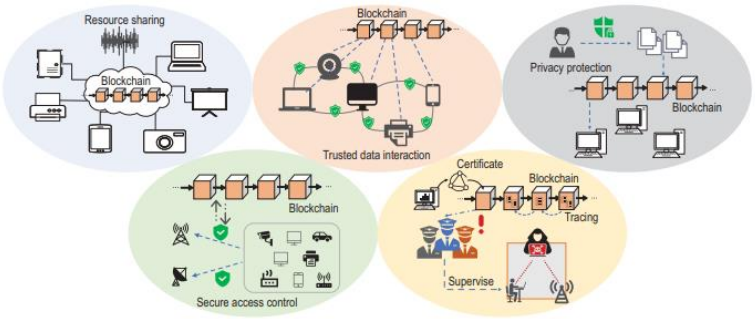


Figure 1 Applications of Blockchain Network in Networking (Wang et al., 2021)

WSNs architecture can be in form of centralized where a single server will authenticate the transaction or in decentralized architecture the 3rd party networks like blockchain are used to authenticate the transaction. The blockchain networks are subject to confirmation of minimum 6 network confirmations to perform a single transaction. The hash values are used in blockchain and hashes are non-tempered.

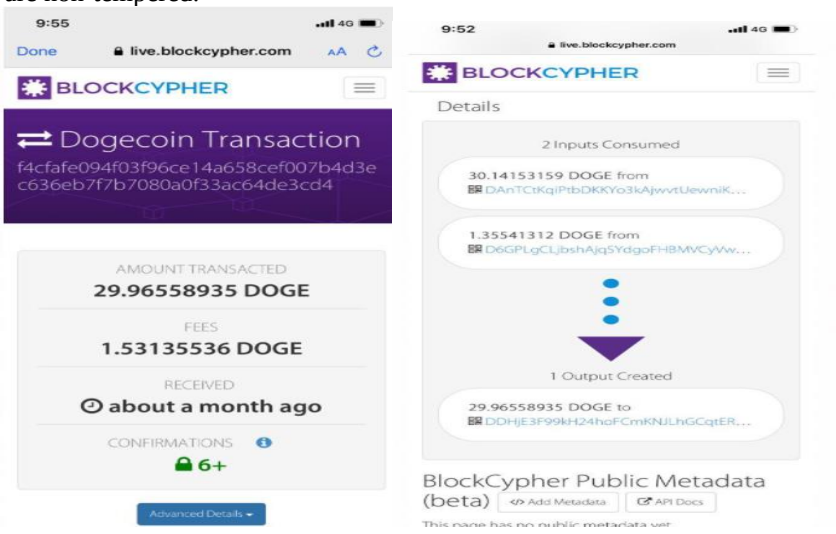


Figure 2 Blockchain Transaction Network Hash and Confirmation Representation (Orignal Image Used)

In figure 2 the block chain network transaction is represented which require atleast 6 network confirmations and the hash value are used to transact with a mining fees. Further in this paper various existing trust based methods are discussed with performance analysis.

LITERATURE REVIEW

(De Filippi et al., 2020) had identified the various problems associated with the governance of the trust in a blockchain based wireless network. The author states that a blockchain network is not based on trust but based on the confidence value. Simulating any block chain network will helps in gaining the confidence of transaction to legal node, whereas there is not any defined mechanism to recover the transactions which halt due to network traffic or low mining fees. There are multiple consensus available in blockchain network, but all can ensure the confidence not trust. The usage of blockchain network for applications like WSNs and IoT will helps in gaining confidence in transactions using de-centralized architecture.

(Moinet et al., 2017) had proposed a trust and authenticated routing scheme for wireless sensor network. In this method the WSN is connected to a decentralized block chain network to authenticate all the network transactions. The proposed model is also using a cryptographic method to encrypt the data to transfer among the nodes. The network routing decisions are taken on basis of real time dynamic computations which consume more energy but ensure the selection of best trusted path. The proposed method had shown efficient results for energy usage, delay, and security.

(Putra et al., 2021) had proposed a trust-based routing to authorize the transactions and nodes in IoT. The proposed trust and reputation (TRS) scheme use and access control method which provide the score to each node on the successful transmission of the data. The score is dynamically computed in the beginning of every round for the selection of best path. The proof-of-concept consensus mechanism is used in the blockchain network. The experimental results have shown that the proposed scheme achieves better latency and ensure better security.

(Jayasinghe et al., 2019) had proposed a trust chain model for privacy preservation with edge computing. The proposed model utilizes he power of blockchain network with trust computation concept. The train chain model is used in edge computing to ensure better delay and privacy. The trust chain network includes the control layer to decrease the ledger size. The work used a encryption scheme with efficient key exchange method for the data transfer. The trust chain model supports the data sharing with authentication from the decentralized server.

(She et al., 2019) had proposed a blockchain trust-based model for malicious node detection in WSN. In this work author proposed a trust model based on blockchain. Initially to detect the malicious node the data structure is used. The proof of work (POW) consensus mechanism is used but the constraint of high energy usage was improved by combining the POS with POW. The hash formation with every transaction ensures the data will be delivered to the authenticated node only. There is different has for each node, the sybil attack is also not possible as simulation the behavior of same hash is not possible.

(Kim et al., 2019) proposed a trust management model based on blockchain method. The main aim of the model is to enhance the trust among nodes to detect and eliminate the malicious node in wireless sensor network. The model uses a behavioral based trust scheme which evaluate the behaviors of the nodes and allocate a trust value to each node. The proposed scheme uses various factors like honesty, closeness, and intimacy to compute the trust value. The proposed model had shown better results in term of true and false positive rate than various existing schemes.

(Li et al., 2020) had proposed a scheme for trust management in wireless and mobile communications. In this trust management mechanism, a trust trunk system is used which uses the efficient storage system to detect the malicious nodes. The computed trust for all nodes in each round is normalized and risk is also measured. The risk measure needs to store the multiple transactions, which increase the demand of storage space. But still the model performs well in securing the network.

(Yang et al., 2019) had proposed a trust-based routing scheme with reinforcement learning using the blockchain in wireless sensor networks. The tampering in the network is not possible as the blockchain network is using the scheme of POW which requires the authentication for at least 50% of the network nodes to commit any transaction. The reinforcement learning is used in the wireless routing that learn dynamically for the network each round. The process of routing improves in every round and the decisions of path selection improves after every round. The work is validated on basis of various parameters like throughput, efficiency, and latency.

Discussion

The various existing methods are studied in the previous literature review section. Based on the studies the Table I discussed with comparison between WSNs and Blockchain networks. In Tale II the existing methods proposed by various authors to ensure trust based routing are discussed:

TABLE I
COMPARISON BETWEEN BLOCKCHAIN AND WSN NETWORK

Factor	WSN	Blockchain
Privacy	Privacy Lack and Nodes Compromised	All participating nodes in network ensure privacy
Architecture	Centralized	Decentralized
Bandwidth	Limited Bandwidth Requirement	High Bandwidth
Scalability	Medium Scalable	Highly Scalable
Resources	Restricted Resources	More Resources
Energy	Limited	Limited or Chargeable
Latency	Low Latency	High Latency
Security	Less Secure	Highly Secure

TABLE II
COMPARATIVE STUDY OF VARIOUS EXISTING TRUST BASED SCHEMES IN WSNS

Author	Method	Objective
(Feng et al., 2018)	Collocation storage architecture using blockchain for secure data processing	To improve the security and reliability of data in wireless sensor networks using best resource allocation scheme
(Moinet et al., 2017)	Blockchain based monitoring with trust computation to detect the malicious node in the network	To make the routing process highly secure and to enhance the overall network performance
(Buldin et al., 2019)	Implement the usage of block chain method for industrial process to improve the efficiency in highly scalable processes	To enhance the process of data exchange in industrial devices with minimum energy consumption by selecting the best optimal paths
(Casado-Vara, 2019)	Stochastic variable based model using blockchain network to predict the network attacks before the malicious node create its own network	To improve data security
(Ren et al., 2018)	Build wireless sensor network model for the data storage of the nodes using block chain network	Efficient use of storage space
(Youssef et al., 2019)	Design a system architecture for the dam surveillance using the blockchain network	Enhanced the data security to reduce the energy consumption

CONCLUSION

In this paper many existing models used in WSNs to ensure efficient routing and security are discussed. The blockchain method is one of the trending methods used nowadays to ensure complete security. Using blockchain the energy consumption is higher, but the network will become immune to all the internal and external attacks. Further many authors integrate the blockchain method with WSN to authenticate the transactions which ensure security and for better path selection variety of learning algorithms are applied to make routing energy efficient and to reduce the delay. There is further need to enhance such models with better efficient learning and data aggregation to reduce the network load, and also the selection of blockchain consensus mechanism is one of the big question, the optimal selection of consensus mechanism as per need to be selected for transactions confirmations.

REFERENCES

1. Buldin, I. D., Gorodnichev, M. G., Makhrov, S. S., & Denisova, E. N. (2019). Next Generation Industrial Blockchain-Based Wireless Sensor Networks. *2018 Wave Electronics and Its Application in Information and Telecommunication Systems, WECONF 2018*, 1–5. https://doi.org/10.1109/WECONF.2018.8604408
2. Casado-Vara, R. (2019). Stochastic approach for prediction of WSN accuracy degradation with blockchain technology. In *Advances in Intelligent Systems and Computing* (Vol. 801). Springer International Publishing. https://doi.org/10.1007/978-3-319-99608-0_58
3. De Filippi, P., Mannan, M., & Reijers, W. (2020). Blockchain as a confidence machine: The problem of trust & challenges of governance. *Technology in Society*, *62*(June), 101284. https://doi.org/10.1016/j.techsoc.2020.101284
4. Feng, L., Zhang, H., Lou, L., & Chen, Y. (2018). A Blockchain-Based Collocation Storage Architecture for Data Security Process Platform of WSN. *2018 IEEE 22nd International Conference on Computer Supported*

Cooperative Work in Design ((CSCWD)), 75–80.

5. Jayasinghe, U., Lee, G. M., MacDermott, Á., & Rhee, W. S. (2019). TrustChain: A Privacy Preserving Blockchain with Edge Computing. *Wireless Communications and Mobile Computing*, *2019*, 1–17.
6. Kim, T. H., Goyat, R., Rai, M. K., Kumar, G., Buchanan, W. J., Saha, R., & Thomas, R. (2019). A Novel Trust Evaluation Process for Secure Localization Using a Decentralized Blockchain in Wireless Sensor Networks. *IEEE Access*, *7*, 184133–184144. https://doi.org/10.1109/ACCESS.2019.2960609
7. Li, F., Wang, D., Wang, Y., Yu, X., Wu, N., Yu, J., & Zhou, H. (2020). Wireless communications and mobile computing blockchain-based trust management in distributed internet of things. *Wireless Communications and Mobile Computing*, *2020*. https://doi.org/10.1155/2020/8864533
8. Moinet, A., Darties, B., & Baril, J.-L. (2017). *Blockchain based trust & authentication for decentralized sensor networks*. 1–6. http://arxiv.org/abs/1706.01730
9. Putra, G. D., Dedeoglu, V., Kanhere, S. S., Jurdak, R., & Ignjatovic, A. (2021). Trust-based Blockchain Authorization for IoT. *IEEE Transactions on Network and Service Management*, 1–12. https://doi.org/10.1109/TNSM.2021.3077276
10. Ren, Y., Liu, Y., Ji, S., Sangaiah, A. K., & Wang, J. (2018). Incentive Mechanism of Data Storage Based on Blockchain for Wireless Sensor Networks. *Mobile Information Systems*, *2018*. https://doi.org/10.1155/2018/6874158
11. She, W., Liu, Q., Tian, Z., Chen, J. Sen, Wang, B., & Liu, W. (2019). Blockchain trust model for malicious node detection in wireless sensor networks. *IEEE Access*, *7*, 38947–38956. https://doi.org/10.1109/ACCESS.2019.2902811
12. Wang, J., Ling, X., Le, Y., Huang, Y., & You, X. (2021). Blockchain-enabled wireless communications: a new paradigm towards 6G. *National Science Review*, *April*. https://doi.org/10.1093/nsr/nwab069
13. Yang, J., He, S., Xu, Y., Chen, L., & Ren, J. (2019). A trusted routing scheme using blockchain and reinforcement learning for wireless sensor networks. *Sensors (Switzerland)*, *19*(4). https://doi.org/10.3390/s19040970
14. Youssef, S. B. H., Rekhis, S., & Boudriga, N. (2019). A Blockchain based Secure IoT Solution for the Dam Surveillance. *IEEE Wireless Communications and Networking Conference, WCNC*, *2019-April*, 1–6. https://doi.org/10.1109/WCNC.2019.8885479

FLOW BASED PROGRAMMING: APPLICATIONS FOR FOG COMPUTING

Kirandeep Kaur[#1], Arjan Singh[#2], Anju Sharma[#3]
[#1]*Computer Science and Engineering Department, Punjabi University*
[#2]*Mathematics Department, Punjabi University*
[#3]*Computer Science and Engineering Department, PTU*
[1]kirandeep_rs17@pbi.ac.in
[2]arjan@pbi.ac.in
[3]anjusharma@thapar.edu

ABSTRACT— As the world is getting smarter each day with the advent of smart technologies like IoT, Edge Computing, Fog Computing and Cloud Computing. Our ambience is changing gradually to a smart ambience ranging from individual homes/offices to schools, factories, hospitals, cities, and critical services. As fog computing pushes resources such as compute and storage to the network edge, IoT-based services are taking aids from their propinquity to provide better performances. However, developing applications for such a smart ambience is not candid as, such applications need to support dynamic nature and large scale of the system with complex logics of QoS maintenance. In this paper, we present a prototype model for the development of IoT applications from the perspective of fog computing using open-source platform Node-Red. Specifically, we show how applications for fog computing can be developed, deployed, and used to create dynamic datasets for research.

KEYWORDS— Fog Computing, Internet of Things (IoT), Node-red, Cloud Computing, Programming

INTRODUCTION

With the prevalent development in the field of Internet of things (IOT) from 2015 onwards, it is prophesied that quantity of devices linked on internet will jump from 27 billion in 2017 to 125 billion in 2030 [1]. These devices just not include smart phones or laptops but connected cars, sensors, cameras, meters, wearables etc. These devices will use various types of connections like wireless personal area network, wireless local area network, Low-power wide area network, wired, cellular, 5G, wireless neighborhood area networks etc. to connect themselves either to the internet or the neighboring device. The companies like Microsoft and Google have introduced new microcontrollers and operating systems for consumer electronics goods to make better and secure IoT devices. But till date, there has been less work on application development for these devices. As developing application for a fog-based scenario is a difficult process because of the distributed and diversified nature of the IoT and fog computing [2] [3] [4]. There are many cloud-platform based applications available, but they face the problems like higher latency and turn-around time. Moreover, sending all the data of IoT sensors directly to cloud is impractical, as it needs more bandwidth, time, processing power and storage, and is not required, as this huge data will create redundancy and high cost to the user. Thus, the data coming from these IoT sensors need to be processed and only the required data need to be sent to the cloud [5].

In this paper a prototype model for fog computing is presented using flow-based programming approach. The flow-based programming approach basically explain data flows by message passing scenario [5] [6]. There are more than 40 flow-based programming environments are available on internet [7] but Node-red gained interest by most of the authors for IoT application development [5].

RELATED WORK

In this section we survey some existing models that works on Fog computing and dataflow models. T. Szydlo et al. [5] presented the concept of application framework development for fog computing and IoT devices using Node-red. N. K. Giang et al. [8] proposed a model for IoT that utilizes computing infrastructure across Fog and Cloud. They implemented the model in Node-red. They have used MQTT broker for device-to-device communication. N. K. Giang et al. [9] presented a case study, depicting decomposition and development of fog applications to a geographically distributed infrastructure using Node-red. They have addressed the issues of large scale, dynamic nature and context -dependent application logic of IoT and fog computing. G. Tricomi et al. [10] presented the FaaS services, to provide users to visualize applications and utilize the resources of IoT platform. S.-V. Oprea and A. Bâra [11] proposed an adaptive direct load optimization and control framework. In this the authors have aimed to minimize the electricity expense by deploying an application in Node-Red. This framework is used to overcome grid congestion, power capacity scarcity and forecast errors. Here, authors have simulated dataset of 114 single family houses.

APPLICATION FOR FOG COMPUTING

The idea of fog computing is to let billions of IoT devices to get connected right at the network boundary [12], [13], [14]. It processes the data closer to where it is produced and needed. The basic purpose of fog computing system is to perform data preprocessing, filtration, maximize utilization of resources, bridging between end-user and cloud, reduce the cost of service and bandwidth [15], [16], [17]. Here, we have implemented the concept of fog computing in Node-red as shown in Figure 10

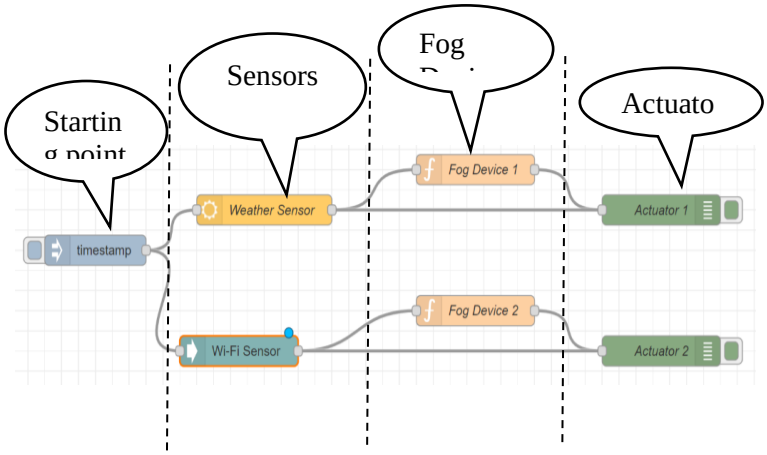


Figure 10 Concept of Fog Computing application Using Node-red

FF. Scenario 1:

In first scenario, the weather sensor is sensing the live weather of the selected city. The data of live weather has been taken from https://openweathermap.org/. The API of the OpenWeather has been passed into this node. The message of this sensor is given to Fog Device 1 to process the data so that only the required data is sent to the user and if required to the cloud. Here in this scenario, we have given information to the user about weather if it is clear or rainy. And in the last the Actuator 1 will display the message payload on debug screen.

The output of the Actuator 1 is shown in Figure 11 is the output after processing from Fog Device 1 and Figure 12 shows the output of Actuator 1 without any processing.

Figure 11 Output of Actuator-1 through Fog Device 1

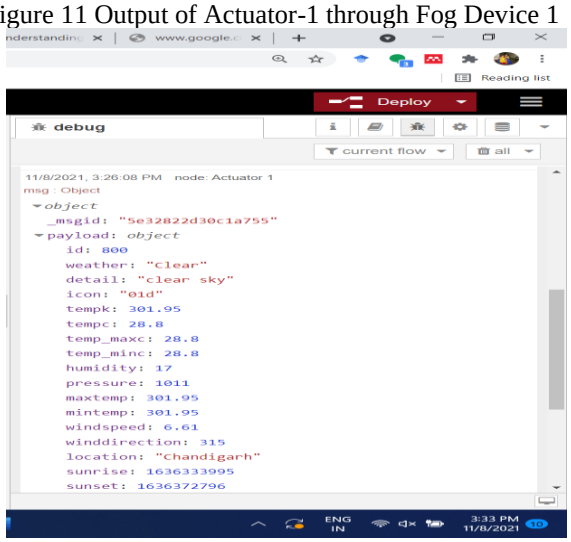


Figure 12 Output of Actuator-1 directly from sensor

GG. Scenario 2

In second scenario, the Wi-Fi sensor is sensing the connections of Wi-Fi signals available to the device. And if the device is having the strong signal of the safe Wi-Fi connection, then only Fog Device 2 will allow to get connected to the safe connection. The message of this sensor is given to Fog Device 2 to process and check for the Wi-Fi connection. Then in the last the Actuator 2 will display the message payload on debug screen. The output screen of the Fog Device 2 and the output message of the direct signal from Wi-Fi sensor is shown in Figure 13

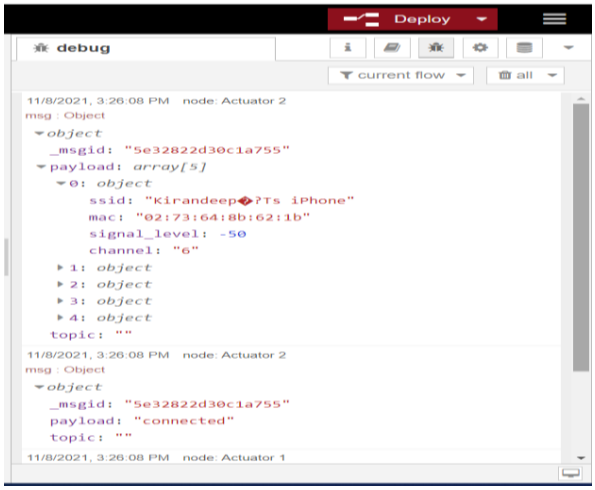


Figure 13 Output of Actuator 2

CONCLUSIONS

Fog computing has arisen as an outstanding solution to the problem of data processing for the IoT applications. Fog layer assists in processing the data closer to the IoT devices rather than shifting the data to the cloud layer. In this paper, we have presented a prototype model of application for fog computing using IoT sensors. This concept will help the researchers to build virtual applications to collect datasets for simulation purpose. Also, this framework will can be extended to make more complex applications. In our future work we are going to extend this model to make dynamic applications in fog computing for smart home and smart city.

REFERENCES

[1] I. Markit, The Internet of Things: A movement, not a market, 2017.

[2] P. Patel, A. Kattepur, D. Cassou and G. Bouloukakis, "Evaluating the Ease of Application Development for the Internet of Things.," hal-00788366, 2013.

[3] A. Awan, S. Jagannathan and A. Grama, "Macroprogramming heterogeneous sensor networks using cosmos.," ACM SIGOPS Operating Systems Review, vol. 41, no. 3, pp. 159-172, 2007.

[4] L. Mottola and G. Picco, "Programming wireless sensor networks," ACM Computing Surveys, vol. 43, no. 3, pp. 1-51, 2011.

[5] T. Szydlo, R. Brzoza-Woch, J. Sendorek, M. Windak and C. Gniady, "Flow-based programming for IoT leveraging fog computing," in IEEE 26th International conference on enabling technologies: infrastructure for collaborative enterprises (WETICE), Poznan, Poland, 2017.

[6] J. P. Morrison, Flow-Based Programming: A new approach to application, Scotts Valley, CA: CreateSpace, 2010.

[7] Mike, "48 Open Source Flow Based Programming Software Projects," Open Source Libs, [Online]. Available: https://opensourcelibs.com/libs/flow-based-programming. [Accessed 31 October 2021].

[8] N. K. Giang, M. Blackstock, R. Lea and V. C. Leung, "Developing IoT applications in the Fog: A Distributed Dataflow approach," in International Conference on the Internet of Things (IOT), Seoul, Korea, 2015.

[9] N. K. Giang, R. Lea and V. C. Leung, "Developing Applications in Large Scale, Dynamic Fog Computing: A case study," Software: Practice and Experience, vol. 50, no. 5, pp. 519-532, 2020.

[10] G. Tricomi, Z. Benomar, F. Aragona, G. Merlino, F. Longo and A. Puliafito, "A NodeRED-based dashboard to deploy pipelines on top of IoT infrastructure.," in IEEE International Conference on Smart Computing (SMARTCOMP), Bologna, Italy, 14-17 Sept. 2020.

[11] S.-V. Oprea and A. Bâra, ""Edge and fog computing using IoT for direct load optimization and control with flexibility services for citizen energy communities," Knowledge-Based Systems, vol. 228, p. 107293, 2021.

[12] F. Bonomi, R. Milito, J. Zhu and S. Addepalli, "Fog Computing and Its Role in the Internet of Things," in Proceedings of the first edition of the MCC workshop on Mobile cloud computing. ACM, 2012., Helsinki, Finland, 17 August, 2012.

[13] R. Mahmud and R. Buyya, "Fog computing: A taxonomy, survey and future directions," arXiv preprint arXiv:1611.05539 (2016)., 2016.

[14] I. Stojmenovic and S. Wen, "The Fog Computing Paradigm: Scenarios and Security Issues," in 2014 Federated Conference on Computer Science and Information Systems (FedCSIS), IEEE, Warsaw, Poland, 7-10 Sept. 2014.

[15] M. Aazam and E.-N. Huh, "Dynamic resource provisioning through Fog micro datacenter.," in 2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), Gwangiu, South Korea, 24-27 March 2015.

[16] T. H. Luan, L. Gao, Z. Li and L. Sun, "Fog Computing: Focusing on Mobile Users at the Edge," arXiv Preprint arXiv:1502.01815, p. 11, 2015.

[17] O. Skarlat, S. Schulte and M. Borkowski, "Resource Provisioning for IoT Services in the Fog," in 2016 IEEE 9th International Conference on Service-Oriented Computing and Applications (SOCA), Macau, China, 4-6 Nov. 2016.

REVIEW OF EMERGING IMAGE FUSION TECHNIQUES FOR REMOTE SENSING APPLICATIONS

Perminder Kaur[#1,2], Raman Maini[#3], Sartajvir Singh[#4]
[1,3]*Department of Computer Science and Engineering,*
Punjabi University,
Patiala, Punjab, 147 002, India
er.perminderkaur@gmail.com
research.raman@gmail.com
[2]*Department of Computer Science and Engineering,*
Baba Banda Singh Bahadur Engineering College,
Fatehgarh Sahib, Punjab, 147 002, India.
er.perminderkaur@gmail.com
[4]*Chitkara University School of Engineering and Technology,*
Chitkara University,
Himachal Pradesh, 174 103, India.
sartajvir.singh@chitkarauniversity.edu

ABSTRACT— Image Fusion (IF) is the process of integrating different source images to create a composite image that contains more comprehensive and beneficial information for future study and subsequent applications. The main objective of fusing multi-sensor information into a single image is to enhance the image quality by maintaining the coherency of the important features. It is being widely applied in various real-time application domains like intelligent robots, medical imaging, satellite imaging, manufacturing process monitoring and electronic circuit design and inspection, etc. IF has been a well-discussed research topic in the remote sensing community because of the ever-increasing versatility of the earth observation sensors, but are not capable enough to extract the assorted land features in one image so, for the effective utilization of various forms of image data, its essential to enhance the spatial as well as the spectral quality of the image. Nowadays, many different and emerging IF methods are available, this article intends to review the state-of-the-art IF techniques and initiates with the introduction of IF. Categorization of IF techniques based on the application of fusion method at different levels of fusion have been discussed and advanced techniques in the field of IF have been explored. Finally, applications of different fusion techniques with respect to satellite imaging are presented.

KEYWORDS: *Image Fusion, Remote Sensing, Spatial and Spectral quality, Multi-sensor, Fusion level.*

I. INTRODUCTION

More data is indeed accessible for advanced research in the field of remote sensing due to the increasing number of sensors in use. This is the reason why the need to analyze and blend the various data sources to extract the most meaningful information is increasing [1]. When the data is restricted to only image data it is termed image fusion which is one component of data fusion. In remote sensing, image data is captured from sensors having different physical characteristics lying at a distance, and remote sensing image fusion (RS-IF) thus enhances the understanding of our surroundings [2]. RS-IF has been widely used for the analysis and extraction of fine features of the large land-covered area for change detection, classification, building extraction, traffic monitoring, network tracking and to improve the resolution of the low-resolution image of one sensor with the help of the high-resolution image of the same or different sensor, which is termed as Pan-sharpening [3].

The fast growth in the amount and availability of data from a variety of sources poses significant problems related to the efficient storage and processing of data. There are numerous remote sensing and supplementary data sets available for a particular remote sensing application; this raises a difficulty in finding the best combination of the data sets to get the ideal results [3], [4]. It is, for this reason, that fusion of data acquired from multiple sources has attracted so much attention in recent years in the context of remote sensing. There exists a wide range of approaches in RS-IF, capable of preserving the spatial and spectral information of the input images, providing the most meaningful information without generating any variations in the fused image. For the fused image to be more suitable for further analysis, it is very essential to remove the geometrical differences between the source images by mapping various images with respect to the reference image [5]. This form of mapping is performed as the first step of the IF process which is known as image registration, followed by the steps required to extract the information of interest using the appropriate technique.

Many state-of-the-art IF techniques have been developed nowadays and have been generally categorized into three main classes: (i) Pixel Level, (ii) Feature Level, and (iii) Decision Level [6], [7]. IF techniques at the pixel level incorporate information from incoming images directly for additional computer processing activities using arithmetic operations and transformations, whereas feature level IF combine the similar features recognized in various source images. On the other hand, in decision level IF methods for information extraction, each of the input images is analyzed separately.

In [8] some IF approaches in the remote sensing field have been reviewed and a sub-pixel level fusion have been proposed to overcome the difficulties of multi-sensor IF. Pohl C in [9] explained the concept and importance of pre-processing and knowledge of the sensor characteristics in multi-sensor IF applications. Solanky et al. in [10] provided a comparative review of the most popular and recent IF techniques. In [11] various multi-resolution pixel-level fusion methods have been explored

and evaluated in comparison to the traditional methods that resulted in better accuracy and improved image quality. Several papers have been published in recent years that focus on enhancing the fusion quality and advancing the application areas of image fusion. In [12] Joshi et al. reviewed the application of optical and radar data fusion for the discovery and mapping of land-use areas. Shah E et at., applied some pixel-level fusion techniques on the Synthetic Aperture Radar (SAR) and optical images to identify various ice features in the Antarctic region [13]. [14] demonstrated the advantages of IF on multi-sensor imagery using Ehlers fusion technique coupled with Rotation Forest machine learning algorithm for land use and land cover classification. Kulkarni et al., in [15] surveyed many traditional and hybrid approaches of IF on SAR and optical imagery for several remote sensing applications and provide an insight into the major advancements in the field of IF.

Although many papers have been published on the history and development of image fusion in various applications, the recent emergence of multi-sensor satellite image fusion in remote sensing has not been covered. The goals of this study are to provide an overview of recent trends in multi-sensor satellite image fusion, with an emphasis on principal remote sensing applications.

This paper is organized as follows: Section II presents a brief description of the IF categories based on the level of fusion, Section III gives an overview of various traditional RS-IF techniques, Section IV presents recent trends in RS-IF and finally, their applications are discussed in Section V.

II. CATEGORIZATION OF IMAGE FUSION TECHNIQUES

IF refers to the process of extracting data from the satellite images that have multi-spectral and or multi-temporal characteristics by the integration of images collected from satellite-based instruments to produce a more meaningful and complete image of the observed environment. With respect to the application of the fusion process, a variety of methodologies and procedures have been developed throughout the years. These approaches are commonly categorized into three types, according to [16] as shown in figure 1.

A. Fusion at Pixel Level

This type of fusion takes place at the lowest level of processing i.e. on a pixel-by-pixel basis and thus this type of fusion handles the highest amount of data. In this method, a new data value is calculated for each pixel of the source image by using several mathematical operators to combine the information at the same pixel in other images. Prior to fusion, all the images have to be carefully registered. The resulting fused image has more detailed and accurate information about the object of interest like vegetation cover [17].

B. Fusion at Feature Level

This kind of fusion involves extracting the features from multiple images and then merging them into a single image [18]. Before applying fusion, the distinguishing features from the different source images of the scene are identified, segmented, and extracted from several domains and then combined to get the resultant fused image. This image contains features that help in identifying objects of interest within a heterogeneous environment. This method is usually used for mapping the growth and changes in farmland and forests. Accurate selection and extraction of similar features from multiple source images is the prerequisite of this method.

C. Fusion at Decision Level

Fusion is applied at the advanced level of image processing that involves feature and decision making. In this process extracted features are combined using the external decision rules to reinforce common interpretation [19]. Decision rules are generated for the features collected from each individual image by mixing distinct feature extractions and interpretations using one or more mathematical operators. As the final decision-making step is driven by the type of features used, this type of fusion highly depends upon the quality of features extracted and can be applied for land cover/land use classification [20].

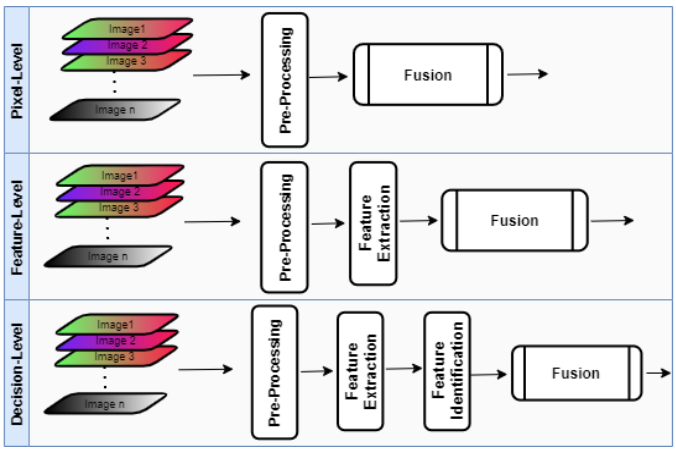


Fig. 1: Description of IF process at different levels.

III. REMOTE SENSING IMAGE FUSION TECHNIQUES

Several fusion strategies have been presented over the last two decades. The majority of these methods are centered on striking a balance between intended spatial improvement and spectral consistency. Various methods such as intensity-hue-saturation (IHS), high pass filtering (HPF), Brovey transform (BT), and artificial neural networks (ANNs) are commonly used in the domain of IF [21]. A general overview of the selected methodologies is presented in this paper with a focus on current developments in RS-IF.

*IHS: T*his technique involves converting the histogram of the high-resolution source image according to the statistical parameters of the intensity component of the low-resolution multi-spectral image which is converted from RGB to IHS color space. The intensity component is then replaced by the histogram-modified high-resolution image. Finally, inverse IHS transformation is applied to get the resultant fused image [22].

HPF: In this method, a high-resolution image is scanned using a high pass filter in the frequency domain [23]. The resulting image is then applied to low-resolution multispectral bands to get the enhanced spatial quality fused image.

PCA: It is a technique extensively used in the image processing field to decorrelate multicomponent datasets, in order to efficiently compact the energy of the input vectors in a reduced number of components of the output vectors, which are called the Principal Components. Based on this basic idea, a multi-spectral image is transformed to uncorrelated feature space to obtain those features that best represent the spatial information of the image, called the initial Principal Component (PC1). This component is then replaced by the histogram-matched high-resolution image. In the end, by using inverse PCA transformation, a sharpened image is obtained [22].

BT: In this technique, a high-resolution image is divided by the average of multispectral bands to get the normalized bands that need to be fused, and the resultant fused image having the desired spatial resolution is generated after multiplying it with the high-resolution band [24].

Wavelet and Pyramid techniques of IF rely upon the low pass version of the input source images which are generated by decomposing them at different scales using an iterative decomposition scheme [25]. After the decomposed images are subjected to some fusion rule, their sub-bands are then transformed using the inverted transformation techniques to get the fused image.

Wavelet Transformation: This technique decomposes the high-resolution source image into a series of low-resolution images with the corresponding wavelet coefficients at each level [26]. The individual band of multi-spectral low-resolution source image then replaces the low-resolution images at the same resolution level and high spatial details are injected by applying the inverse wavelet transformation on each band of the multi-spectral low-resolution image to get the final fused product.

Pyramid Transformation: These techniques basically decompose the source images using Gaussian or Laplacian Pyramid, in which an image pyramid is formed as a set of lowpass versions of the original image and each version represents a pattern of information of a different scale. At each level of the pyramid most meaningful features are extracted after performing feature enhancement and injected into the low-resolution source image using some fusion rule and finally enhanced image is reconstructed using inverse transformation [27].

ANN-based technique: A pair of images is divided into multiple blocks. The resulting feature vectors can be extracted from the relevant blocks. The first step in training neural networks is to identify the various components to assess the fusion effect. Usually, these are spatial frequency, edge, and visibility. The ANN model can recall a functional relationship. This capability can be used for further calculations once it has been trained [28]. For these reasons, the ANN approach has been used to create nonlinear models for multi-sensor data fusion.

Hybrid Techniques: As the name suggests different fusion methods can be combined to get a better quality of result which may not be achieved through individual techniques. So, developers nowadays find it a useful strategy to combine a method resulting in an improved spatial quality image with the one generating better spectral quality to get the benefits of both[29]. Many researchers have focused on the fusion of wavelet transform with the IHS method to achieve better spatial resolution and spectral quality. However, it is also necessary to find the optimal fusion strategy to achieve the best combination.

IV. RECENT TRENDS

Latest developments in the field of RS-IF are progressing in several directions because of the improvements in data processing tasks which have to broaden the application areas of image fusion across many disciplines. In this section, recent trends in IF have been discussed.

Data Mining: Massive amounts of data from various vehicle-borne, UAV-borne, airborne and space-borne sensors are available with an increased spatial and spectral resolution, waiting for proper extraction of important information. Data mining is playing an important role in this direction. Automatic object identification through data mining can become helpful in identifying objects of interest in the fused data sets. In [30] the concept of data mining has been utilized to classify land use and land cover areas in the fused data sets.

Big Data: Every day, remote sensing satellites collect vast amounts of data that met the three criteria of big data: variety, volume, and velocity. It can be used in various data-driven applications. As a result, we are living in the Age of Big Remote

Sensing Data [31], [32]. The problems and prospects of large data processing for remote sensing applications are investigated in [31] and [32]. Methodology of big data processing like High-Performance Computing is being utilized for extracting the information from large remote sensing fused datasets in RS-IF as the conventional fusion techniques are no more supported by big remote sensing data.

Deep Learning: Deep learning can be used to fuse multi-sensor data for remote sensing applications due to its ability to extract and represent data. A method based on convolution neural network (CNN) was proposed in [33] to predict the missing information from the collected data. Schmitt et al. [3] introduce a deep learning algorithm that learns to colorize SAR images. The algorithm learns by training the data with co-registered optical data. The algorithm achieves its goal by improving the interpretability of the data, which is difficult to do with traditional SAR and optical IF techniques. Some researchers have proposed establishing deep learning-based fusion techniques for processing large datasets at the feature and decision level [33], [34]. Liu et al. [35] provide a comprehensive overview of deep learning-based pixel-level fusion techniques.

Higher-level fusion methods: The use of high-level fusion techniques for improving image accuracy is becoming the norm in fusion technologies. For image classification and data extraction, the use of multi-features is required. To achieve a high level of fusion, multi-dimensional features such as spectral content and structural context from multi-source images are fused at feature and decision levels. Probability theory, evidence theory, fuzzy and possibility theory, neural networks, and other mathematical tools are currently available for high-level fusion approaches [36]. Modern machine-learning techniques are becoming more prevalent. Support Vector Machine (SVM) and ensemble learning, as well as other recent advancements in machine learning, can also be applied to achieve high-level fusion [37].

V. REMOTE SENSING IMAGE FUSION APPLICATIONS

By mapping and identifying the various features of the Earth's surface, remote sensing techniques can help monitor and maintain the planet's atmosphere and its various terrestrial features. Due to the advancements in sensor technology, the amount of data that can be collected from remote sensing devices continues to expand at an astounding rate. Multi-sensor image fusion is frequently utilized in remote sensing for many applications to improve image interpretation. Some of the major image fusion applications are briefly discussed in this section.

Classification: One of the most important functions of remote sensing applications is classification. When multi-source image data is used in the processing of remote sensing imagery, the classification accuracy improves. Images from microwave and optical sensors provide additional information that aids in-class differentiation [38]. Optical and microwave sensors data help in the classification of different kinds of land covers. The data can be used to distinguish between classes of land features based on their spectral characteristics.

Object Identification: Feature augmentation is a capability that can be utilized in image fusion to provide better images than the original data. Fused images are valuable products for maximizing the amount of information recovered from satellite image data. A method for detecting urban building structures using the Dempster-Shafer fusion theory was presented in 2004 [39]. A classification method based on the Dempster-Shafer theory [40] of data fusion distinguishes the classes tree, grassland, and bare soil. Image fusion techniques could also help with the identification of linear objects like roads.

Urban Monitoring: Urbanization and industrialization is the process that increases the number of people living in cities. This contributes to various environmental problems such as the greenhouse effect and groundwater depletion [41]. Thus, monitoring of the urbanization process using the remotely sensed data plays an important role in urban pollution, energy consumption, and risk reduction in natural hazards and climate change analysis. In urban contexts, detailed and up-to-date information sources are required for planning, maintenance, and resource management. This is where RS-IF comes in to help get the most information out of the images. As spatial resolution got improved, it became possible to map complicated urban areas in great detail.

Agriculture: Food security has become a major issue in agriculture as the world's population grows and cultivatable land becomes scarce. RS-IF is a useful tool for mapping cropland and crop monitoring. Crop mapping for resource management in terms of crop types, vegetation health, and changes in farmland extent is a major RS-IF application in agriculture [42]. It is important to acquire data at specific moments in time to derive relevant parameters, which poses the challenge of cloud cover and limits image acquisition. To address these restrictions, multi-temporal and multi-sensor image fusion is a useful tool.

Change Detection: Change Detection is a technique that combines multiple sensors' images to detect changes in an area over time. By combining these images, it is possible to detect deviations in the observed phenomenon. Multitemporal and multi-sensor images taken for a certain area of the earth provide complementary information, and fusing this data is always helpful in detecting changes in that area [43].

Forest Monitoring: Natural resources and habitats exist in forests. Human exploitation and natural factors such as fire have resulted in extensive deforestation. Forest monitoring and management are critical for maintaining ecological balance. Because of the additional information provided by high-resolution and multi-spectral remote sensing imagery, one of the most successful remote sensing technologies in forest monitoring is IF for the study of deforestation, biomass estimation and watershed conservation, etc [44].

Natural Hazards and Disasters: A hazard is a natural activity or phenomenon that has the potential to have a negative influence on society. Information gathered using the available electromagnetic spectrum at high spectral and spatial resolution with a high repetition rate at global, regional, and local scales is critical for protecting life and predicting future natural calamities [24], [45].

Security and Defense: Multi-sensor remote sensing plays a key role in national security and military applications. The information is part of geospatial and visual intelligence that is used in crisis management missions and operations, as well as maritime surveillance and border control. Automatic target detection and mapping of locations and developments of interest are made possible by the RS-IF approaches [46].

Archaeology: By combining data from reflective infrared for changes in vegetation, thermal infrared for temperature changes, and active microwaves like ground-penetrating radar (GPR) for temperature changes, one can detect and identify many more features of such archaeological sites than with any single sensor. The typical image characteristics of tone, texture, structure, pattern, shape, size, shadow, orientation, and associated features from the fused images are used by archaeologists using remote sensing techniques to discover new sites or to monitor existing known sites [47].

VI. CONCLUSION

Image fusion is a process that combines multiple sensor data streams to obtain deeper insights from complex images. It is commonly used for developing image-based applications. Among the hundreds of image fusion approaches, IHS, PCA, BT, and wavelet transform are the most extensively utilized techniques. The major problem with algorithms like IHS, PCA, and BT which have reduced complexity and shorter processing time, is color distortion [48]. Wavelet-based schemes are more advantageous in terms of reducing color distortion than other traditional methods. However, they require more complexity in terms of computation and parameter setting. Due to the complexity of the data processing involved in the processing of hyperspectral satellite sensor data, it is one of the main obstacles that the present fusion techniques need to overcome. ANN appears to be one feasible solution to deal with the hyper-spectral satellite sensor data's high dimension nature. Hybrid methods integrate the best features of these methodologies. Fusion approaches that combine decomposition-based methods and component replacement methods provide a superior fused product than standalone methods, according to the findings in [49].

RS-IF research is moving in the direction of deep learning, data mining, big data, and cloud computing due to the various changes in the sensor geometry, polarization, and resolution of an image, image fusion methods are very specific to the data set and require fine-tuning of the algorithm's parameters. Aside from that, there are several challenges in multi-sensor image fusion, such as multi-sensor picture registration, source image noise, and computational complexity that also need to be addressed. With the launch of new microwave and optical remote sensing satellites with higher resolutions, fusion of multi-sensor images remains a hot topic that will be valuable for a variety of remote sensing applications.

REFERENCES

[1] K. C. Bhataria and B. K. Shah, "A Review of Image Fusion Techniques," *Proceedings of the 2nd International Conference on Computing Methodologies and Communication, ICCMC 2018*, pp. 114–123, Oct. 2018.

[2] D. Mishra and B. Palkar, "Image Fusion Techniques: A Review," *International Journal of Computer Applications*, vol. 130, no. 9, pp. 7–13, Nov. 2015.

[3] M. Schmitt and X. X. Zhu, "Data Fusion and Remote Sensing: An ever-growing relationship," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 6–23, 2016.

[4] P. Ghamisi *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1, pp. 6–39, Mar. 2019.

[5] I. Amro, J. Mateos, M. Vega, R. Molina, and A. K. Katsaggelos, "A survey of classical methods and new trends in pansharpening of multispectral images," *Eurasip Journal on Advances in Signal Processing*, vol. 2011, no. 1. Springer International Publishing, 2011.

[6] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Information Fusion*, vol. 33, pp. 100–112, Jan. 2017.

[7] V. K. Mishra, S. Singh, S. Kumar, and A. P. J. Abdul, "Evaluation of Image Fusion Algorithms Mapping of Urban Area for Smart City Planning from Remotely Sensed Images View project LULC Applications View project Evaluation of Image Fusion Algorithms," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2016, [Online]. Available: https://www.researchgate.net/publication/297002320

[8] F. Imed, "A Multi Views Approach for Remote Sensing Fusion Based on Spectral, Spatial and Temporal Information," *Image Fusion*, Jan. 2011.

[9] C. Pohl, "Challenges of Remote Sensing Image Fusion to Optimize Earth Observation Data Exploitation," *European Scientific Journal*, Dec. 2013.

[10] V. Solanky and S. K. Katiyar, "Pixel-level image fusion techniques in remote sensing: a review," *Spatial Information Research*, vol. 24, no. 4. 2016.

[11] R. Sivagami, V. Vaithiyanathan, V. Sangeetha, M. Ifjaz Ahmed, K. Joseph Abraham Sundar, and K. Divya Lakshmi, "Review of image fusion techniques and evaluation metrics for remote sensing applications," *Indian Journal of Science and Technology*, vol. 8, no. 35, 2015.

[12] N. Joshi *et al.*, "A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring," *Remote Sensing*, vol. 8, no. 1. MDPI AG, 2016.

[13] E. Shah, P. Jayaprasad, and M. E. James, "Image Fusion of SAR and Optical Images for Identifying Antarctic Ice Features," *Journal of the Indian Society of Remote Sensing*, vol. 47, no. 12, pp. 2113–2127, Dec. 2019.

[14] R. Padmanaban, A. K. Bhowmik, and P. Cabral, "Satellite image fusion to detect changing surface permeability and emerging urban heat islands in a fast-growing city," *PLoS ONE*, vol. 14, no. 1, Jan. 2019.

[15] S. C. Kulkarni and P. P. Rege, "Pixel level fusion techniques for SAR and optical images: A review," *Information Fusion*, vol. 59, pp. 13–29, Jul. 2020.

[16] C. Pohl and J. van Genderen, "Remote Sensing Image Fusion: A Practical Guide," *Remote Sensing Image Fusion*, Oct. 2016.

[17] V. R.Pandit and R. J. Bhiwani, "Image Fusion in Remote Sensing Applications: A Review," Int. J. Comput. Appl., vol. 120, no. 10, pp. 22–32, Jun. 2015.

[18] J. Zhang, "Multi-source remote sensing data fusion: Status and trends," *International Journal of Image and Data Fusion*, vol. 1, no. 1. Taylor and Francis Ltd., pp. 5–24, 2010.

[19] A. Vijan, P. Dubey, and S. Jain, "Comparative Analysis of Various Image Fusion Techniques for Brain Magnetic Resonance Images," *Procedia Computer Science*, vol. 167, pp. 413–422, Jan. 2020.

[20] H. Kaur, D. Koundal, and · Virender Kadyan, "Image Fusion Techniques: A Survey," *Archives of Computational Methods in Engineering*, vol. 1, p. 3, 2021.

[21] H. Ghassemian, "A review of remote sensing image fusion methods," *Information Fusion*, vol. 32. Elsevier, pp. 75–89, Nov. 01, 2016.

[22] V. Vijayaraj, N. H. Younan, and C. G. O'Hara, "Concepts of image fusion in remote sensing applications," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2006, pp. 3781–3784.

[23] R. Virk, "REVIEW OF IMAGE FUSION TECHNIQUES," *International Research Journal of Engineering and Technology*, 2015, [Online]. Available: www.irjet.net

[24] M. Sharma, "A Review: Image Fusion Techniques and Applications." [Online]. Available: www.ijcsit.com

[25] P. Ghamisi *et al.*, "Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 1. Institute of Electrical and Electronics Engineers Inc., pp. 6–39, Mar. 01, 2019.

[26] P. Gaurav, J. Rkdf, A. Shabahat, and H. H. O. D. Rkdf, "Application of Image Fusion Using Wavelet Transform in Target Tracking System." [Online]. Available: www.ijert.org

[27] A. Vijan, P. Dubey, and S. Jain, "Comparative Analysis of Various Image Fusion Techniques for Brain Magnetic Resonance Images," in *Procedia Computer Science*, 2020, vol. 167, pp. 413–422.

[28] K. Rokni, A. Ahmad, K. Solaimani, and S. Hazini, "A new approach for surface water change detection: Integration of pixel level image fusion and image classification techniques," *International Journal of Applied Earth Observation and Geoinformation*, vol. 34, no. 1, pp. 226–234, Feb. 2015.

[29] R. Gharbia, A. H. el Baz, A. E. Hassanien, and M. F. Tolba, "Remote Sensing Image Fusion Approach Based on Brovey and Wavelets Transforms," in *Advances in Intelligent Systems and Computing*, 2014, vol. 303, pp. 311–321.

[30] M. Pugh, A. Waxman, and D. Fay, "Assessment of multi-sensor neural image fusion and fused data mining for land cover classification," *2006 9th International Conference on Information Fusion, FUSION*, 2006.

[31] P. Liu, "A survey of remote-sensing big data," *Frontiers in Environmental Science*, vol. 0, no. JUN, p. 45, Jun. 2015.

[32] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, and Y. Zhu, "Big Data for Remote Sensing: Challenges and Opportunities," *Proceedings of the IEEE*, vol. 104, no. 11, pp. 2207–2219, Nov. 2016.

[33] Z. Shao and J. Cai, "Remote Sensing Image Fusion with Deep Convolutional Neural Network," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1656–1669, May 2018.

[34] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[35] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Information Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[36] S. Jindal and G. Josan, "Neural Network and Fuzzy Logic Approach for Satellite Image Classification: A Review," 2007. [Online]. Available: https://www.researchgate.net/publication/228959509

[37] G. Tsagkatakis, A. Aidini, K. Fotiadou, M. Giannopoulos, A. Pentari, and P. Tsakalides, "Survey of deep-learning approaches for remote sensing observation enhancement," *Sensors (Switzerland)*, vol. 19, no. 18. MDPI AG, Sep. 02, 2019.

[38] L. Yin, P. Yang, K. Mao, and Q. Liu, "Remote Sensing Image Scene Classification Based on Fusion Method," *Journal of Sensors*, vol. 2021, 2021.

[39] F. Rottensteiner, J. Trinder, S. Clode, K. Kubik, and B. Lovell, "Building detection by dempster-shafer fusion of LIDAR data and multispectral aerial imagery," *Proceedings - International Conference on Pattern Recognition*, vol. 2, pp. 339–342, 2004.

[40] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban Change Detection Based on Dempster–Shafer Theory for Multitemporal Very High-Resolution Imagery," *Remote Sensing 2018, Vol. 10, Page 980*, vol. 10, no. 7, p. 980, Jun. 2018.

[41] S. Dahiya, P. K. Garg, and M. K. Jat, "A comparative study of various pixel-based image fusion techniques as applied to an urban environment," *International Journal of Image and Data Fusion*, vol. 4, no. 3, pp. 197–213, 2013.

[42] D. Li, Z. Song, C. Quan, X. Xu, and C. Liu, "Recent advances in image fusion technology in agriculture," *Computers and Electronics in Agriculture*, vol. 191, p. 106491, Dec. 2021.

[43] B. Wang, J. Choi, S. Choi, S. Lee, P. Wu, and Y. Gao, "Image fusion-based land cover change detection using multi-temporal high-resolution satellite images," *Remote Sensing*, vol. 9, no. 8, Aug.

[44] X. Tong *et al.*, "An approach for flood monitoring by the combined use of Landsat 8 optical imagery and COSMO-SkyMed radar imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 136, pp. 144–153, Feb. 2018.

[45] Y. Byun, Y. Han, and T. Chae, "Image fusion-based change detection for flood extent extraction using bi-temporal very high-resolution satellite images," *Remote Sensing*, vol. 7, no. 8, pp. 10347–10363, 2015.

[46] S. B. G. Tilak Babu, I. Chintesh, V. Satyanarayana, and D. Nandan, "Image Fusion: Challenges, Performance Metrics and Future Directions," in *Lecture Notes in Electrical Engineering*, 2020, vol. 686, pp. 575–584.

[47] D. D., A. Agapiou, D. G., and A. Sarris, "Remote Sensing Applications in Archaeological Research," *Remote Sensing - Applications*, Jun. 2012.

[48] S. Bhat and D. Koundal, "Multi-focus image fusion techniques: a survey," Artif. Intell. Rev. 2021, pp. 1–53, Feb. 2021.

[49] S. Harinkhede and M. S. Mishra, "A Comparatively Analysis of Various Hybrid Image Fusion Techniques," *International Journal of Engineering Sciences & Research Technology*, 7(2), pp.51-55, Feb. 2018.

HYBRID APPROACH FOR SANSKRIT TO ENGLISH TRANSLITERATION

Anupama Sharma [1], Dr Dhavleesh Rattan [2], Dr Madan Lal[3]
[1,2,3]*Department of Computer Science and Engineering, Punjabi University*
[1]anupamasharmakaushal@gmail.com
[2]dhavleesh@gmail.com
[3]madanlal@pbi.ac.in

ABSTRACT— In this paper, we have proposed a hybrid approach for transliterating Sanskrit words to English. Transliteration is the conversion of the text from one script into another script. The proposed approach is a combination of grapheme based approach and phoneme based approach. We have taken a dataset of 206 words. We have evaluated our technique using various parameters like precision, recall and F-Measure. This system aims at maximum approximation of source language pronunciation. We have achieved high precision and recall.

KEYWORDS—NLP, Transliteration, Sanskrit, English, Hybrid approach

INTRODUCTION

A. Transliteration

Transliteration is representation of character from one script to another. Transliteration differs from translation as the translation is the process of conversion of one natural language text to another natural language text in a way that its meaning doesn't change. In transliteration process script gets changed and the pronunciation of the text remains same or close to same in case the same is not possible [16]. Phonetic translation across the language pairs which have different alphabets and sound systems is called transliteration. Forward transliteration is transliteration of a language from its origin to foreign language. On the other hand backward transliteration is transliteration of the text from foreign language to back its origin. Backward transliteration is also called as back-transliteration. It is transliterating a word from its transliterated version back to the source language [27].There are two directions of transliteration. Given a pair (x,y) in which x is a word of source language and y is a word of target language that is transliterated. Forward transliteration is phonetically converting x into y and Backward transliteration is generating x from y [3]. Machine transliteration is a process of converting the given text from one alphabetical system into other alphabetical system [18].

B. Need of transliteration

1) The main advantage of transliteration is that if a person doesn't know any particular language, even then he/she can read the text in that particular language so it overcomes the language barriers.
2) It is difficult to translate names and technical terms so, transliteration is used in this case.
3) Machine transliteration is important for natural language application. For example in information retrieval (Cross lingual information retrieval system) and machine translation systems where proper nouns and jargons are used [18].

The Transliteration system should retain the phonetic characteristics of source text after transliterating it in the target language text.

C. Motivation

1) Study of Sanskrit text deserves more attention than it has received till now in our country. This language has ancient heritage, religious and intellectual life [30].
2) Sanskrit language is considered to perfectly fit for the computer [24].
3) If a person wants to read a text that is written in Sanskrit but that person is not familiar with Devanagari script, then he/she can read the transliterated text (written in English).

BACKGROUND

Difference between Sanskrit and English Transliteration is the subfield of computational linguistics, and its requirement of language processing makes it language-specific [27]. The difference between the Sanskrit language and English language is shown in the below table.

TABLE XIII
DIFFERENCE BETWEEN SANSKRIT AND ENGLISH.

Basic	SANSKRIT	ENGLISH
Alphabet	42 characters	26 characters
Script	Devanagari	Roman
Number	Three: Singular plural and dual	Two: Singular and plural
Vowels	Nine vowels	Five vowels
Consonants	33 consonants	21 consonants
Sentence Order	Free order	Subject-verb-object

Approaches of Machine Transliteration

Following are the approaches that are used for machine transliteration.

No more than 3 levels of headings should be used. All headings must be in 10pt font. Every word in a heading must be capitalized except for short minor words as listed in Section III-B.

1) Grapheme-based approach

It includes mapping of source graphemes to target grapheme without any phonetic understanding of source language words. Hence, this approach is also known as direct method [32].

1)Phoneme-based approach

It includes the mapping of source graphemes to source phonemes and then from source phonemes to target graphemes [6]. In this approach pronunciation is main factor. This approach is also known as pivot method as it uses source language phonemes as a pivot while converting the source language graphemes to target language phonemes. This approach follows the following two steps:

- Transformation of source grapheme (Sg) to source phoneme (Sp).
- Transformation of source phoneme (Sp) to source grapheme (Sg) [32].

4) Hybrid based approach

This approach uses both the source graphemes and source phonemes to create transliteration model. Out of the two approaches discussed above, Grapheme based approach is considered to be better than the Phoneme based approach because the later approach involves more transformations. In some cases one more transformation is applied that is between source phoneme to target phoneme and target phoneme to target grapheme. Hence error at earlier steps makes it hard to generate correct transliteration. On the other hand if source language words have unlike pronunciation from their spellings then grapheme based approach cannot generate correct transliteration. So the combination of both the approaches is used which is known as hybrid approach [32].

RELATED WORK

Rathod et al [1] showed transliteration for Hindi to English and Marathi to English languages by using SVM (Support Vector Machine).The system architecture of the transliteration system is divided into three modules. That is Pre-processing, Training of bilingual corpus, testing of additional data. Pre-processing phase is used to convert the input into format that is acceptable for the system. In this phase Syllabification is done. Syllabification refers to the segmentation of text. They utilized SVM as a machine learning algorithm to classify patterns. SVM creates the hyper plane using linear polynomial function which divides the data into two partitions. Classification in the training phase is done based on the n-grams and they concluded that, 5-gram is best suitable size for Hindi and Marathi language to English language transliteration of named entity.

Dhore et al [2] discussed Hindi language to English language machine transliteration of proper nouns, place names and organization names using CRF (Conditional random field). They defined Conditional probability distributions P(Y |X) of target text provided the source language text. Y is word of target language and X is a word of target language. They also discussed the issues due to which the direct transliteration of Hindi language to English language is very difficult. They used concept of syllabification, in this process named entity written in Devanagari script is divided into basic units that are called aksharas or syllabic unit. They took one akshara in Devanagari as one syllabic unit in English. They received accuracy of 85.79%.

Lehal [3] discussed transliteration that is classified into two directions. Provided a pair (s,t) in which s is source language word and t is target language word that is transliterated. Forward transliteration is phonetically converting s into t and backward transliteration is generating s from t. The architecture of machine translation is split into three stages i.e. pre-processing, processing, post-processing. In pre-processing stage, the Gurmukhi word is normalized and prepared for transliteration according to Shahmukhi word (spellings and pronunciation). In processing stage transliteration from Gurmukhi to Shahmukhi is done by using rule based approach. In post- processing stage, Shahmukhi words are corrected by using Shahmukhi corpus. The transliteration accuracy for his system was 98.6%.

Chinnakotla and Damini [4] discussed Character Sequence Modeling (CSM) in transliteration research. They discussed the transliteration into English from resource scant languages. They have done transliteration from Hindi and Persian into English. Their system uses a mapping table that maps characters between the Hindi and English letters and uses CSM for ranking of the generated candidates. They improved performance of CSM by following changes: They enlarged the alphabet because CSM allocates lower probability to the candidates that are longer and higher probability to the shorter candidate.

Ganesh et al [5] described a statistical transliteration technique that utilizes HMM that is Hidden Markov Model alignment and CRF that is Conditional Random Fields. Using this technique, desired numbers of transliterations are produced for a specified word. They built and presented the Hindi-English transliteration system. Transliteration's approach is divided into two phases. In first phase, character alignment is done and the second phase utilizes statistics for transliterating the

specified words. They also demonstrated that transliteration done using both HMM alignment and CRF gives better results than using HMMs alone for the sample of cross language information retrieval (CLIR).

Chinnakotla et al [6] discussed forward and backward transliteration. Their system uses Character Sequence Modeling (CSM) on source side to identify word origin, and then again uses the CSM on the target side to rank the outputs. They also discussed various approaches of machine transliteration such as Grapheme based approach, phoneme-based approach and hybrid approach other approaches like Rule based machine transliteration and Statistical machine transliteration. They discussed average entropy of dataset, which is uncertainty involved in mapping characters of language1 to characters of language2. They did transliteration from Hindi-to-English and then English-to-Hindi and compared the both. They concluded that Hindi-to-English transliteration is easier than English-to-Hindi transliteration.

Vijay MS et al [7] discussed the process of transliteration in two steps: source string is split into transliteration units and are then related with the target language units by providing different combinations of alignments and mappings. They also discussed Challenges faced in English-Tamil Transliteration. The process of transliteration is divided into three stages: pre-processing, training and transliteration.

Jiang et al [8] discussed an approach to enhance the named entity translation by joining transliteration approach with web mining. They discussed two approaches of transliteration. Rule-based approach assumes linguistic rules for translation. Statistics-based transliteration approach select the most predictable translations based on information gained from the training data. They build three-level transliteration model: English string is converted to Chinese Pinyin string, Chinese Pinyin string is converted to Chinese character string and Chinese character language model.

Rani and Laxmi [9] discussed Punjabi to Hindi transliteration. Punjabi and Hindi languages are not mutually comprehensible in written form, but are in spoken form. Mutual comprehensible languages is dependent on some factors like degree of similarities of phonemes, morphemes, syntax and lexemes. They discussed character to character mapping i.e. Direct mapping of Punjabi consonants and vowels into Hindi consonants and vowels respectively.

Matthews [10] described and evaluated an automatic transliteration system built by using Moses. Moses is a Statistical Machine Translation system which is used to automatically train translation models. SMT systems are trained using large parallel corpora and CLIR uses transliterations of unknown words and proper names. He discussed transliteration of English-Chinese and Arabic-English language pairs. For transliteration pre-processing of data is done then each corpora is split into training, development and test sets.

Das et al [11] discussed five different transliteration models that transliterates the text from an English word into an Indian language text. The transliteration models are Trigram Model (Tri), Joint Source-Channel Model (JSC), Modified Joint Source-Channel Model (MJSC), Improved Modified Joint Source-Channel Model (IMJSC) and International Phonetic Alphabet Based Model (IPA). In the first four models, the probability of transliterated text is calculated from the database If, value is not found, then a very small value is assigned to the probability. In the last model, a pre-processing module is used to check if the source English word is a valid dictionary word. Phoneme based transliteration module handles the dictionary words.

Das et al [12] developed a transliteration system that automatically learns the mappings from the bilingual NEWS training set which has knowledge of linguistic. The output obtained by mapping is a list which has the transliterated words with their probabilities. In the post processing step transliteration rules are applied. Some other rules are applied to produce more spelling variations.

IMPLEMENTATION

It describes the mapping scheme followed and approach selected in the present study. The mapping system, approach and rules are selected on the basis of literature review presented in Background.

HH. IAST

IAST is an acronym for International Alphabet of Sanskrit Transliteration. Although Sanskrit is officially related to Devanagari, is also written in several regional scripts of India and also in Roman (IAST) [21]. This standard is used for the transliteration of Sanskrit to Roman script. So it is a considered as standard for the Romanization of Sanskrit. This model uses diacritics. Diacritic is a sign, such as an accent or cedilla, which is written above or below a letter and indicates different pronunciation from the same letter when unmarked or differently marked. For example द is written as d and ड is written as ḍ. The mapping of Sanskrit characters (vowels and consonants) is done according to the following tables.

TABLE 2
IAST MAPPING OF VOWELS.

अA	आĀ	इI	ईĪ	उU	ऊŪ	ऋ Ṛ	ऌ Ḷ	एE	ऐAI
ओO	औAU								

TABLE 3
IAST MAPPING OF CONSONANTS

क$_{K}$	ख$_{KH}$	ग$_{G}$	घ$_{GH}$	ङ$_{\dot{N}}$	Velar
च$_{C}$	छ$_{CH}$	ज$_{J}$	झ$_{JH}$	ञ$_{\tilde{N}}$	palatal
ट$_{\underline{T}}$	ठ$_{\underline{TH}}$	ड$_{\underline{D}}$	ढ$_{\underline{DH}}$	ण$_{\underline{N}}$	retroflex
त$_{T}$	थ$_{TH}$	द$_{D}$	ध$_{DH}$	न$_{N}$	dental
प$_{P}$	फ$_{PH}$	ब$_{B}$	भ$_{BH}$	म$_{M}$	bilabial
य$_{Y}$	र$_{R}$	ल$_{L}$	व$_{V}$		
श$_{\acute{S}}$	ष$_{\underline{S}}$	स$_{S}$	ह$_{H}$		

II. Methodology

Sanskrit to English machine transliteration is done by using hybrid approach. Hybrid approach utilizes properties of both grapheme based and phoneme based approaches.

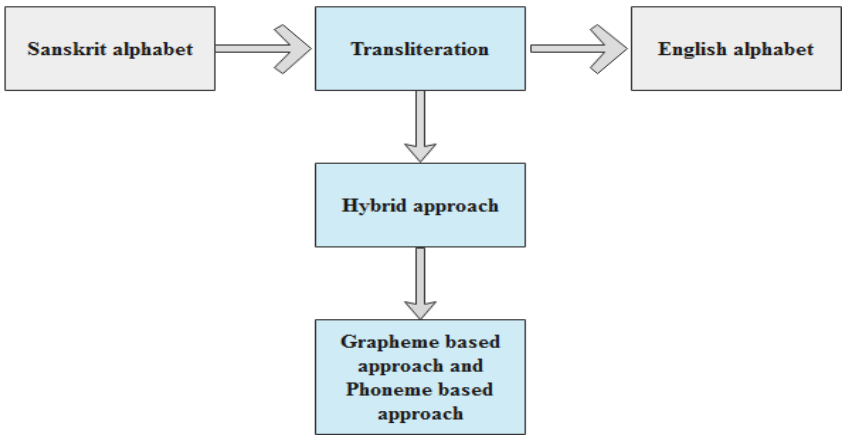


Fig. 7 Sanskrit to English machine transliteration by using hybrid approach

JJ. Flow chart

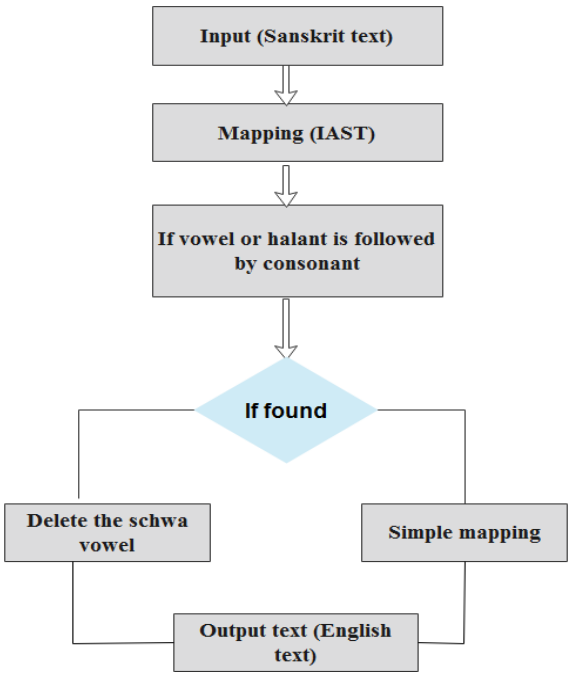


Fig. 8 Flow chart of Sanskrit to English Machine Transliteration process.

Each consonant letter not only represents only consonantal sound but also has an inherent vowel, for example क= क्+अ. Hence U+0915 Devanagari letter क not just represents 'k' rather 'ka'. When dependent vowel is present the inherent vowel that is related to the consonant letter gets overridden by the dependent vowel. Consonant letters can also be given as half-forms. These forms are used to represent the initial consonant. The half-form of a consonant does not have inherent vowel. These half forms are same as full consonant but the missing part is 'halant' [20]. Our system takes input the Sanskrit text and then the alphabets are mapped with English alphabet. If dependent vowel follows a consonant then its inherent schwa vowel is removed and only sound of dependent vowel is associated with the consonant and if there is no vowel that follows consonant then simple mapping is done.

RESULTS

The evaluation of the algorithm and mapping rules is shown in this section.

Precision and Recall

Precision and Recall are the two measures that are widely used to evaluate machine translation system. Precision is the measure of exactness and recall is the measure of completeness. Precision describes how accurate the system is and recall tells how complete the system is.

Precision (P) = C/(C+W)

Where C = Number of Correct analysis.

W = Number of Wrong analysis.

Recall (R) = C/(C+M)

C = Number of Correct analysis.

M = Number of Missed analysis.

F-measure

It is weighted harmonic mean between the precision and recall. It is weighted because in some applications one can care more about precision and recall. It is harmonic because it is conservative that is lower than arithmetic and geometric mean.

F= 2PR/ (P+R) which is equal to

2/(1/R+1/P)

Results

The results of the developed tool are evaluated by taking संस्कृत पुस्तक (छठी श्रेणी के लिए) पंजाब स्कूल शिक्षा बोर्ड , as Subject system. The accuracy is calculated by taking first ten chapters of the book.

TABLE 4

INDIVIDUAL RESULTS

Chapter	Total words taken	Number of right words	Number of wrong words	Accuracy (%)
Chapter-1	9	9	0	100%
Chapter-2	20	20	0	100%
Chapter-3	21	21	0	100%
Chapter-4	23	23	0	100%
Chapter-5	29	29	0	100%
Chapter-6	14	12	2	85.71%
Chapter-7	21	20	1	95.2%
Chapter-8	20	20	0	100%
Chapter-9	27	26	1	96.29%
Chapter-10	22	21	1	95.45%
Total	206	201	5	

TABLE 5

RESULTS

Precision	Recall	F-measure
97.5%	100%	98%

CONCLUSION AND FUTURE SCOPE

In this research work, we proposed an algorithm for transliteration from Sanskrit to English. The experiment results show that the algorithm has accuracy of 97.5%.

Conclusions

The following conclusions are drawn on the basis of experimental observations and analysis:

1) Mostly the two scripts differ from each other. In some cases, some sounds are missing and in some cases they are extra for target language. In those cases, we have to map the phonemes that are missing or extra to the letter which is

phonetically similar to it, *e.g.*, in Devanagari script, alphabet "घ् " that is not as such present in English script. So a letter that is similar in sound or letter combinations are used to map such sounds. So in case of '' घ्'' we use combination of letters "GH". The alphabets are mapped according to IAST.

2) Hybrid approach is used for transliterating Sanskrit to English.
3) Some rules like handling of inherent vowels are applied to our system.

Scope for future

In future, the present work may be extended on the following lines:

1. Backward transliteration that is transliteration from English to Sanskrit.
2. The system should be trained to identify the correct transliteration for the symbol ◌̇. Our system transliterates this symbol as Ṃ. The system should be trained to identify when to transliterate this symbol as Ṃ and when to transliterate it as N.

REFERENCES

[1] P. H. Rathod, M L Dhore, R. M. Dhore,Hindi and marathi to english machine transliteration using svm, International Journal on Natural Language Computing (IJNLC), 2(4), 2013, pp. 57-71.

[2] M.L. Dhore, S.K. Dixit, T.D. Sonwalkar, Hindi to English Machine Transliteration of Named Entities using Conditional Random Fields,International Journal of Computer Applications(0975 – 8887), 48(23), 2012, pp. 31-37.

[3] G.S. Lehal, A gurmukhi to shahmukhi transliteration system, in: proceedings of ICON-2009: 7th international conference on Natural Language Processing, 2009, pp.167-173.

[4] M.K. Chinnakotla, O.P. Damani, Character Sequence Modeling for Transliteration, in: proceedings of ICON-2009: 7th international conference on Natural Language Processing, 2009.

[5] S.Ganesh, S.Harsha, P.Pingali, V.Varma,Statistical Transliteration for Cross Langauge Information Retrieval using HMM alignment and CRF, in: proceedings of The 2nd International workshop on Cross lingual information access, 2008, pp. 42-47.

[6] M. K. Chinnakota, O.P. Damini, and A. Satoskar, Transliteration for Resource-Scarce Languages, ACM Transactions on Asian Language Information Processing, 9(4), Article 14, 2010.

[7] Vijaya MS, Shivapratap G, Dhanalakshmi V, Ajith VP, Soman KP, Sequence labeling approach for English to Tamil Transliteration using Memory-based Learning, in: proceedings of ICON – 2008: 6th International Conference on Natural Language Processing, 2008

[8] L. Jiang, M. Zhou, L.F. Chien, C. Niu,Named Entity Translation with Web Mining and Transliteration, in: proceedings of IJCAI-07 International Joint Conference on Artificial Intelligence, 2007, pp. 1630-1634.

[9] S. Rani and V. Laxmi, A Review on Machine Transliteration of related languages: Punjabi to Hindi, International Journal of Science, Engineering and Technology Research (IJSETR), 2(3), 2013, pp. 733-736.

[10] D. Matthews, Machine Transliteration of Proper Names, Masters Thesis, School of Informatics, University of Edinburg, 2007.

[11] A. Das, T.Saikh, T. Mondal, A. Ekbal, S. Bandyopadhyay,English to Indian Languages Machine Transliteration System at NEWS 2010, in: proceeding of The 2010 Named Entities Workshop, ACL 2010, Sweden, 2010, pp. 71–75.

[12] A. Das, A.Ekbal, T.Mondal, S.Bandyopadhyay,English to Hindi Machine Transliteration System at NEWS 2009, in: proceedings of the 2009 Named Entities Workshop, ACL-IJCNLP, Singapore, 2009, pp. 80–83.

[13] http://jrgraphix.net/r/Unicode/0900-097F, referred on 27th April, 2021.

[14] V.K. Gupta, N. Tapaswi, S. Jain, Knowledge Representation of Grammatical Constructs of Sanskrit Language Using Rule Based Sanskrit Language to English Language Machine Translation, in: Proceedings of International Conference on Advances in Technology and Engineering (ICATE), 2013.

[15] N. Tapaswi, S. Jain, Treebank Based Deep Grammar Acquisition and Part-Of-Speech Tagging for Sanskrit Sentences, in: Proceedings of CSI Sixth International Conference on Software Engineering (CONSEG), 2012.

[16] N. Tapaswi, S. Jain, Morphological and Lexical Analysis of the Sanskrit Sentences, MIT International Journal of Computer Science & Information Technology, 1(1), 2011, pp. 28-31.

[17] Shahnawaz, Conversion between Hindi and Urdu, in: Proceedings of International Conference on Computing, Communication and Automation(ICCCA2015), 2015, pp. 309-313.

[18] Antony P. J, Soman K P, Machine Transliteration for Indian Languages: A Literature Survey, International Journal of Scientific & Engineering Research, 2(12), 2011.

[19] D.A. Anand, S. Jana, Chronology of Sanskrit Texts: An Information- Theoretic Corroboration, in: Proceedings of National Conference on Communications (NCC), 2013.

[20] www.unicode.org, referredon April 7, 2021.

[21] D.Mishra, K.Bali, G.N.Jha, Syllabification and Stress Assignment in Phonetic Sanskrit Text, in : Proceedings of International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013.

[22] S. Bhavatu, Proposed Vedic Sanskrit Coding Scheme: Some suggestions, in: National Seminar on oral tradition and written text of Indian Languages, Mumbai, 2007.

[23] http://tdil.mit.gov.in/, referred on May, 2021.

[24] Rick Briggs, ”NASA article on Sanskrit in AI.” Spring , 1985.
[25] J. Ray, A review of terminological work being done in indian languages, in: Proceedings of Term banks for tomorrow's world : translating and the computer 4 : a conference jointly sponsored by Aslib, the Aslib Technical Translation Group, and the Translators' Guild of the Institute of Linguists, London, 1982.
[26] J. K. Raulji, J.k. R. Saini, Sanskrit Machine Translation Systems: A Comparative Analysis,International Journal of Computer Applications (0975 – 8887), 136(1), 2016.
[27] S. Karimi, F.Scholer, A. Turpin, Machine Transliteration Survey,ACM Computing Surveys, 43(3), 2011.
[28] G. S. Josan, J. Kaur,International Journal of Information Technology and Knowledge Management, 4(2), 2011, pp. 459-463.
[29] A. Kumar, Development of statistical approach based Hindi to Punjabi Machine Translation System, thesis, Punjabi University, 2014.
[30] http://www.gutenberg.org/files/41563/41563-h/41563-h.htm, referred on May, 2021.
[31] http://jrgraphix.net/r/Unicode/, referredon May, 2021.
[32] T. Kumar, Development of Shahmukhi to Gurumukhi Transliteration System, Thesis, Punjabi University, 2011.

BRAIN TUMOR DETECTION FROM MRI IMAGES USING DEEP LEARNING MODELS ANN AND CNN.

*Ravi Kumar Verma[*1], Dr Lakhwinder Kaur[#2],ER Navneet Kaur[#3]*
[#]Department of Computer Science and Engineering Punjabi University, Patiala, Punjab,India
[1]ravisadhak@gmail.com
[2]mahal2k8@yahoomail.com
[3]navneetmavi88@gmail.com

ABSTRACT:——Cancer in brain can be due to cells inside the brain which are abnormal. Magnetic Resource imaging (MRI) scans are general methods for the detection of brain tumor. Abnormal growth regarding tissues is detected from images of MRI. Algorithms of deep learning(DL) and machine learning(ML) are used to detect and identify brain tumor in many research papers. Fast prediction of brain tumor is achieved after using these algorithms as well as accuracy is very high as a result treatment of cancer patients is also fast. Doctors can take decision very fast after getting such type of predictions of ML or DL. In this paper we are implementing prediction models based on deep learning in which we are using Artificial Neural Network(ANN) and Convolutional Neural Network(CNN) to detect brain tumor in MRI images. At the end we will analyse performance of these two algorithms.

KEYWORDS: Artificial Neural Networks, Convolutional Neural Networks, Machine Learning, Deep Learning.

1. INTRODUCTION

Most important organ of body is brain, functionality of all the other organs is controlled by it and decision making process of body also relies upon it [1]. Overall control of central nervous system is in hands of brain and activities of human body whether they are voluntary or involuntary responsibility of their performance is on brain. Unwanted growth of tissues leads to a mesh which is fibrous, it is called tumor, it multiplies itself in an unrestricted way. Proper understanding of brain tumor and all of its stages plays an important role in the prevention and treatment of this life threatening disease. Brain tumors [2] are analysed by doctor or radiologist with the help of MRI magnetic resonance imaging [3]. In this paper we have applied deep learning algorithms [4] to identify whether the brain is normal or diseased and result is analysed. Normal and Abnormal brain is classified in this paper using Artificial neural networks and Convolutional Neural Networks. Artificial neural networks are based on internal functioning of human brain that's why interconnections inside ANN's are in large amount and processing units which are simple are applied on training set to train the neural network as well as knowledge is stored which is experimental in nature. In a neural network there are many hidden layers of neurons which are connected with each other [5]. In Deep Learning process a training data set is applied from which a model is built by neural network algorithm. Hidden or processing layers may be of any number [6] as it depends upon requirement but there is only one input and one output layer. In each layer neurons have their weight and bias which is totally dependent upon input features and for hidden and output layers it is dependent on previous layers. Activation function is applied on input features to train the model so that learning can be enhanced and output can be achieved as per expectation. In Ann layers are fully connected and amount of processing is very large and in the paper the dataset consists or images that's why we are also focusing on CNN as a Deep Learning Algorithm. Convolution is mathematical operation which is linear and used in Convolutional Neural Networks [7]. For every layer dimensionality reduction is performed for the image and information is not lost and performance of training process is not compromised. Model is created by applying different layers like Dense, Convolutional, dropout, flatten. The focus of this paper is on creating a model with ANN and CNN and comparison of these two algorithms is also performed after applying both on dataset of brain tumor images.

2. Literature Review

In this paper ANN is used to detect and classify Brain tumor [8]. Extraction of features, segmentation of images, enhancement of images and histogram equalization these techniques of image processing are used. In proposed work ANN is used as a classifier so that images can be classified and efficiency of ANN is more when compared with other classification algorithms. Accuracy is improved with specificity and sensitivity. Results which are acquired are good with not much computation time required.

[9]In this paper ,implementation of CNN Convolutional neural network is performed, accuracy obtained is more than 90% and recall is 99%,81% and 88% for pituitary, glioma and meningioma tumor respectively. Slices of MRI images are analysed with Deep Learning in which CNN as a classifier is used with convolutional layer which are used which are 2d in nature and three different types of tumors are classified. Acquisition of data, pre-processing of data, optimization of model and tuning of hyper parameters are applied in this paper. Generalizability of the model is checked by performing cross validation which is 10 fold on the whole dataset.

[10]Hough voting is the basic fundamental behind the method proposed in this paper, in Hough Voting localisation and anomalies of interest segmentation is completely automatic. Segmentation method which is used which is based on learning techniques its robustness, flexibility, multi-region properties and adaptation to different modalities cannot be under estimated. Final results are predicted by applying more than two amount of training data as well as different dimensionalities like 2d,3d etc. are applied. Image is analysed using Convolutional neural networks [10], CNN with Hough voting, classification which is Voxel-vise and evaluation which is patch wise using CNN.

[3] Most essential organ in body is brain the control and coordination of working of other parts of body is dependent on brain. It acts as a centre of control for nervous system. All the activities of human body which are voluntary or involuntary, brain is responsible for their performance. Brain tumor is a mess like structure which multiplies itself in the way which is unstoppable. Magnetic Resonance Imaging(MRI) is used by doctors to identify whether there is tumor or not. After analysis it is revealed that whether tumor is present or not. When MRI images are properly analysed, brain tumor can be cured at fat speed and society will be benefitted.

3. Dataset

Dataset which we have used is taken from Kaggle website. In this dataset we have MRI images. Data set is comprise of two folders name Yes and No respectively in first one we have images with brain tumor and in second one image without brain tumor. Total number of images in the dataset are 253.In Figure 1 two images are shown one with tumor and another without tumor. There 155 tumorous images and 98 non tumours images. The images are of different shapes (eg. 172X201,630X630) these images are resized to 128x128.No of image for training were 202 and for testing were 51.

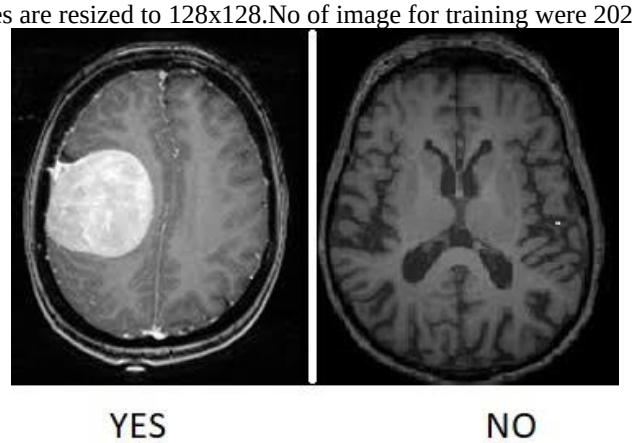


Figure 1 Tumorous and Non Tumorous Brain

4. Implementation

We have applied two Deep Learning Techniques Artificial Neural Networks and Convolutional Neural Networks and then we analyse their performance i.e. which one is better in classifying brain tumor images.

A. *Following are the steps while implementing ANN*

1)All the python packages which are needed are imported.
2)Images from the dataset are imported from yes as well as no folder.
3)Then images are labeled using one hot encoding i.e 1 for having brain tumor and 0 for not having brain tumor.
4)Size of all the images will be changed to 128x128
5)Normalization of images is performed.
6)Splitting of Data into training, test and validation set is performed.
7)Model creation using python Keras Library
8)Compilation of model.
9)Model applied on training data.
10)Then model evaluation is performed on test set.

In our ANN implementation there are 7 layers. First layer in our model in Dense layer in which activation function is relu and input shape is 128x128x3 as well as 32 neurons. Then we have flatten layer as our second layer using which images are converted into array of single dimension. Our next four layers are dense layers with 64,64,64,32 neurons respectively and these are hidden layers of our ANN model and the last layer of our model is output layer in which there are 2 neurons and activation function as Softmax.Compilation of model is done using adamax as optimizer and loss function is categorical crossentropy. Generation and training of model is based on validation images and training images. After training testing is performed using test data.

B. Convolutional Neural Network is then implemented on the same data which was applied on ANN.

Steps which are taken to implement CNN are:

1)All the python modules which are needed are imported.
2)Dataset folder is imported having subfolders Yes and No for tumorous and non-tumorous images respectively.
3)Class Labels are given to images i.e. 1 Brian Tumor and 0 Non Tumor using one hot encoding.
4)Resizing all the images into (128x128)
5)Image Normalization
6)Dataset divided into training, validation and test data.
7)Sequential model using keras library of python is created with convolutional layers.
8)Compilation of model.
9)Model Applied on Training Data and evaluation is performed by Validation data.
10)Test data images are used for model evaluation.
11)Training Accuracy and Validation accuracy is visualised with the help of graph.

Different layers are implemented to create a sequential model for CNN.Shape of image is changed to 128x128.Input image is processed by convolutional layer in which activation function is relu. To make the output image looks same as input image padding in the convolve layer is set to same. Filters in convolutional layers are 32,32,64,64 respectively. Maxpooling layer is implemented in which window size is 2x2.There are two layers with dropout function with dropout value 0.25 and 0.5 respectively. A Flatten layer is added so that there is conversion of feature into array which is one dimensional. Then we have a dense layer which is fully connected in which number of neurons are 900.In our output layer we have two neurons and softmax activation function is used. Our CNN Architecture is showed in Figure 2. Our code is implemented with Keras Deep learning library of Python programming language and we have used Jupiter Notebook for writing and executing our code. Modal is trained using 30 epochs. After storing the history of model execution we have plotted graphs of Loss and accuracy to understand the model well.

5. Analysis of Experimental Results

A variable is created and all image data is stored in it in the form of numpy arrays. Then labels for images that is 1 for tumor and 0 for no tumor are stored in variable result using one hot encoder. Whole data is then appended to a list named data. Division of dataset set is performed into training, testing and validation dataset. In Figure 3 loss and accuracy of ANN model is presented when applied on training and validation data. After applying ANN model on training data, with 30 epochs accuracy is 99.1% and for validation data accuracy is 67.86%.

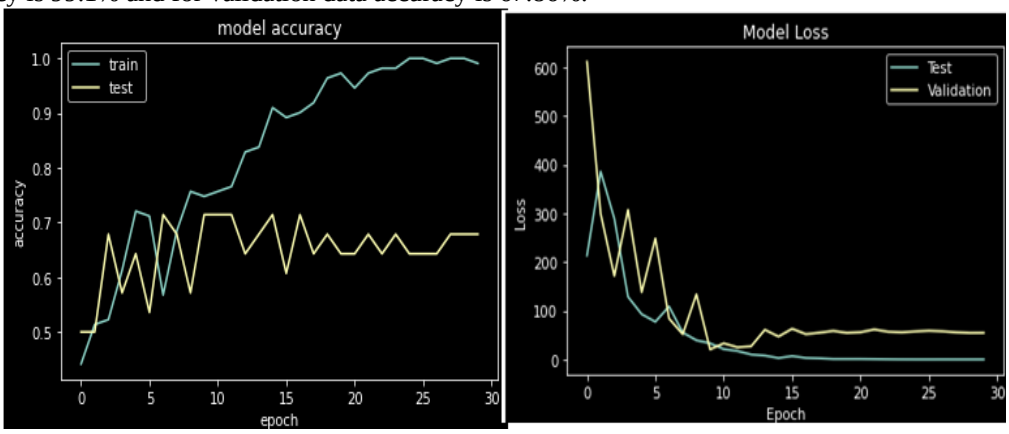


Figure 3 : Comparision of Loss and Accuracy of Training/validation after implementing ANN Model

In Figure 4 loss and accuracy of CNN model is presented when applied on training and validation data. After applying CNN model on training data, with 30 epochs accuracy is 100% and for validation data accuracy is 82%.

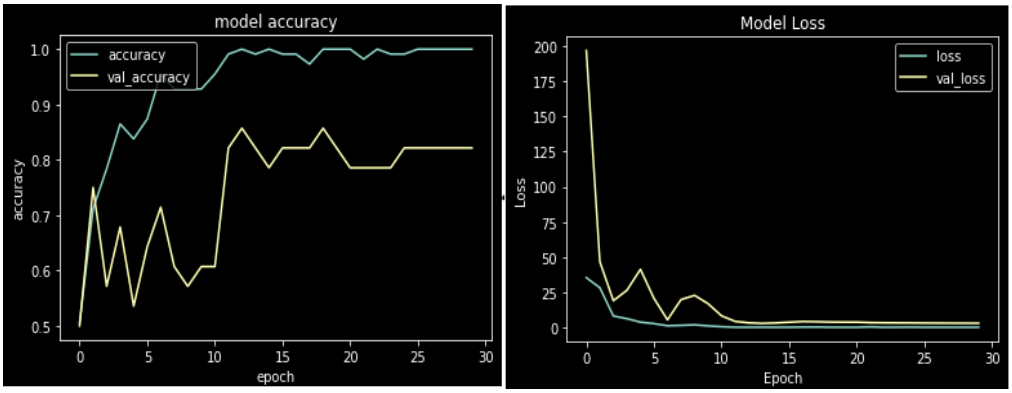


Figure 4 : Comparision of Loss and Accuracy of Training/validation after implementing CNN Model.

6. Testing model practically

A. ANN

In figure 5 we have inputted an image from no folder to out ANN model

```
94]: def ann_result(imgx):
      imgarray = np.array(imgx.resize((128,128)))
      imgarray = imgarray.reshape(1,128,128,3)
      resultx = model.predict_on_batch(imgarray)
      classdata = np.where(resultx == np.amax(resultx))[1][0]
      imshow(imgx)
      print(str(resultx[0][classdata]*100) + '% Sure This Is ' + info(classdata))

imgx = Image.open(r"C:\Users\DELL\Desktop\Final\brain_tumor_dataset\no\n17.jpg")

ann_result(imgx)
```

```
100.0% Sure This Is No, Its not a tumor
```

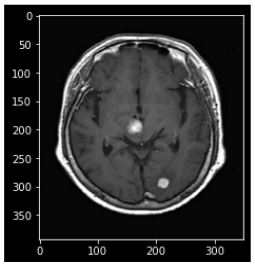


Figure 5 : Read operation on image from no folder using ANN

If we look at the output in figure 5 our model I predicting that there is no tumor.

In figure 6 we have inputted an image from yes folder to out ANN model

```
95]: def ann_result(imgx):
      imgarray = np.array(imgx.resize((128,128)))
      imgarray = imgarray.reshape(1,128,128,3)
      resultx = model.predict_on_batch(imgarray)
      classdata = np.where(resultx == np.amax(resultx))[1][0]
      imshow(imgx)
      print(str(resultx[0][classdata]*100) + '% Sure This Is ' + info(classdata))

imgx = Image.open(r"C:\Users\DELL\Desktop\Final\brain_tumor_dataset\yes\y3.jpg")

ann_result(imgx)
```

```
100.0% Sure This Is Its a Tumor
```

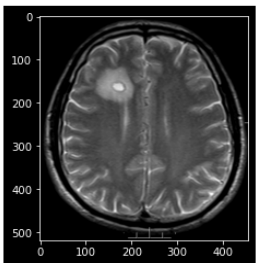


Figure 6 : Read operation on image from Yes folder using ANN

If we look at the output in figure 6 our model is predicting that there is tumor

B. CNN

In figure 7 we have inputted an image from no folder to our CNN model

```
]: def cnn_result(imgx):
      imgarray = np.array(imgx.resize((128,128)))
      imgarray = imgarray.reshape(1,128,128,3)
      resultx = model.predict_on_batch(imgarray)
      classdata = np.where(resultx == np.amax(resultx))[1][0]
      imshow(imgx)
      print(str(resultx[0][classdata]*100) + '% Sure This Is ' + info(classdata))

imgx = Image.open(r"C:\Users\DELL\Desktop\Final\brain_tumor_dataset\no\n17.jpg")

cnn_result(imgx)
```

```
100.0% Sure This Is No, Its not a tumor
```

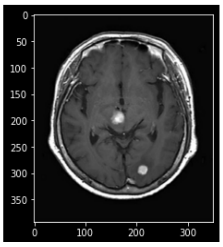


Figure 7 : Read operation on image from no folder using CNN

If we look at the output in figure 7 our model is predicting that there is no tumor.

In figure 8 we have inputted an image from yes folder to our CNN model

```
]: def cnn_result(imgx):
    imgarray = np.array(imgx.resize((128,128)))
    imgarray = imgarray.reshape(1,128,128,3)
    resultx = model.predict_on_batch(imgarray)
    classdata = np.where(resultx == np.amax(resultx))[1][0]
    imshow(imgx)
    print(str(resultx[0][classdata]*100) + '% Sure This Is ' + info(classdata))

imgx = Image.open(r"C:\Users\DELL\Desktop\Final\brain_tumor_dataset\yes\y3.jpg")

cnn_result(imgx)
```

```
100.0% Sure This Is Its a Tumor
```

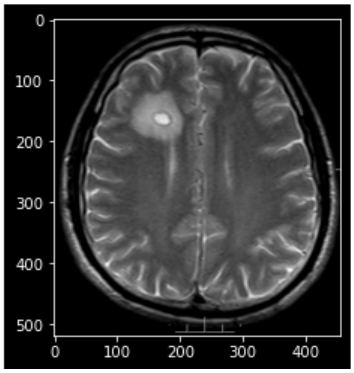


Figure 8 : Read operation on image from no folder using CNN

If we look at the output in figure 8 our model is predicting that there is tumor.

7. Conclusion

Whenever there is need to analyse a dataset in which there are images CNN is the best choice as it is basically deigned to classify images. In CNN actual size of the image is reduced but there is no compromise with the information provided by the image which is required in the prediction process. In CNN we get testing accuracy of 82%. After applying ANN we get testing accuracy of 67%, data must be increased to get more accuracy. We can also take help of image augmentation in which similar images can be generated from already existing images. We have used trial and error approach in our implementation. In future work we can use optimization techniques to decide the optimal number of filters and layers. In current work accuracy of CNN is high as compared to ANN on given dataset.

REFERENCES

[1] Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., Ashburner, J. T., & Penny, W. D. (2004). *Human brain function*. Elsevier.

[2] Kaye, A. H., & Laws, E. R. (1995). *Brain tumors: An encyclopedic approach*. Saunders.

[3] Raheleh Hashemzehi Seyyed Javad Seyyed Mahdavi Maryam Kheirabadi Seyed Reza Kamel 2020 *Detection of brain tumors*

[4] A. Shrestha and A. Mahmood, *"Review of Deep Learning Algorithms and Architectures,"* in IEEE Access, vol. 7, pp. 53040-53065, 2019, doi: 10.1109/ACCESS.2019.2912200.

[5] Alzubaidi, L., Zhang, J., Humaidi, A.J. *et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions*. *J Big Data* **8,** 53 (2021). https://doi.org/10.1186/s40537-021-00444-8

[6] Panchal, F.S., & Panchal, M. (2014). *Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network*.

[7] N. Aloysius and M. Geetha, *"A review on deep convolutional neural networks," 2017 International Conference on Communication and Signal Processing (ICCSP)*, 2017, pp. 0588-0592, doi: 10.1109/ICCSP.2017.8286426.

[8] Rajeshwar Nalbalwar Umakant Majhi Raj Patil Prof.Sudhanshu Gonge 2014 *Detection of Brain Tumor by using ANN*

[9] Fausto Milletari Seyed-Ahmad Ahmadi Christine Kroll Annika Plate Verena Rozanski Juliana Maiostre Johannes Levin Olaf Dietrich Birgit Ertl-Wagner Kai Bötzel, Nassir Navab 2016 *Hough-CNN: Deep learning for segmentation of deep brain regions in MRI and ultrasound* Elsevier Inc 164 92-102.

[10] Dena Nadir George Hashem B Jehlol Anwer Subhi Abdulhussein Oleiwi *2015 Brain Tumor Detection Using Shape features and Machine Learning Algorithms* International Journal of Scientific & Engineering Research 6 12 454-459.

INTRUSION DETECTION USING DEEP LEARNING TECHNIQUES: A REVIEW

Er.Navroop Kaur[1], Meenakshi Bansal[2],Sukhwinder Singh[3]
Research Scholar, Punjabi University patiala[1]
Punjabi University, YCOE, Guru Kashi Campus, Talwandi Sabo[2]
Punjabi University, YCOE, Guru Kashi Campus, Talwandi Sabo[3]

ABSTRACT:— A network intrusion detection system is an analytical and demanding element of every internet-connected system because of attacks from external and internal sources. The most catastrophic cybercrime is those which cause loss of an organization's intellectual property and disruption's to a nation's national infrastructure. A considerable amount of attention has been given to machine learning and deep learning techniques for detecting known and unknown attacks on the computer system. This paper provides an overview of machine learning and deep learning techniques and the numerous datasets employed in intrusion detection systems. By evaluating several research papers in which Deep Learning has been successfully integrated into intrusion detection systems, this review distinguishes Deep Learning and machine learning models. This paper also presents recent advancements in the IDS datasets, which can be utilized as a manifesto by many research communities for generating efficient and effective new IDS datasets.

KEYWORDS: Deep Learning, Intrusion Detection Systems, Anomaly Based Detection

1. INTRODUCTION

With the growth of technology and globalization, the challenge to keep the system safe is more considerable. Because of these internet attacks, numerous damages and losses of billions of dollars happened. So cyber security especially intrusion detection systems have higher importance for dealing with these types of attacks. A network intrusion detection system(NIDS) used along with a firewall is a software application that monitors the network traffic, detects attacks and security bugs, and gives its information to the administrator. As per the CISCO network gauge report, the overall organization traffic in 2016 was 96 EB/month and is required to arrive at 278 EB/month in 2021 (Viegas et al., 2019). In the future, there are plans to support a bandwidth of 400 Gbps. Intrusion detection systems (IDS) use a sample data distribution to create an intrusion detection model to detect unauthorized actions on systems and computer networks that may constitute a threat to information. When such attacks are discovered, an IDS sends out an alarm. An IDS' major goal is to distinguish between malicious and non-malicious network traffic, something that standard firewalls can't do. An IDS can be actualized as signature-based, anomaly-based IDS (Liao et al., 2013). **In signature-based** IDS intrusions are detected by comparing monitored behavior with pre-defined intrusion patterns. It is also known as misused detection or knowledge-based detection due to the use of knowledge collected from previous intrusions and vulnerabilities Nonetheless, this technique isn't adequate to recognize obscure interruptions and variations of known ones (Li et al., 2017). While Anomaly-based IDS focuses on knowing normal behavior to identify any deviation (Liao et al., 2013). Different techniques are used to detect anomalies, such as statistical-based, knowledge-based, and machine learning techniques. Anomaly-based discovery comprises three general modules parameterization, training, and detection (García-Teodoro et al., 2009). The challenges with anomaly-based intrusion detection are that it needs to deal with novel attacks for which there is no prior knowledge to identify the anomaly.

Network monitoring is used broadly for forensics, anomaly detection, and security. Several issues have been created in recent years and become a barrier for NIDS and some of these areas following Volume, Diversity, Accuracy, Low-frequency attacks, Dynamics, Adaptability. Machine learning has been used highly in computer science for face, image, and voice recognition. Various reasons inhibit its practical application of machine learning to the IDS. First, the required accuracy is high because the false detection risk is very high on IDS. Second, the large number of alerts generated by intrusion detection can be a significant burden for internal teams. Most intrusion detection systems come with a set of pre-defined alert signatures, but for most organizations, these are insufficient and extra effort is needed to baseline activities explicit to each environment. Third, An IDS is only as good as its signature library. If it isn't updated frequently, it won't register the latest attacks and it can't generate alerts about them, so the latest attacks will always be a big concern.

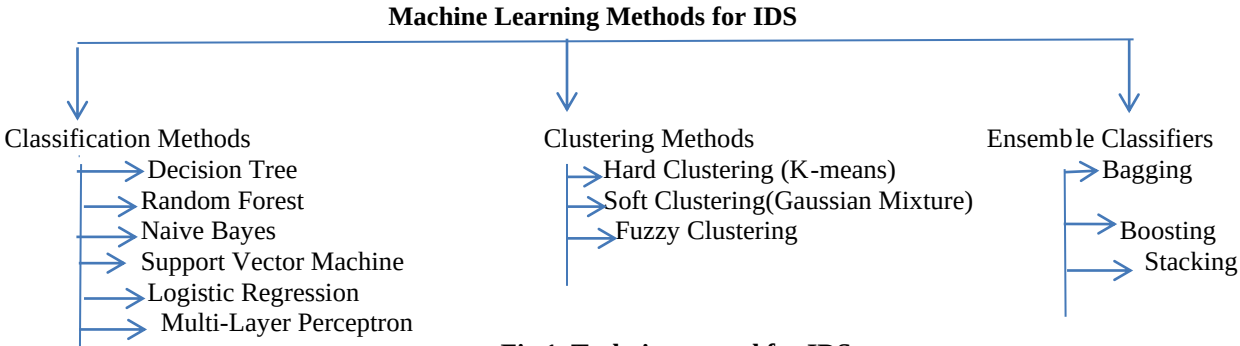


Fig 1. Techniques used for IDS

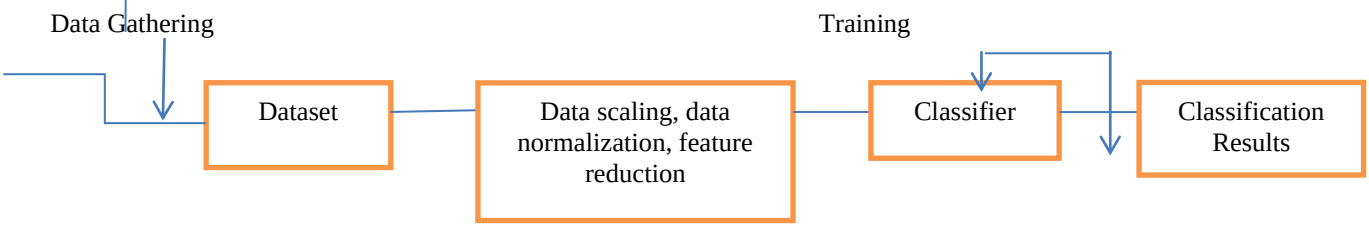


Fig2.Classification Procedure

Classifiers

In Fig 1. various machine learning techniques are classified. Machine learning classifiers are decision tree, random forest, support vector machine, naive Bayes, logistic regression, etc and deep learning classifiers are autoencoder, convolutional neural network, artificial neural network, etc. For the classifier, two significant datasets, for example, a training dataset and a testing dataset are required. The training dataset is used to train the classifier and the testing dataset is used to test the performance of the classifier(Kwon et al., 2019).

Fig 2. represents the classification procedure, the dataset is firstly normalized, feature scaling and feature extracting is done before training the dataset, after training testing is done to classify the data (binary or multiclass). Ensemble learning is a machine learning mechanism where multiple weak classifiers are trained to solve the same problem by reducing bias and variance and then combined to get more accurate results. Bagging and boosting are both considered homogeneous weak classifiers but stacking considers heterogeneous weak learners. In bagging and stacking weak classifiers learns parallelly but in boosting weak learners sequentially. In bagging and boosting output is combined using some deterministic strategy, in stacking training a meta-model to output a prediction based on the different weak models' predictions

1.2.1 Bayesian Network

A Bayesian Network breaks up a probability distribution based on conditional independencies. Networks that use Bayesian probability theory to control model complexity, optimize weight decay rates, and automatically find the most important input variables(Azuaje, 2006).

1.2.2 Naive Bayes

Naive Bayes is a kind of Bayesian network and is a commonly used machine learning algorithm (Azuaje, 2006). It is a basic probabilistic-based technique that calculates the probability to classify or predict the cyber-attack class in a given dataset. This method assumes each feature's value as independent and considers the correlation or relationship between the features. Naive Bayes includes two probabilities - one is the conditional probability, and another one is class probability. Class probability is determined by dividing the frequency of each class instance by total instances. Conditional probability is the ratio of the occurrence of each attribute for a given class and the occurrence of samples for that class. Naive Bayes is faster than other classifiers. The number of parameters required by Naive Bayes classifiers is linear in the number of variables in a learning problem.

1.2.3 Decision Tree

The decision tree is one of the most popular classifications and prediction algorithms in machine learning (X. Z. Wang et al., 2006). A Decision Tree is a tree-like structure, in which an internal node represents attributes, and branches represent the outcome, and a leaf represents a class label. For classification and regression, Decision Trees (DTs) are a non-parametric supervised learning method. The goal is to learn simple decision rules from data attributes to develop a model that predicts the value of a target variable. It's easy to comprehend and interpret. Decision Trees can handle missing values successfully, and they are resistant to outliers. Overfitting of data causes Decision Tree to make incorrect predictions. To deal with the data, it creates new nodes regularly, resulting in an overly complex tree. Overfitting increases the likelihood of large variance and thus inaccuracy. When a new data point is introduced, the tree must be rebuilt, and all nodes must be computed and reconstructed. With a modest quantity of noise, the Decision Tree becomes unstable and produces incorrect predictions.

1.2.4 Random Forest

Random Forest is a classifier comprising of decision trees operated as ensemble learning. The reason is that it combines both the different sets of data called bootstrap aggregation and also numerous features selection, to predict the outcome. Additionally, random forest is the mix of single model trees, where every hub contains k haphazardly picked credits in the tree (Alqahtani et al., 2020). DT has high variance and low bias leading to undesired outputs. RF can focus on both, the observations and variables of training data for evolving distinct decision trees and take maximum voting for classification and the total average for regression problems respectively. RF makes use of the bagging technique that randomly considers observations and chooses the columns incompetent of signifying noteworthy variables at the root for all the DTs.

1.2.5 Support vector machine (SVM)

SVM maps real-valued input feature vector to a higher dimensional features space through nonlinear mapping and can provide real-time detection capability, deal with large dimensionality of data, can be used for binary –class as well multiclass classification. Support Vector Machines are a collection of supervised learning algorithms for regression, classification, and outlier detection. The novelty of SVM resides in the way it selects the decision border that maximizes the distance between the classes under the study's adjoining data points. The maximum margin hyperplane is the decision boundary that is constructed. The data points on one side of the created line represent one category, while the data points on the other side of the line represent another. SVM is incapable of handling text structures. A fundamental disadvantage of SVM is the presence of multiple kernels, which makes it difficult to pick the best decision (Deepa & Kavitha, 2012).

2. DEEP LEARNING

Because harmful attacks are so quantified and complicated, some flaws in classic machine learning algorithms have been reinforced, such as the emphasis on processing low-dimensional data and the lack of reaction to high-dimensional data, as well as the reliance on manual feature selection. Deep learning has eliminated the manual feature selection procedure, which has become a feasible option for machine learning jobs that process high-dimensional data. Deep learning techniques in the field of intrusion detection have advanced substantially in recent years. Deep learning was motivated by the structure and depth of the human brain. The reason behind the learning of the network to map the input features to the output is the multiple levels of abstraction. Major focus on deep networks in the domain of deep learning, where classification training is conducted by training with multiple layers in hierarchical networks using unsupervised learning. Deep network intrusion detection systems can be classified based on how the architectures and techniques are being used for classification or feature reduction or both. Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and the Automatic Encoder (AE) are examples of algorithms that have enhanced the accuracy and simplicity of intrusion detection (Lin et al., 2018). Profound learning designs are arranged into three primary classifications: generative, discriminative, and hybrid structures.

2.1 Generative Architectures

Generative Models are a class of unsupervised learning models where training data is given and we aim to try and generate some new data points from the same distribution with some variation. The most common architectures which are coming under this category are explained below.

2.1.1Multilayer Perceptron (MLP)

MLP is a type of multiple hidden layer architecture with a fully connected feed-forward network in which each layer has nonlinear neural units. Because of this non-linearity, it becomes very difficult to train a deep MLP. MLP based models get re-established recently because of various techniques given by the deep learning community (Lin et al., 2018). This includes Stochastic Gradient Descent (SGD), Adam streamlining agent(Kingma & Ba, 2015), cluster standardization (Deepa & Kavitha, 2012), and Dropout. Initially, Deep Neural Network (DNN) is developed by forming a stack of linear classifiers which is a type of most basic type of multilayer perceptron. Any model that consists of more than 3 layers is called a deep network. In DNN, Inputs are taken by the model, multiplied with weights, and then passed to the activation function. In DNN this whole process is repeated over multiple layers.

2.1.2 Restricted Boltzmann Machine (RBM)

The Boltzmann machine (BM) is a probabilistic neural network introduced by Hinton and Sejnowsk (Ackley et al., 1985). A BM network consists of binary units paired symmetrically and decides which of them are activated. However, there are several connections among units, which ends in terribly slow learning. RBM is a unidirectional model projected by Smolensky in 1986 to unravel problems arising from the quality of BM. The principle behind RBM is to eliminate or restrict the connections among neurons within the same layer. Each Boltzmann machine and RBM incorporates a layer of hidden units connected to a layer of visible units. Feature distribution from input variables is learned by hidden layers. RBM can be used as a feature extractor and also as a classifier. Deep Boltzmann machine (DBM) is defined as when more than one RBM is cascaded. When stacking of RBM is implemented, trained in a greedy layer-wise fashion and each RBM is trained on top of the previous one, where the input to the next RBM is going through each hidden layer of the previous RBM is called a deep belief network (DBN).

2.1.3 Auto Encoder (AE)

An AE may be a deep neural network introduced by Holden (Atefinia & Ahmadi, 2021), generally used for reducing the dimension of raw data for producing improved data representation. Anne has both input and output layers with an equal number of feature vectors. An AE has some feature vector in addition to a hidden layer with low dimensional feature representation. The encoder extracts the features which are raw and it learns data representation by converting the input into low-dimensional abstraction. The decoder then reconstructs the original features from low dimensional representation. There are several AE extensions, stacked AE (SAE), sparse AE, and de-noising AE.

2.1.4 Recurrent Neural Network (RNN)

RNN may be a dynamic feed-forward neural network introduced by Hopfield in 1982. It's distinguished by its ability to find out sequent information over time steps. The output of every hidden layer in RNN is predicated on the present time step input and also the output of the previous time step. Every hidden unit encompasses a feedback circuit that passes the unit output back to an identical unit to be related to consecutive timesteps (Lipton et al., 2015). RNN can be used for either supervised or unsupervised learning. RNNs are extended with totally different memory unit variants, as well as long-short time memory (LSTM) and gated recurrent unit (GRU).

3. DISCRIMINATIVE ARCHITECTURES

Discriminative architectures are mainly supervised architectures used for labeled information to distinguish patterns for prediction tasks. Following architecture are the most common discriminative deep learning architectures.

Convolutional Neural Network (CNN)

CNN's were introduced to handle rigorous or complex connections between deep neural network (DNN) layers. CNN's train multiple layers with nonlinear mappings to classify high-dimensional input data into a collection of different categories and different sets of classes at the output layer. A CNN is represented by convolutional layers and pooling layers, followed by totally connected layers. Convolutional layers include filters which are used to represent smaller dimensional slices of input data. The filters deform over the full contribution to make feature maps. The pooling layer then operates over the feature maps to perform sub-sampling that reduces the spatiality of the feature maps (Cleary, 2019).

4. HYBRID ARCHITECTURES

Hybrid architectures incorporate both generative and discriminative models. This takes advantage of generative features in early phases and discriminative features in later stages to distinguish data.

Generative Adversarial Network (GAN)

A GAN may be a hybrid deep network introduced by a gaggle of researchers at Google Brain in 2014. A GAN is an inner cycle of 2 networks: a generative network and a discriminative network. A GAN is sort of a minimax game during which one network seeks to maximize the performance price and also the different tries to attenuate it. In every antagonistic spherical, the generator produces irregular examples from the commotion (Goodfellow et al., 2020).

5 PERFORMANCE METRICS FOR INTRUSION DETECTION SYSTEMS AND DATASETS FOR NETWORK INTRUSION DETECTION

- True Positives (TP) - These are the correctly predicted positive values which mean that the value of the actual class is yes and the value of the predicted class is also yes.
- True Negatives (TN) - These are the correctly predicted negative values which mean that the value of the actual class is no and value of the predicted class is also no.
- False Positives (FP) – When the actual class is no and the predicted class is yes.
- False Negatives (FN) – When actual class is yes but predicted class in no.

Generally, equations (1) and (2) Detection Rate (DR) and False Alarm Rate (FAR) are used as the metrics of IDS evaluation. The ratio between the number of accurately predicted attacks and the total number of attacks is used to calculate it. The TPR is 1 if all intrusions are detected, which is exceedingly unlikely for an IDS. A Detection Rate is another name for TPR (DR)

$$Detection\ Rate(DR) = \frac{True\ Positive(TP)}{True\ Positive + False\ Negative(FN)} \quad (1)$$

$$False\ Alarm\ Rate = \frac{False\ Positive\ \ (FP)}{False\ Positive(FP)\ + True\ Negative(TN)} \quad (2)$$

$$Accuracy = \frac{True\ Poditive(TP) + True\ Negative(TN)}{True\ Positive(TP) + True\ Negative(TN) + False\ Positive(FP) + False\ Negative(FN)} \quad (3)$$

The ratio between the number of normal instances mistakenly categorized as an attack and the total number of normal instances is used to determine the False Positive Rate (FPR). In equation (3) the classification rate or accuracy assesses how well the IDS detects regular and abnormal traffic patterns. It's defined as the proportion of accurately predicted cases among all instances.

There are several publicly available datasets for intrusion detection but the most generally utilized dataset is the KDD99 dataset. The KDD99 dataset has been gotten in 1999 from the DARPA98 network traffic dataset. It was the benchmark dataset utilized for the worldwide knowledge discovery and data mining tools competition, and the most well-known dataset that has ever been utilized in the interruption location field. Every TCP association has 41 labels with a name that indicates the status of an association as either being ordinary or a particular attack type. There are 38 numeric labels and 3

Table 1. Overview of Network-Based datasets

Data Set	Developed By	Year of Traffic Creation	Features	Realistic traffic	Label data	Attacks
KDD99	University of California	1998	41	Yes	Yes	DoS, Privilege escalation, probing
NSL-KDD	University of California	1998	41	Yes	Yes	DoS, Privilege escalation, probing
AWID	University of AEGEAN	2015	155	Yes	Yes	Popular attacks on 802.11 like authentication requests, ARP flooding
UNSW-NB15	Cyber Range Lab of UNSW Canberra	2015	42	Yes	Yes	Backdoors, DoS, exploits, fuzzes, generic, port scans, reconnaissance, shellcode, spam, worms
CIC-IDS-2017	Canadian Institute of Cyber Security	2017	80	Yes	Yes	Brute force, portscan, Botnet, Dos, DDos, Web Infiltration
CSE-CIC-2018	Canadian Institute of Cyber Security	2018	80	Yes	Yes	Brute force, portscan, Botnet, Dos, DDos, Web Infiltration

non-numeric labels. Table 1 list the datasets which are mostly used in network intrusion detection. According to (Aldweesh et al., 2020) 34% of researchers used the KDD99 benchmark dataset for intrusion detection systems and 37% of researchers used NSL-KDD for intrusion detection systems. Only 5% used real-time data either from the simulated or real environment. The basic and most important characteristics for developing a dataset are network configuration. network traffic, labeled dataset network, network interaction, capturing the traffic, protocols, and feature metadata.

6 Deep Learning-Based Intrusion Detection Systems

Table 2 below lists the work done in the ongoing years for performing intrusion detection, utilizing different profound learning-based structures.

Table 2. Deep Learning-Based Intrusion Detection Techniques in the Recent Years

Author	Year	Dataset used	Deep learning Model used	Results
(Yu & Bian, 2020)	2020	Binary and multiclass classification, using 1% NSLKDD train + for training and UNSWNB15 dataset.	Proposed multi-stage deep feature learning for intrusion detection using DNN, CNN as embed function for feature extraction and dimensionality reduction with random sampling technique	Achieved high accuracy of 92.34% for KDD test+ and 85.75% for KDD test 21. The detection rates for U2R and R2 L were increased from 13 to 81.50 and 44.41 to 75.93%.
(Huang & Lei, 2020)	2020	NSL-KDD, UNSW-NB15, CICIDS 2017	Proposed IGAN to tackle the class imbalance problem in intrusion detection by generating samples only for the minority class	The proposed IDS perform best as compared to the other 15 machine learning and deep learning IDS
(Kwon et al., 2019)	2019	NSL-KDD data set	Proposed deep learning model by using fully connected network (FCN)for analyzing network traffic and for improving accuracy	Accuracy was above 90% with FCN that is much improved in comparison to 50.3-82.5 by using SVM, Random forest, and ADA boosting

(Y. Wang et al., 2018)	2018	NSL KDD data set	Used random split and fuzzy C-means (FCM) clustering technique with 3 layers of SVM	The accuracy of the Proposed model was 99.41% with FCM and 99.16 with a random split.
(Diro & Chilamkurti, 2018)	2018	ISCX2012	10 Layers of LSTM, 30 embedding layers and sigmoid output	Accuracy=99.91% Precision=99.85% Recall=99.96%
(Xu et al., 2018)	2018	KDD99,NSL-KDD	GRU and bidirectional GRU,	BGRU gives best results with fast convergence. On NSL-KDD: Acc = 99.24%, Rec = 99.31%, FAR = 0.84%
(Moraboena et al., 2020)	2018	KDD99(10% subset)NSL-KDD(5 and 13 class classification)	Stacked AE(unsupervised training) followed by RF AE+RF	Accuracy = 85.42%, Precision =100%, Recall = 85.42%, F1 = 87.37%, FAR = 14.58%
(Abeshu & Chilamkurti, 2018)	2018	NSL-KDD ,Binary Classification	Stacked auto Encoder followed by softmax	Accuracy = 99.2%, Recall = 99.27%, FAR = 0.85%
(Tang et al., 2016)	2016	DNN for SDN environment	DNN with 3 hidden layers	Accuracy = 75.75%

7. CONCLUSION

The paper has examined the varied Deep Learning Models to assist the detection of malware and unwanted traffic. Because of the increase in network traffic and data, technology switches from data mining to machine learning and machine to deep learning. It will be seen from the classification that Autoencoders and Deep Neural Networks are performing well. RNN based mostly strategies will be incorporated in models for improved accuracy. In the future researchers can work in adversarial deep learning algorithms and ensemble deep learning algorithms for improving the accuracy of an intrusion detection system.

REFERENCES

1. Abeshu, A., & Chilamkurti, N. (2018). Deep Learning: The Frontier for Distributed Attack Detection in Fog-To-Things Computing. *IEEE Communications Magazine*, *56*(2), 169–175. https://doi.org/10.1109/MCOM.2018.1700332
2. Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, *9*(1), 147–169. https://doi.org/10.1016/S0364-0213(85)80012-4
3. Aldweesh, A., Derhab, A., & Emam, A. Z. (2020). Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues. *Knowledge-Based Systems*, *189*, 105124. https://doi.org/10.1016/j.knosys.2019.105124
4. Alqahtani, H., Sarker, I. H., Kalim, A., Minhaz Hossain, S. M., Ikhlaq, S., & Hossain, S. (2020). Cyber intrusion detection using machine learning classification techniques. *Communications in Computer and Information Science*, *1235 CCIS*, 121–131. https://doi.org/10.1007/978-981-15-6648-6_10
5. Atefinia, R., & Ahmadi, M. (2021). Network intrusion detection using multi-architectural modular deep neural network. *Journal of Supercomputing*, *77*(4), 3571–3593. https://doi.org/10.1007/s11227-020-03410-y
6. Azuaje, F. (2006). Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques 2nd edition. *BioMedical Engineering OnLine*, *5*(1), 1–2. https://doi.org/10.1186/1475-925x-5-51
7. Cleary, M. (2019). Deep Learning A Practioner's Approach. In *Journal of Chemical Information and Modeling* (Vol. 53, Issue 9).
8. Deepa, A. J., & Kavitha, V. (2012). A comprehensive survey on approaches to intrusion detection system. *Procedia Engineering*, *38*, 2063–2069. https://doi.org/10.1016/j.proeng.2012.06.248
9. Diro, A., & Chilamkurti, N. (2018). Leveraging LSTM Networks for Attack Detection in Fog-to-Things Communications. *IEEE Communications Magazine*, *56*(9), 124–130. https://doi.org/10.1109/MCOM.2018.1701270
10. García-Teodoro, P., Díaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers and Security*, *28*(1–2), 18–28. https://doi.org/10.1016/j.cose.2008.08.003
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144. https://doi.org/10.1145/3422622
12. Huang, S., & Lei, K. (2020). IGAN-IDS: An imbalanced generative adversarial network towards intrusion detection system in ad-hoc networks. *Ad Hoc Networks*, *105*. https://doi.org/10.1016/j.adhoc.2020.102177
13. Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.
14. Kwon, D., Kim, H., Kim, J., Suh, S. C., Kim, I., & Kim, K. J. (2019). A survey of deep learning-based network anomaly detection. *Cluster Computing*, *22*, 949–961. https://doi.org/10.1007/s10586-017-1117-8

15. Li, C., Xu, K., Zhu, J., & Zhang, B. (2017). Triple generative adversarial nets. *Advances in Neural Information Processing Systems*, *2017-December*, 4089–4099.
16. Liao, H. J., Richard Lin, C. H., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications*, *36*(1), 16–24. https://doi.org/10.1016/j.jnca.2012.09.004
17. Lin, S. Z., Shi, Y., & Xue, Z. (2018). Character-Level Intrusion Detection Based on Convolutional Neural Networks. *Proceedings of the International Joint Conference on Neural Networks*, *2018-July*, 1–8. https://doi.org/10.1109/IJCNN.2018.8488987
18. Lipton, Z. C., Berkowitz, J., & Elkan, C. (2015). *A Critical Review of Recurrent Neural Networks for Sequence Learning*. 1–38. http://arxiv.org/abs/1506.00019
19. Moraboena, S., Ketepalli, G., & Ragam, P. (2020). A deep learning approach to network intrusion detection using deep autoencoder. *Revue d'Intelligence Artificielle*, *34*(4), 457–463. https://doi.org/10.18280/ria.340410
20. Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2016). Deep learning approach for Network Intrusion Detection in Software Defined Networking. *Proceedings - 2016 International Conference on Wireless Networks and Mobile Communications, WINCOM 2016: Green Communications and Networking*, 258–263. https://doi.org/10.1109/WINCOM.2016.7777224
21. Viegas, E., Santin, A., Bessani, A., & Neves, N. (2019). BigFlow: Real-time and reliable anomaly-based intrusion detection for high-speed networks. *Future Generation Computer Systems*, *93*(Ml), 473–485. https://doi.org/10.1016/j.future.2018.09.051
22. Wang, X. Z., Buontempo, F. V., Young, A., & Osborn, D. (2006). Induction of decision trees using genetic programming for modelling ecotoxicity data: Adaptive discretization of real-valued endpoints. *SAR and QSAR in Environmental Research*, *17*(5), 451–471. https://doi.org/10.1080/10629360600933723
23. Wang, Y., Meng, W., Li, W., Li, J., Liu, W. X., & Xiang, Y. (2018). A fog-based privacy-preserving approach for distributed signature-based intrusion detection. *Journal of Parallel and Distributed Computing*, *122*, 26–35. https://doi.org/10.1016/j.jpdc.2018.07.013
24. Xu, C., Shen, J., Du, X., & Zhang, F. (2018). An Intrusion Detection System Using a Deep Neural Network with Gated Recurrent Units. *IEEE Access*, *6*, 48697–48707. https://doi.org/10.1109/ACCESS.2018.2867564
25. Yu, Y., & Bian, N. (2020). An Intrusion Detection Method Using Few-Shot Learning. *IEEE Access*, *8*(1), 49730–49740. https://doi.org/10.1109/ACCESS.2020.2980136

DDOS DETECTION IN SDN: A REVIEW

Mukesh Kumar, Abhinav Bhandari
Department of Computer Engineering, Punjabi University, Patiala
Mukesh.hcl.noida@gmail.com
bhandarinitj@gmail.com

ABSTRACT— SDN has changed the Network industry in the last decade because of its benefits like decoupling of control plane and data plane, programmability, customization, etc. Security is one domain where it needs to improve continuously for the better. DDoS can be a big problem related to the centralized control plane model as any successful DDoS attack can create lots of damage to SDN-based network by disrupting control plane availability. This paper reviews recent works in DDoS Detection for SDN-based networks. SDN Controllers are vulnerable to DDoS attacks like TCP SYN Flood and HTTP Flood. By having a fast DDoS threat response system, SDN can be deployed securely. This paper helps in understanding how different types of DDoS attacks can be detected and mitigated to deploy SDN securely.

KEYWORDS— SDN, DDoS, Controller, Attack, Security

I. INTRODUCTION

Software Defined Network or SDN is one of the most popular state-of-the-art technologies that have spread over different sectors of networking including Enterprises, Service Providers, Data Centers and many more. The reason being the wide-spread benefits offered by such eminent technology. Since old-fashioned technologies were getting obsolete, a student from Stanford University initiated the project under the name of "Clean Slate Project". The main intention was to re-design the network according to modern needs of the network which were not met through conventional ways. Therefore, a new type of network design has been proposed by disintegrating of Control and Data Planes.

The concept of SDN had been welcomed by almost all the major network providers, such as Cisco, Huawei, Juniper, Google, Microsoft and several others. The SDN uses the centralization technique to have a centralized controller and data plane devices that work on the instruction given by their controller. Because SDN has been advancing rapidly, it has been decategorized into several parts: SDWAN or Software Defined WAN, Software Designed Radio (SDR), Software Designed Access (SDA) and SDS or Software Defined Security.

OpenFlow protocol has been used to communicate between the controller and the data plane. The controller is the actual working brain of network. Unlike older network technologies, control plane is being developed by routing protocols which is eminent part of SDN framework. Apart from this, Control Clustering has been used to protect the network from problems encountered by controller, where primary control takes the charge. In DC, SDN controllers can be installed that will work on either Virtual Machine or on Hardware Devices like Servers having Linux distributions such as Red Hat, Ubuntu, Mint and so on. In the following figure, the architectural comparison has been discussed that demonstrate traditional as well as Software Defined Network designs.

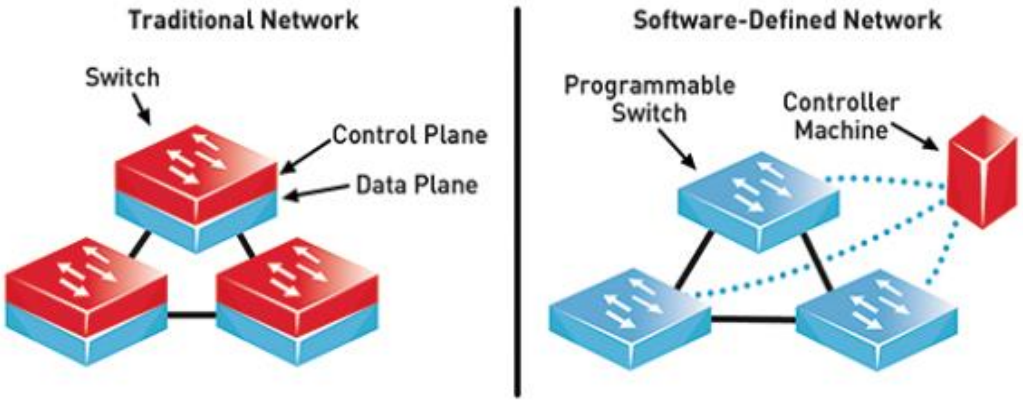


Figure 1 – Traditional Network v/s SDN

A. SDN ARCHITECTURE

SDN Architecture mainly consists of three different layers: Application, Control and Infrastructure Layer as shown in the figure 1.2

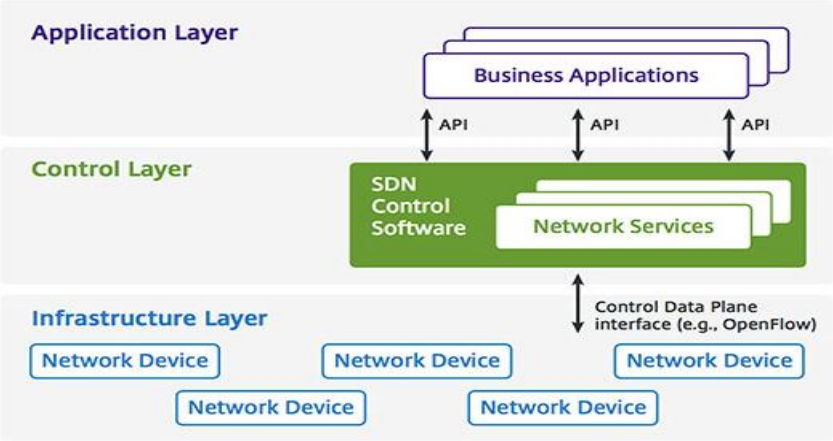


Figure 2 – SDN Architecture

1) *Application Layer*: It contains the programs which communicate to the controller using APIs or Application Programming Interfaces. Several applications can be used in a network, such as network management, analytics or large-scale applications for data centers.

2) *Control Layer*: This layer acts as the intermediatory layer of SDN architecture. It receives the information from the application later and uses it for network components in infrastructure layer. Other than this, it receives the information from the different network devices as well that passes the information to application layer further.

3) *Infrastructure Layer*: The infrastructure layer has the information regarding the data plane, which gets passed the data according to the information it gets from the controller.

B. Benefits of SDN Architecture

1) *Enhanced Configuration:* The configuration of the network is quite challenging as well as risky task because any mistake can devastate the whole network while leaving the vulnerability in the network. Traditionally, routers needed to get configured individually and that to after testing if there is change in the network devices, so that network may not get affected by the changes. In other words, manual configuration had been initiated on new devices in order to work smoothly. However, in SDN, controller takes care of such operations automatically that mitigates the configuration time.

2) *Customized & Programmable:* Programming of network is possible in SDN network through which new functionality and applications can be deployed in a network in an efficient manner. In earlier network technologies, propriety-based software can be used, for instance, in case of purchase of new router, base operating system can be used only that comes with that device. In case of additional features, licensing had been mandate, and there were no options for customizing or programmed the system. Besides this, experimenting was quite difficult. Nonetheless, using SDN framework, advanced features can be implemented through its programmable interface. Having this, more flexibility in the network can be achieved. Consequently, several data center organizations including Facebook and Google have SDN systems rather than having traditional systems.

3) *Low Cost Infrastructure:* SDN network is cost effective when compared to traditional network as there no issues of propriety-based hardware and software. Also, there is no need to have control plane in each device. Most of the companies use multiple controllers in large-sized organizations.

4) *Security of Granularity:* Having plenty of network devices connected to worldwide web and cloud services, there are escalation in challenges in terms of network scalability and security. However, this can be mitigated by controller that provides security because of its centralized nature while making the network policies easier and efficient.

5) *Better Visualization*: while having a centralized controller, the visualization and management of devices are far better. There is ease of finding the bugs and problems because there is no longer need to have monitoring system for each device just like in the older network technologies.

6) *Reliability and less Downtime:* The troubleshooting and implementation of SDN is easy in centralized system. Earlier, in case of problems and bugs, link had to be located in order to resolve the issue, and in large organization, it was quite daunting. However, there is ease of troubleshooting the network devices in case of SDN infrastructure. By using the SDN clustering feature, Active-Active Controller can be implemented, which make the network redundancy more flexible and robust. Besides this, load-balancing on controllers are easy and efficient way of utilization of resources.

C. Problems with SDN

1) Resolving On-demand up gradation: There are certain technologies including cloud computing, IoT, Machine Learning and so on which are creating lot of issues in the industry. Resultantly, there is mammoth data size which needs to be processed without having the delay or latency in the network.

2) Automation of devices: Since there is automation of network as well as server, the challenges associated with it are complex and problematic, especially in data centers. There are certain types of Application Programming Interfaces are being used by the application layer which needs to be handled carefully. OpenAPIs needs to be implemented for this to have the better visibility coupled with better understanding of controllers and topologies used in SDN architecture.

3) Security: Security is one of the primary challenges associated with SDN systems. In order to protect the controller from unauthorized access and other problems, such as DDOS attacks and malwares, security is paramount. Since SDN controllers can have open sources application, third-party software may hinder the functionality and cause harm to the system as the application that are being used in controller are not verified and tested. Therefore, these applications may create vulnerabilities in the SDN controller working. There are several types of attacks which are discussed ahead:

- Third party application software can be malicious and cause damage to the SDN controller because they are untrusted and cannot be verified.
- DDoS attacks are quite common these days initiated by hackers which are meant to halt the operations of network, application and service availability. Such attack disables the controller from performing its tasks. Larger-sized SYN, ICMP, TCP, UDP, HTTP request can be generated using various bots to disrupt the controller.
- Man-in-the-Middle attacks are also vogues that are meant for to breach the confidentiality. Using the OpenFlow protocols, Controller controls the data plane switches for the selection of paths. MITM attacks can be conducted by miscreant attackers to hinder the communication between these two and try to gain the access when the authentication is missing.
- Spoofing of the controller can also be performed in order to control the data plane switches using rogue controller.

II. LITERATURE REVIEW

Securing SDN Network is one of the major focuses of organizations deploying SDN or its sub-technologies like SD-WAN, SD-Access, SD-Security etc. The controller based model where control plane and data plane is decoupled left some security issues mainly related to DDoS. Recent works in the field of DDoS Detection and Mitigation is explained in a detailed review in table on next page:

TABLE I
Recent Researches done by the authors

Author	Year	Description
Cui, Y et al.	2021	Author provided a comprehensive survey in two kinds: (1) DDoS targeted at the SDN network and (2) DDoS aimed at the service providers. Five different categories of DDoS – machine learning based, statistics based, combination of multiple methods, threshold based and other methods. As per the conclusion, machine learning based methods are the most popular ones to detect DDoS attacks.
Li, R et al.	2020	Author proposed a lightweight early detection scheme for DDoS attacks in SDN on the basis of entropy. IP Address is extracted from the flow table and entropy value is calculated and then a comparison is made against the threshold value used to detect if a DDoS attack occurred.
S. Saharan et. al.	2019	Author states DDoS as the threat that disrupts the availability of the network. DoS attacks involves a single source and DDoS is performed by botnet. Different types of DDoS attacks are explained by the author with major focus on the DNS related DDoS attacks. These attacks can be prevented or the impact can be reduced by using application layer protocols. SDN adds the intelligence of programmability and customization to better tackle issues related with the DDoS.
S.Dong et.al	2019	Author stated SDN and Cloud Computing are widely adopted technologies in IT industry. Security is the major issue related with these technologies. Author reviewed various DDoS attack scenarios in SDN and Cloud along with the detection schemes which can be used to reduce the impact of DDoS in both the environments.

Abimbola Sangodoyin et. al.	2018	Author explained SDN as a progress in networks that make networks more interoperable than before along with innovative approaches. DDoS is one of the biggest security issue related with Networks and it is not changed in SDN and can be disastrous in case the controller controlling the network becomes unavailable taking whole network down. In his research work, author used Mininet Emulator, ODL controller and network testing and traffic generating tools like iperf and hping and hit the victim with three different DDoS attacks related with TCP, UDP and ICMP. Results of their experimentation displays that there is throughput fall from 233Mbps to 87Mbps during the attack.
B. H. Lawal and A. T. Nuray	2018	SDN is emerging rapidly in network industry due to heap of benefits it brings. Security is the only big challenge that it faces. Author used sFlow Analyzer to detect and control the DDoS attack and analyzes the traffic on the network and filters traffic on the basis of rules before processing it through the controller. All the experimentation is performed on Mininet Emulator inside a VM.
Muhammed Tahir et. al.	2018	Author has worked in order to find and explain current tools in Linux Availability Protection System also known as LAPS, which is mainly used to deny or filter DDoS traffic. Three type of attacks were performed i.e. TCP, UDP and ICMP on web server application. Author has used Linux based firewalls like IPTables and EPTables by adding the policies in them to filter the traffic.
Bawany et. al.	2017	DDoS is a big challenge to the security of SDN. Author reviewed some of the recent works related with DDoS detection and mitigation. Author categorized the techniques on the basis of detection solutions and then presented a framework which can be used to detect and mitigate DDoS attacks in SDN. ProDefense Framework is proposed by the author for proactive DDoS detection and mitigation in SDN to improve the security.
Xu Xiaoqiong et. al.	2017	Author described DDoS as one of the biggest and rising threat on Internet Security. Decoupling of data and control plane brings some options to solve the DDoS related problem in network industry. But there is one issue related with the centralization as a vulnerable controller can be used to take down whole network by pushing a successful DDoS attack. Author presented different types of DDoS and threats related with it in traditional and SDN.
Prajakta M. et. al.	2017	Author has proposed a DDoS detection solution with the help of Intrusion Detection System (IDS) in SDN. Author performed a DDoS attack on the target machine first and used IDS to detect the DDoS attack and isolates the traffic from different sources and states that SDN by itself is not secure at all and in order to make it secure, one has to perform best security practices.
Huseyin Polat & Onur Polat	2017	Author states that SDN has plethora of benefits that it brings to the network industry. Using SDN,all the network functions related with routing, switching, security and qos can be centralized as control plane is managed only by controller. Author performed bandwidth based DDoS attack on ODL and Pox and found that there is bandwidth reduction as the time passes and it results in rhigher response time. It is also found that as flow table disrupts because of high response time, it becomes difficult to add the flow table again as most time of the controller revolves around handling of error and packet-in events.

Zhanogang Shu. et. al	2016	Author reviewed issues and solutions related with Software Defined Networks. The major focus of author is on the security related problems at different layers of SDN architecture. Author also defined some of the countermeasures related with security problems.
Mohammed A. Saleh et. al.	2015	Author has proposed a framework to solve issues related with HTTP based DDoS attack in the network and named it Flexible, Collaborative, Multiplayer DDoS Prevention Framework or FCMDPF. As per author, the framework works well for the web applications that are mainly vulnerable for HTTP based DDoS attack. Although not as accurate as some of the previous works, author is working in future to solve the accuracy issue.
Nasir Shahzadet et. al.	2015	Author explains the tasks of network engineers in handling the network traffic in internet service providers and data centers. Video and Voice traffic is rising rapidly and SDN brings the multiple options related with programmability and customization that helps in ensuring better data flow than traditional networks. Author, however, stated some concerns on the centralization of the control plane as it can bring the whole network down, in case the controller becomes unavailable with DDoS attack. Malware and Spoofing are another two security concerns as controller can be spoofed and can be malware infected if any third party application is integrated in the application layer without checking it in any test case.
D Kreutz et. al.	2015	Traditional networks are complex and manageability takes lots of toll specially when there are some large deployments or troubleshoot a large network. As per author, SDN eliminates those issues and make a customizable network. There are eight fundamental facets described by the authors when traditional network was compared with SDN and i.e. hardware infrastructure, integration of controller and data plane device through southbound interface, network virtualization, various network OS, integration of applications and controller via northbound interface, different network slicing options, and network programming.
Nick Feamster et. al.	2015	Author presents the history of customizable and programmable network which is quite same as SDN, but using only programmable network does not contain a proper process or path. Using SDN, decoupling of control and data plane was the major focus and OpenFlow has worked perfectly in integration with the programmable networks. Authors also put emphasis on making understand that SDN is more of a technology that helps in solving network management problems.
Z. Anwar. et. al.	2014	Author in his work simulated a DDoS attack in data center networks. According to author, automation is partially used in data centers and large scale data centers or Cloud companies like Amazon, Google or Microsoft etc. have a staff of around 50-60 in the data center and major chunk of work is through the automation tools and scripts. Author stated that there is a possibility that DDoS attack can be used on control management apps that brings cooling system disruption.
Gagandeep G et. al.	2014	SDN is the future of the network industry and brings lots of benefits like customization because of the programmability, easy configuration, faster troubleshooting and centralization etc. Although mainly used for data center industry, SDN has evolved widely in other network industries like ISPs and

		Enterprise Networks with the technologies like SD-WAN, SD-Access, SD-QoS etc. In the review, author also stated that openness in the SDN helps writing programs for controllers and limits the need of licenses for different features.
Diego Kreutz et. al.	2013	The major focus of the work is to design and deploy SDN networks with security in mind. Author explored different threats in SDN based networks and some of the best security measures which can be taken to ensure secure SDN deployment.

III. CONCLUSION

SDN is rapidly changing the deployment methods of world networks, industry is accepting the change with open hands because of the benefits which are provided by the SDN, but still there are security challenges related with SDN as SDN by default is not secure and we need to configure best security practices in order to make it secure. DDoS is one of the biggest vulnerability that Controller has because of the centralized design it has. Due to centralization, SDN brings lots of benefits, but with benefits it also has a drawback that a successful DDoS attempt on controller can disrupt the control plane availability. IDS can be used to detect the DDoS attack and generate the traffic alert in case any critical service/port is communicated. Recent works have not used Intrusion Detection Systems for DDoS detection on SDN controllers. Snort is an open-source IDS and also has a proprietary version comes under Cisco. Using Snort can integrate a proven alert and detection system with SDN.

REFERENCES

[1] AbimbolaSangodoyin, TshiamoSigwele, Prashant Pillai, Yim Fun Hu, IrfanAwan and Jules Disso: DoS Attack Impact Assessment on Software Defined Networks. ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2018.

[2] B. H. Lawal and A. T. Nuray: Real-time detection and mitigation of distributed denial of service (DDoS) attacks in software defined networking (SDN) In :*26th Signal Processing and Communications Applications Conference (SIU)*,Izmir, pp. 1-4, 2018.

[3] Muhammad Tahir,Mingchu Li, NaeemAyoub, Usman Shehzaib and AtifWagan: A Novel DDoS Floods Detection and Testing Approaches for Network Traffic based on Linux Technique. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 9, No. 2, pp 341-357, 2018.

[4] NarmeenZakariaBawany, Jawwad A. Shamsi1 & Khaled Salah: DDoS Attack Detection and Mitigation Using SDN: Methods,Practices, and Solutions. Arab J SciEng(Springer) ,Vol. 42, pp 425–441,2017.

[5] Prajakta M. Ombase, Nayana P. Kulkarni, Sudhir T. Bagade and Amrapaliv V. Mhaisgawali (2017): Survey on DoS Attack Challenges in Software Defined Networking. International Journal of Computer Applications (0975 – 8887), Vol. 173, No.2, September 2017.

[6] Huseyin POLAT, Onur POLAT: The Effects of DoS Attacks on ODL and POX SDN Controllers In: 8th International Conference on Information Technology (ICIT),2017.

[7] ZhaogangShu, Jiafu Wan, Di Li, Jiaxiang Lin, Athanasios V. Vasilakos, and Muhammad Imran: Security in Software-Defined Networking: Threats and Countermeasures In: *Mob. Netw. Appl.* Vol. 21, No. 5, pp 764-776, Oct 2016.

[8] Mohammed A. Saleh1 and AzizahAbdulManaf : A Novel Protective Framework for Defeating HTTP-Based Denial of Service and Distributed Denial of Service Attacks Web Links.Hindawi Publishing Corporation, The Scientific World Journa,Vol. 2015, pp 1-19, 2015.

[9] Saleh, M.A. and A. Abdul Manaf: A novel protective framework for defeating http-based denial of service and distributed denial of service attacks. The Scientific World Journal, Vol. 2015, pp 1-19, 2015.

[10] Nasir Shahzad, GhulamMujtaba and ManzoorElahi: Benefits, Security and Issues in Software Defined Networking (SDN). NUST Journal of Engineering Sciences, Vol. 8, No. 1, pp. 38-43, 2015.

[11] Gagandeep Gargand RoopaliGarg: Review On Architecture & Security Issues of SDN. International Journal of Innovative Research in Computer and Communication Engineering,Vol. 2, No. 11, November 2014, ISO 3297: 2007.

[12] Diego Kreutz, Fernando M. V. Ramos and Paulo Verissimo: Towards Secure and Dependable Software-Defined Networks: HotSDN'13, ACM, 2013.

[13] D. Kreutz, F. Ramos, P. Verissimo, C. Rothenberg, S. Azodolmolky, and S. Uhlig: Software-Defined Networking: A Comprehensive Survey In: Proceedings of the IEEE, Vol. 103, No. 1, pp. 14-76, January 2015.

[14] Alexander Gelberger, Niv Yemini, RanGiladi: Performance Analysis of Software-Defined Networking (SDN): IEEE, 2013.

[15] XU Xiaoqiong, YU Hongfang, and YANG Kun: DDoS Attack in Software Defined Networks: A Survey: ZTE COMMUNICATIONS ,2017

[16] Nick Feamster, Jennifer Rexford, Ellen Zegura: The Road to SDN: An Intellectual History of Programmable Networks: Princeton, USA,2015.

[17] Scott Shenker, Martin Casado, TeemuKoponen, Nick McKeown,: The future of networking, and the past of protocols: Open Networking Summit, Vol. 20, pp 1-30, 2011.
[18] Xerxes DOS Tool, https://github.com/zanyarjamal/xerxes [Accessed on 20 June 2019].
[19] PyLoris DOS Tool, https://sourceforge.net/projects/pyloris/ [Accessed on 20 June 2019].
[20] Hping3, https://tools.kali.org/information-gathering/hping3 [Accessed on 20 June 2019].
[21] OpenDaylight project, https://www.opendaylight.org[Accessed on 20 June 2019].
[22] GitHub of OpenDaylight Integration Project, https://github.com/opendaylight/integration. [Accessed on 20 June 2019].
[23] Best DOS Attacks and Free DOS Attacking Tools, https://resources.infosecinstitute.com/dos-attacks-free-dos-attacking-tools/#gref[Accessed on 20 June 2019].
[24] SDN Explained Article https://commsbusiness.co.uk/features/software-defined-networking-sdn-explained/[Accessed on 20 June 2019].
[25] MaxTech,"SDN Architecture", http://learning.maxtech4u.com/software-defined-networking/S. Dong, K. Abbas and R. Jain, "A Survey on Distributed Denial of Service (DDoS) Attacks in SDN and Cloud Computing Environments," in *IEEE Access*, vol. 7, pp. 80813-80828, 2019. doi: 10.1109/ACCESS.2019.2922196
[26] M. Ficco, F. Palmieri, "Introducing fraudulent energy consumption in cloud infrastructures: A new generation of denial-of-service attacks", *IEEE Syst. J.*, vol. 11, pp. 460-470, Jun. 2017.
[27] Z. Anwar, A. W. Malik, "Can a DDoS attack meltdown my data center? A simulation study and defense strategies", *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1175-1178, Jul. 2014.
[28] S. Saharan and V. Gupta, "Prevention and Mitigation of DNS based DDoS attacks in SDN Environment," *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, Bengaluru, India, 2019, pp. 571-573. doi: 10.1109/COMSNETS.2019.8711258
[29] Y. Liu, B. Zhao, P. Zhao, P. Fan and H. Liu, "A survey: Typical security issues of software-defined networking," in *China Communications*, vol. 16, no. 7, pp. 13-31, July 2019. doi: 10.23919/JCC.2019.07.002
[30] S. Dong, K. Abbas and R. Jain, "A Survey on Distributed Denial of Service (DDoS) Attacks in SDN and Cloud Computing Environments," in *IEEE Access*, vol. 7, pp. 80813-80828, 2019. doi: 10.1109/ACCESS.2019.2922196
[31] M. Ficco, F. Palmieri, "Introducing fraudulent energy consumption in cloud infrastructures: A new generation of denial-of-service attacks", *IEEE Syst. J.*, vol. 11, pp. 460-470, Jun. 2017.
[32] T.Lohman, DDoS is Cloud's Security Achilles Heel, Sep. 2011, [online] Available: http://www.computerworld.com.au/article/401127.
[33] D. Sher, "Gartner: Application layer ddos attacks to increase in 2013", 2013.
[34] Z. Anwar, A. W. Malik, "Can a DDoS attack meltdown my data center? A simulation study and defense strategies", *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1175-1178, Jul. 2014.
[35] D. Anstee, D. Bussiere, G. Sockrider, "Arbor special report: Worldwide infrastructure security report", 2012, [online] Available: http://pages.arbornetworks.com/rs/arbor/images/wisr2012_en.pdf.
[36] Understanding the SDN architecture. SDXCentral.
[37] Sdn architecture. Open Networking Foundation, 2014.
[38] Vahid Ahmadi, Ahmad Jalili, and Mostafa Khorramizadeh. Multi-objective controller placement problem: issues and solution by heuristics. *International Journal of Computer Science and Information Security*, 14(8):543, 2016.
[39] M. Isard, "Autopilot: Automatic data center management," *Operating Syst. Review*, vol. 41, no. 2, pp. 60–67, 2007.
[40] R. Miller. (2008, Jan. 18). The economics of data center staffing. [Online]. Available: www.datacenterknowledge.com
[41] https://noviflow.com/wp-content/uploads/Image-1.jpg
[42] Tan, Liang, Yue Pan, Jing Wu, Jianguo Zhou, Hao Jiang, and Yuchuan Deng. "A new framework for DDoS attack detection and defense in SDN environment." *IEEE Access* 8 (2020): 161908-161919.
[43] Li, R., & Wu, B. (2020, June). Early detection of DDoS based on $\varphi $-entropy in SDN networks. In *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)* (Vol. 1, pp. 731-735). IEEE.
[44] Cui, Y., Qian, Q., Guo, C., Shen, G., Tian, Y., Xing, H., & Yan, L. (2021). Towards DDoS detection mechanisms in Software-Defined Networking. *Journal of Network and Computer Applications*, 103156.

A REVIEW ON ZERO TRUST NETWORK

Mukul, Madan Lal
Department of Computer Engineering
Punjabi University Patiala, 147002
Patiala, Punjab, India.
Email Id: - 12091009@csepup.ac.in, madanlal@pbi.ac.in

ABSTRACT— Zero trust Network is a new concept that involves the provision of organization resources, to the subjects without relying on any implied trust. It is a security model that helps the organizations to restrict unauthorized access into the network. The zero-trust model grant access only to authorized nodes, making this a very safe and secure model, unlike the legacy technologies in which any user or node inside the network is considered trusted. In this work, we have reviewed latest research papers related to zero trust network and their security challenges are summarized.

KEYWORDS— *Zero Trust Network, Zero Trust Model, No trust, zero knowledge, zero control, Mobile, Perimeter*

1. INTRODUCTION

In today's world most of the data is present online, whether it is in the form of audio, video, text or any other form. Every bit of a data that is available online could be insecure. So, securing data must be a crucial part for any organization. Back in times we have trusted our traditional security system, but as the technology is growing day by day, we can't be fully sure about the security of our data. So, to solve this problem, zero trust network has been introduced. This model follows a whole different kind of approach that we will discuss in the paper.

1.1 Zero Trust Network

In 2010, John Kindervag, has proposed a term called as zero trust. Zero Trust or Zero Trust Network is a security model that works on "Never trust, always verify" approach. This model provides a secure environment inside the network for all the nodes that are connected to each other. Zero trust network checks that if all the nodes that are connected inside the network is trusted or not. This model works on "least privilege access", so therefore it only grants access tothose nodes that are trusted and can lead to a secure and seamless process between the sender and the receiver. It combines strict security rules of verification for every device.

Zero Trust Network isn't based on a single technology. It combines multiple security technologies to work without any interruptions. Zero Trust Network assumes that the security of the network is compromised and to avoid this situation zero trust network take a strong approach by protecting the inside network and building a strong authentication and authorization approach to gain the access. "Zero trust is viewed as based on no presumptive trust, and a risk-based approach to trust, along with verification of trust on a continuous basis" [1].

The main goal of zero trust network is to gain control access based on identity and the behavior of the nodes. "Zero Trust network benefits to prevent data breaches caused due to exploitation of privileged credentials by stamping out the concept of trust from an organization's network architecture" [2].This short paper proceeds by first reviewing prominent thinking in zero trust and briefly discussing zero trust proof system [1].

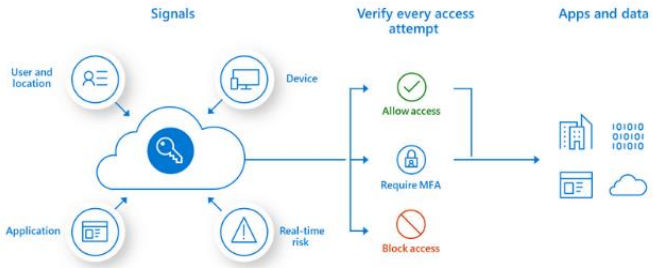


Figure 1: Zero Trust Network Model [9].

The zero-trust security model takes verification multiple steps further. The model incorporates strict security protocols, with multi-factor authentication as a minimum, as well as inspecting and logging all traffic. Access requests originating on a local area network (LAN) are treated with the same level of suspicion as if they had come from a wide area network (WAN) [13].

1.2 Zero Trust Architecture

Zero Trust Architecture works on the same principle as the Zero Trust Network of "Never trust, always verify". "The core idea of Zero Trust Architecture is that no person/device/system inside or outside the network should be trusted by default, and the trust foundation of access control needs to be reconstructed based on authentication and authorization" [3].

The zero-trust architecture reduces malicious access and attacks by employing least privilege policies and strictly enforcing access control policies. It detects and logs all network traffic and continuously tracks the user behavior [4].

The zero-trust security framework mainly includes the following ideas.

- Using identity as the basis of access control.
- Using 'least privilege' principle for resource allocation.
- Real-time calculation of access control strategy.
- Continuous evaluation of trust level from multiple data sources.

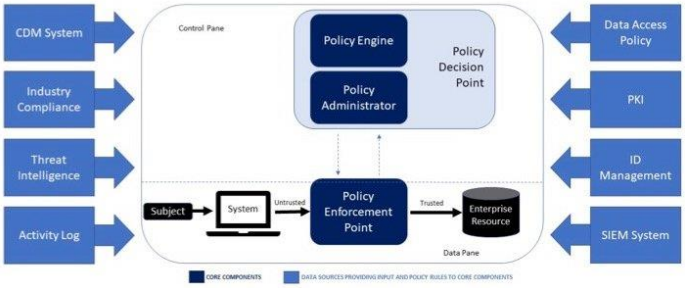


Figure 2: Zero Trust Architecture [10].

There are 3 main components of zero-trust network architecture:

1. Policy Engine (PE): The Policy Engine handles the main decision to grant, deny or to revoke the access.
2. Policy Administration (PA): The main aim of the policy administration is to establish, maintain and terminate the sessions in the data plane.
3. Policy Enforcement Point (PEP): The policy enforcement point is responsible for handling the enable operation, monitoring and terminating the connections.

Zero Trust Architecture also contains three steps of security:

1. Verify the user
2. Verify the device
3. Verify access privileges

These three layers of security are accomplished through multiple compliance checks based on the attributes of each device. These compliance checks range from device encryption to pattern behavior. In case if any device fails to complete any compliance check, then it will label the user or device as "non-compliant" and deny access to that user or device [8].

1. The user initiates an access request through the control plane and then it gets authenticated and authorized by the trust evaluation engine.
2. Once the request is granted, the system firmly configures the data plane., the access agent initiate the data traffic from the access object and establish a secure connection between sender and the receiver.
3. The trust evaluation engine supplies the data to the access control engine for zero trust.

Zero trust is intended to provide a scalable security infrastructure which can be applied across many different types of organizations. A fundamental principle of zero trust involves guaranteeing secure access to all resources, regardless of location, and assuming all network traffic is a threat until it is authorized, inspected, and secured [11].

The National Cyber Security Centre (NCSC) published some guidelines of zero trust in 2019. According to NCSC the nodes that connects to the traditional security framework has a possibility of getting hacked. In Zero Trust Model, it assumes that the security of the network is always compromised, hence no one is trusted. The Zero Trust Technology never grants an access to the resources until they are verified by reliable authentication and authorization.

The National Cyber Security Centre (NCSC) proposes 10 principles for zero trust: [1] knowledge of architecture, devices and services, [2] the creation of single strong identity, [3] a strong device identity, [4] authenticate everywhere, [5] knowledge of the health of devices and services, [6] monitoring devices and services, [7] policies based on the value of the devices and services, [8] control access to device and services, [9] don't trust the network including the local network, and [10] choose services designed for zero trust.

Zero-trust network security has already been used and implemented by few companies. The whitelist-based policy model supports zero-trust security architecture. These policies permit, deny and logs the traffic between two endpoints.Zero-trust is intended to provide a scalable security infrastructure which can be applied across many different types of organizations [11].The zero-trust model provides security to various sectors like: cloud environment, on-premises data centers, users, devices etc.

The zero-trust concept has been intended to create a new access control policy that embraces the modern environment and protects individual devices and users beyond their perimeter, which is free from network support micro-segmentation. In the present cloud environment, the security policy proposed using the zero-trust framework would grant access to significant network traffic for the use of distributed cloud resources [14].

Zero Trust is characterized by segmented, parallelized, and centralized network. It is based on three key concepts that empower secure networking:

1. Easy to manage segmented networks: Hierarchical networks are difficult to segment. Zero Trust recommends new ways of segmenting networks by addressing segmentation at the core of the network.
2. Built with multiple parallel switching cores: Zero Trust recommends distributed processing by breaking the core switch into multiple smaller and less expensive cores. By using the concept of parallelization, Zero Trust segregates network traffic into smaller network segments.
3. Centrally managed from a single console: Central management of all networking elements is a key feature of Zero Trust. It recommends a platform to centrally manage the network components and segment network traffic [15].

In today's time, everything work is being done with the help of cloud as all the data is now migrating to cloud for better convenience of the user. By migrating the data on cloud, there will be many advantages of the same. [1] 24*7 availability of the data, [2] Less resources needed, [3] Security of data is more, [4] User can work mobile, [5] Less storage required and so on.

Cloud applications have redefined the security perimeters. Employees are bringing their own devices and working remotely. Data is being accessed outside the corporate network and shared with external collaborators such as partners and vendors. Corporate applications and data moving are moving from on-premises to hybrid and cloud environments [7].

The core idea to adopt zero-trust network model is for all the people, devices and applications and to adopt the principle of minimum authority, making network security the most important criterion [12].

2. LITERATURE REVIEW

In this section, latest papers related to zero-trust network are presented in the form of table on the basis of their contribution and year of publication.

Sr No.	Topic	Publisher	Year	Contribution
1.	Establishing a Zero Trust Strategy in Cloud Computing Environment	IEEE	2020	A zero-trust model for security is proposed in order to address modern security challenges that come up with cloud infrastructure. The security is the major challenge in deploying new infrastructure into the cloud. According to the author, zero-trust is proposed as the best-fit for cloud deployment. To Advance cloud security, the Zero-Trust strategy creates a record of what it has in the cloud and accordingly implements strong access control. Zero-Trust framework can better track and block external attackers, while limiting security breaches resulting from insider attacks in cloud paradigm. Zero-Trust can better accomplish access privileges for users and devices across cloud to enable secure sharing.
2.	Zero trust: Never trust, always verify	IEEE	2021	Few Trust algorithms are applied in this paper to evaluate, enforce and to calculate levels of confidence in subject based on contextual elements set as policy by enterprise (time, geo-location, device health, behavior). Zero-trust ensures to protect sensitive information, yet many systems depend on sharing identity and passwords so, an eavesdropper can interfere with the system. To avoid this situation, zero-knowledge proof and grabled circuits are introduced. The zero-knowledge proof approaches are found on the principle of correctness, soundness and zero knowledge. This protocol prevents the unintended disclosure of information during the sharing, and potential problems of compromise where information and privacy may leak. Another way to solve the problem is through grabled circuit. This process involves holding secret inputs which are input into a mathematical computation. This procedure is founded on the properties of validity, privacy and correctness.
3.	A Security Awareness and Protection System for 5GSmart Healthcare Basedon Zero-Trust Architecture	IEEE	2020	Four major dimensions of zero-trust framework for 5G smart healthcare scenarios are proposed: [1] Subject- It is a party that initiates network and resource access request. [2] Object- It includes medical data, smart medical service functions, and service interface. [3] Environment- It includes physical, computing and network environment where medical and network equipment accesses. [4] Behavior- It includes real time security analysis and judgement based on historical access behavior and resource access behavior. The proposed security system is implemented and tested, which proves that it satisfies the needs of end-to-end security

				enforcement of data, users, and services involved in a 5G-based smart medical system.
4.	A small LAN zero trust network model based on Elastic Stack	IEEE	2020	This paper construct a small zero-trust LAN network model based on elastic stack. This model collects and audits the operation records if the computer by running transparent background processes on the computer and sends the information to elastic stack and uses "Kibana" application for real-time log analysis, so as to realize the real time detection or violation to zero-trust LAN network. Kibana isa visualization dashboard software tool for elastic search. It can log data in the form of graphs and charts. The purpose of this model is to try and solve the conflict between file transmission and security of important confidential units. This model ensures the security of internal network and it also solve the problem of using LAN for confidential units.
5.	Performance Analysis of Zero-Trust multi-cloud	IEEE	2021	An architectural test-bed is proposed to perform the analysis. The architectural setup involves two different cloud providers i.e. Google Kubernetes Engine (GKE) and Elastic Kubernetes Services (EKS). In this architecture, a DNS load-balancer is used to route the traffic between two Kubernetes clusters. With this, performance can be benchmarked against two different configurations. This method involves the analysis of CPU and memory system resource focusing on utilisation to identify what percentage of resource is required. It also manages the security of the system through HTTPS and URL requests.
6.	Intelligent Zero Trust Architecture for 5G/6G Tactical Networks	IEEE	2021	An intelligent monitoring, evaluation and decision making (MED), using artificial intelligence (AI) is proposed. The core of ZTA comprises of Policy Enforcement Point (PEP), and Policy Decision Point (PDP). The PEP is the first point of contact for access request. The decision on granting access is made be PDP. The information used by ZTA core to grant and monitor a connection is provided by peripheral modules known as static and dynamic. The static modules include data access policies, PKI and Identity management. It defines the security policy rules for secure communication. The dynamic module includes continuous diagnostics and mitigation (CDM), threat detection, activity logs and security information and event management (SIEM) for collecting information on security state and potential attacks.
7.	Research on the security protection framework of power mobile internet services based on zero trust	IEEE	2021	A zero-trust based mobile power mobile interconnection, business security protection framework is proposed which includes: trusted terminal environment awareness, user equipment application trusted recognition, terminal access agent and access control engine. Its basic components are: [1] Trusted terminal environment awareness, [2] trusted identification, [3] dynamic trust evaluation engine, [4] terminal access agent, and [5] access control engine. This research and analysis conclude about the current security protection risk, combined with the zero-trust concept and architecture. This project designs a zero-trust based power mobile interconnection business security protection framework, and builds an identity-centric for mobile interconnection logic access.
8.	Design and Implementation of a Consensus Algorithm to build Zero Trust Model.	IEEE	2020	A consensus algorithm is proposed in distributed systems so that the system remains consistent and allow concurrency of data. Consensus algorithms work in a manner where a majority of the nodes have to agree before any kind of transaction occurs. The proposed consensus algorithm builds a Zero Trust model as no central authority can be trusted to govern the distributed system. The consensus algorithm is a hybrid algorithm which combines Proof of Work and Proof of Elapsed Time. Data Security is ensured using the RSA algorithm for authentication and authorization. The Proposed consensus algorithm is used in building a zero-trust model which is implemented in the placement system.

9.	Cloud-Based Zero-Trust Access Control Policy:	Springer	2021	A zero-trust based access control policy is proposed to prevent MAC spoofing by ensuring security to the hosts and cloud services. This approach eliminates the threats before spoofing occurred. This approach operates on the open system interface (OSI) layer 3 and layer 4 where an individual TCP packet is captured from the incoming untrusted IP address and retrieves the IP address, port number, and corresponding MAC address of the respective traffic. The proposed algorithm uses the IP trackback and port scanning techniques validation of the TCP packet, which reduces the computational overhead. The use of dynamic threshold stamping by the proposed approach rectifies a legitimate user's traffic before classifying it to the attacker, which reduces the rate of false-positive rate significantly.
10.	The zero-trust supply chain: Managing supply chain risk in the absence of trust	Taylor & Francis	2021	A number of research propositions is proposed to help advance a research agenda in this new area of inquiry. The supply chain becomes more digitised, but it is not only the digital and information flows that need to be controlled, but products, materials, processes, and practices throughout the supply chain that can cause an organisation's reputation. The proposed model practices and philosophy can aid in making sure that the supply chain is secure and resilient to malicious activity.

3. CHALLENGES TO IMPLEMENT ZERO-TRUST

1. **Heavy investment required to migrate onto other security technologies:** It's a very difficult task for an organization to migrate their entire network onto some other security technology, because as an organization, they already invested a large amount on their existing security technology. Now its next to impossible for a low budget organization to relocate their entire network onto some other technology.
2. **Absence to identity governance:** Identity governance can be defined as a set of rules and regulations that every organization must establish and govern in order to keep a track on the access privileges for its employees. Lack of security guidelines and identity governance within an organization can result in unauthorized access breach and can fail a security framework.
3. **Lack to resources to develop a transition plan:** It can be defined as the lack of ability and resources of an organization that makes them think twice before developing a transition plan for transferring their whole network onto some other technologies.
4. **The Retrofit Effect:** Typically, older legacy technology is not compatible with zero trust model. A zero-trust approach requires the ability to control access at granular level and allow "on-the-go" verification. Many old legacies don't support that kind of security.
5. **Integration and Third Party:** When public and private cloud services work together in a unified manner to deliver a service which is not common, then it also breaks the ZTNA model.
6. **Remote Work:** Remote work has been on rise since 2020 pandemic occurs. It forced many employees to work from home with an unsecured internet connect resulting in various types of attacks. Adding zero trust, with its micro-segmentation requirements and decision making that creates another layer of complexity.

4. CONCLUSION

Zero Trust Network is a new concept in the technology market that helps the organization to avoid attacks. It works based on "never trust, always verify". It is useful while working "on-the-go". Zero trust provides the user an access to the resource which the user needed regardless of their physical and network location. Zero trust continually monitors the activity that's going on inside the network as well so that no malicious activity will be there. Zero trust network is very essential from security point of view. So, zero trust is way better than the legacy technology where the possibility and occurrence of an attack is high which is not good for an organization. But no security technology is 100% secure. In this paper, we have discussed about the concept and applications of zero-trust architecture and some of the challenges are reviewed that a user or group of users will have to face while implementing a zero trust in any organization.

REFERENCES

[1] A. Wylde, "Zero trust : Never trust , always verify," *IEEE*, pp. 1–4, 2021.

[2] S. Mehraj, "Establishing a Zero Trust Strategy in Cloud Computing Environment," *IEEE*, pp. 1–6, 2020.

[3] L. Chen, Z. Dai, M. Chen, and N. Li, "Research on the Security Protection Framework of Power Mobile Internet Services Based on Zero Trust," *Proc. - 2021 6th Int. Conf. Smart Grid Electr. Autom. ICSGEA 2021*, pp. 1–4, 2021, doi: 10.1109/ICSGEA53208.2021.00021.

[4] B. Chen *et al.*, "A Security Awareness and Protection System for 5G Smart Healthcare Based on Zero-Trust Architecture," *IEEE*, pp. 1–16, 2020, doi: 10.1109/JIOT.2020.3041042.

[5] M. Campbell, "Trust : Trust Is a Vulnerability," *IEEE*, pp. 1–4, 2020, doi: 10.1109/MC.2020.3011081.

[6] S. Rodigari, D. O'Shea, P. McCarthy, M. McCarry, and S. McSweeney, "Performance Analysis of Zero-Trust multi-cloud," pp. 5–7, 2021, [Online]. Available: http://arxiv.org/abs/2105.02334.

[7] Microsoft, "Zero Trust Maturity Model," *Microsoft Secur.*, pp. 1–7, 2020, [Online]. Available: https://go.microsoft.com/fwlink/p/?linkid=2109181.

[8] K. Delbene, M. Medin, and R. Murray, "The Road to Zero Trust (Security)", pp. 1–10, 2019.

[9] Admin Globaldots, "Zero trust network," 2020. https://www.globaldots.com/resources/blog/zero-trust-explained/ (accessed Oct. 06, 2021).

[10] N. I. of S. and Technology, "[NIST SP 800-207] Zero Trust Architecture," *Nist*, p. 49, 2020, [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207-draft2.pdf.

[11] Decusatis, C., Liengtiraphan, P., Sager, A., & Pinelli, M. (2016). Implementing Zero Trust Cloud Networks with Transport Access Control and First Packet Authentication. *Proceedings - 2016 IEEE International Conference on Smart Cloud, SmartCloud 2016*, 5–10. https://doi.org/10.1109/SmartCloud.2016.22

[12] Kong, C., Liu, J., Xian, M., & Wang, H. (2020). A small LAN Zero trust network model based on elastic stack. *Proceedings - 2020 5th International Conference on Mechanical, Control and Computer Engineering, ICMCCE 2020*, 1075–1078. https://doi.org/10.1109/ICMCCE51767.2020.00236

[13] Greenwood, D. (2021). Applying the principles of zero-trust architecture to protect sensitive and critical data. *Network Security*, *2021*(6), 7–9. https://doi.org/10.1016/S1353-4858(21)00063-5

[14] Mandal, S., Khan, D. A., & Jain, S. (2021). Cloud-Based Zero Trust Access Control Policy: An Approach to Support Work-From-Home Driven by COVID-19 Pandemic. *New Generation Computing*, *0123456789*. https://doi.org/10.1007/s00354-021-00130-6

[15] Dhar, S., & Bose, I. (2021). Securing IoT Devices Using Zero Trust and Blockchain. *Journal of Organizational Computing and Electronic Commerce*, *31*(1), 18–34. https://doi.org/10.1080/10919392.2020.1831870

COMPARATIVE ANALYSIS OF VARIOUS MANET ROUTING PROTOCOLS UNDER DDOS ATTACK: A SYSTEMATIC REVIEW

Isha Sharma[#1], Raman Maini[#2]
Department of Computer Science and Engineering, Punjabi University, Patiala., Department of Computer Science and Engineering, Punjabi University, Patiala.
[1]ishasharma69685@gmail.com
[2] research.raman@gmail.com

ABSTRACT— Mobile ad-hoc network (MANET) is a multi-hop wireless network which consists of various self-coordinated, mobile and wireless topology nodes. There have been several studies and extensions of ad hoc routing protocols focusing on different aspects, such as security, quality of service. MANETs are open transmission and communication media without any security mechanism. So, there are a lot of security attacks especially denial of service attack and distributed denial of service (DDoS) attacks on MANETs. So it is important to analyze the viability of these routing protocols when they are combined with a DoS attack. A review of different research works based on different parameters is presented with a complete analysis and comparison.

KEYWORDS— Routing Protocols, Mobile Ad-Hoc Networks, DDoS, Malicious Node, Network Performance

INTRODUCTION

Wireless ad hoc networks have become popular in recent years for research purposes. Two types of mobile networks are there -Infrastructure and Infrastructure-less mobile ad hoc networks. A MANET is a decentralized form of ad hoc network. Ad hoc networks (MANETs) are independent networks of mobile routers connected by wireless link connectivity. Each router organizes itself in a random manner and can move freely. The network topologies can change spontaneously and rapidly. In MANET, each host act as a router and Data is transferred among hosts using a multi-hop approach.

For ad hoc networks, numerous routing protocols have been introduced. Every routing protocol searches for a path using a different algorithm. In addition to their changing nature, routing protocols are also prone to a variety of attacks, including blackholes, wormholes, packet replication, DOS attacks, floods of unauthorized sessions, spam etc. Because, nodes can join, leave, move, and join the network at any time meaning it is susceptible to attacks coming from within or outside the network. It is known that in the literature various attacks have been used to evaluate the performance of multiple routing protocols under various scenarios, such as mobility, multiple attacker nodes, and varying nodes speeds in order to evaluate their effectiveness.

MANET

A mobile Ad-hoc network (MANET) is network of mobile nodes that communicate over wireless links and agrees to exchange route message between themselves[1]. Every node in MANET contains a wireless ad-hoc routing protocol and are capable to transmit data between the nodes[2] which means that MANET is an infrastructure-less composition of nodes that can switch positions(geographical locations) at their will. When nodes required sharing information with each other, they have to rely on other nodes and have to be in each other's range for communication. Ad-hoc networks depend on the trust and co-operation between nodes which makes it prone to attacks [3]. The important applications of mobile ad-hoc networks are WSN, robot networks, underwater network, internet of things (IoTs) and so on [4]. In MANET routing is generally classified into two types reactive and proactive protocols.

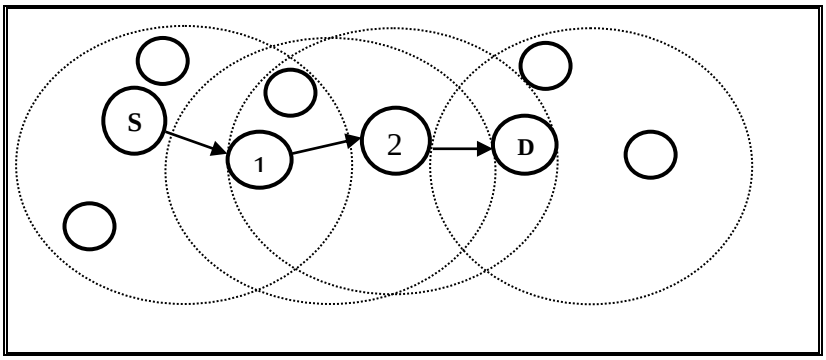


Fig.1 Mobile ad-hoc network (MANET)[1]

MANET BASED ROUTING PROTOCOLS

Routing is when every node has to find the paths to transmit data packets between computing systems in the network. To permit communication within the MANET and to establish routes among participating nodes, a routing protocol is needed.

These routing protocols use links information that is present in the network to perform the task of packet forwarding. Topology-based routing can be categorized into reactive, proactive, and hybrid [5].

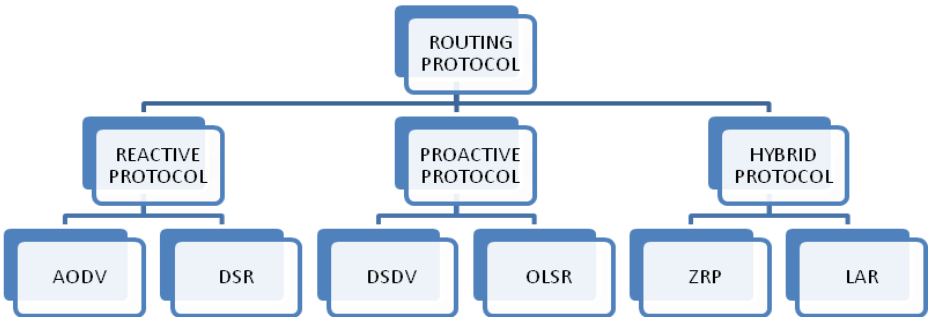


Fig.2 Classification of MANET routing protocols[14]

*A. **REACTIVE ROUTING PROTOCOL***

Reactive protocols, other name for them also includes on-demand-driven reactive protocols. They initiate the route only when it is necessary for a node to communicate with one another [5]. Route discovery and route maintenance are the two main functionalities of the protocol. Discovering a new route is the responsibility of the route discovery function and it occurs by flooding the route request packets. After this phase, the optimal path for transmitting data packets between the source and destination nodes will be identified. The route maintenance function is responsible for the link breakage and repair of the existing links. A variety of on demand-driven protocols have been developed including: Ad-hoc On-Demand Distance Vector (AODV), temporally ordered routing algorithm (TORA), Dynamic Source routing protocol (DSR)[6].

1). *AODV* *(ad-hoc on demand distance vector)* it provides loop-free routing for ad-hoc networks. It is a hop to hop, unidirectional, widely used on-demand routing protocol. it limits the broadcasting process by permitting the on-demand routes. In AODV route establishment occurs in two phases that are route discovery and route maintenance[14]. Each node of AODV maintains a routing table, which contains three critical fields: the node that is the next hop, the sequence number, and the number of hops. In AODV, the route discovery process begins when the source node starts and floods the network with RREQs (Route Request). The consecutive node to the source node acts as an intermediate and sends RREPs (Route Reply) back to the previous node along with the route information by establishing a reverse route in a unicast manner. This process continues unless or until the packet reaches its destination[9]. In response to each RREQ, RREPs (Route Replies) are created.

2). DSR (dynamic source routing) is a reactive (on-demand) routing protocol. It utilizes the concept of source routing[16]. This protocol includes two phases, namely route discovery and route maintenance[6]. Additionally it also comprises three types of route control messages, i.e. Route Request (RREQ), Route Reply (RREP), and Route Error (RERR). In DSR, the process of route request and route reply is similar to that of AODV routing protocol. Source node can use the alternative path to the destination from the cache or it can reinitiate route discovery mechanism. It checks its route cache, if any node wants to communicate to destination node. The DSR protocol permits the network to self-organize and self-configure without any existing infrastructure or administrative support. Among the many benefits of DSR is that it provides excellent performance for routing in multi-hop wireless ad hoc networks, which reduces overhead because no periodic routing advertisements are required[16].

*B. **PROACTIVE ROUTING PROTOCOL***

Proactive routing protocols are called table-driven protocols in which, the routing table is maintained which keeps track of the routes to all the other nodes. Data packets are transmitted over the predefined route specified in the routing table from background information about neighbour nodes (routing updates) [5]. Details about every destination node's routes are stored in every mobile node in the form of routing tables. As the topology in a mobile ad-hoc network is dynamic it gets updated periodically. The limitation it has is that due to the need to maintain route information for all routes in the routing table, it doesn't work well with large networks where the entries become too large. The following are some of the most common table-driven protocols [6]: Optimized Link State Routing protocol (OLSR), Destination sequenced Distance vector routing (DSDV).

1). *OLSR* stands for ***Optimized Link State Routing*** *Protocol*. In mobile ad hoc networking, OLSR is a type of proactive routing protocol [7]. In this, every node regularly floods the status of its links. It provides routes to the requester immediately when demanded [4]. OLSR allows only selected nodes to flood data in the network instead of all of them; Nodes like these are referred to as MultiPoint Relays (MPRs). The upside of OLSR is that By limiting the time stretches

between occasional control message transmissions, OLSR responds rapidly to topological changes.. Only a partial link state needs to be flooded in order for OLSR to provide the shortest path routes. The minimal set of link state information required is, that it is essential all MPRs must declare the links to their MPR selectors [13]. This is the bare minimum set of link state information needed. In the case of more topological information, it MAY be utilized, e.g., as a redundancy function. OLSR is intended to work in a totally disseminated way and doesn't rely upon any central entity.

***2). Destination- Sequenced Distance-Vector Routing Protocol* abbreviated as *DSDV* is a famous routing protocol, which maintains a routing table at each node** [4]**.** It uses the hop count as metric in route selection. **It also maintains information like end-node generated sequence number, number of hops. it is a table driven algorithm [16].** A sequence number is associated with each entry, which makes it easier to identify stale entries. In this way, formation of routing loops can be avoided. Whenever possible, the route associated with the most recent sequence number will be used.

C. HYBRID ROUTING PROTOCOL

As its name suggests, in hybrid routing protocol, reactive and proactive protocols are combined. They take advantage of the other two types of protocols. In nature, these protocols are adaptive and adapt depending on the region and location of the source and destination nodes. As a result, routes are found quickly in the routing network. Zone routing protocol (ZRP), Distributed dynamic routing (DDR), Zone-based hierarchical link-state (ZHLS) are the types of hybrid protocols[14].

1). *ZRP (Zone Wise Routing Protocol)* In ZRP [1]**, every hub works inside the nearby and outside the extension with various communication strategies.**ZRP came into existence to support the larger range of communication between the different zones. ZRP is a combination of the best features of table-driven and on-demand routing protocols [1] that is why underline problems long waiting delays and packet overhead has been steep down. ZRP encourages two types of communication. using IARP (Intra Zone Routing Protocol) it performs Routing with nearby nodes. It uses routing tables that store information about all the paths within the zone before using the proactive method. Message delivery is immediate due to the availability of paths in the routing tables from source to destination. Communication between the farther nodes is performed by the reactive method. IERP (Inter-Zone Routing Protocol) handles the path created between the different zones with the help of intermediate nodes that are present on the boundaries of each zone. they use the reactive technique. In IERP, a request can be handily completed without searching and querying the whole network.

TABLE I
COMPARISON OF DIFFERENT ROUTING PROTOCOLS

PARAMETER	PROACTIVE	REACTIVE	HYBRID
Routing structure	hierarchical structures & flat structure	Mainly flat except for CBRP	Flat
Route discovery	Periodically	On-demand	Both
Control overhead	High	Low	Medium
Periodic updates	Yes	No	Yes
Reaction on failure	Slow	Fast	Fast
Route reconfiguration	Difficult	Easy	Easy
Power requirement	High	Low	Medium
Probability of congestion	Low	High	Low
Bandwidth requirement	High	Low	Medium
Routing information	Stored in routing tables	Does not stored	Provided when requirement is there
Control traffic	High	Low	Lesser than proactive & reactive
Benefit	Rapid establishment of routes & routing information is updated periodically.	Routes are obtained when required, loop free, and does not regularly exchange routing information.	These are more scalable & have updated routing information.
Drawback	Routing information flooded in whole network.	More packet dropping, large delay, and routes are not up-to-date	Requires more resources for larger zones[5].

DISTRIBUTED DENIAL OF SERVICE ATTACK (DDoS)

(DDoS) is presently a typical method for assault that influences truly network security and the quality of online administrations. The anticipation of authorized access to assets or the postponing of time-basic tasks. A distributed denial-of-service (DDoS) attack is a malicious attempt to disrupt the ordinary traffic of a targeted server, service, or network by overwhelming the target or its surrounding infrastructure with a flood of Internet traffic. Most DDoS attacks are intended to consume all accessible network bandwidth or resources on a target network, system, or site. The greyhole attack [8] is carried out at the networking layer and can act as a slow poison on the network end. Greyhole attacks involve malicious nodes advertising themselves as genuine nodes that are it's having a valid route to the destination node in order to intercept data packets.

A black hole attack[12] is an active denial of service attack in which a malicious node can attract all packets by falsely claiming a brand new route to the destination and afterward retaining them without sending them to the objective. A black hole attack happens at the network layer [9].

Wormhole attack[18] Among the most serious attacks on MANETs, this one is particularly dangerous. In order to achieve significant results from wormhole attacks, a minimum of two attackers is needed. Despite residing on different areas of the network, the attackers communicate through the tunnel to communicate with each other. They broadcast the wrong information to the other nodes in the network, indicating that the destination is only single hop away. When the route is discovered on the basis of lowest hop count between the source and destination pair, they also broadcast the incorrect information that they are neighbours. Because of this, the attacker node which is near to the source is selected more easily on the route between the source and destination pair. In the wormhole attack, no data packets are altered or false traffic is generated, so the attack cannot be detected easily.

The flooding attack is easy to perform but it brings out the most problems in the network. In RREQ flooding the attacker floods the RREQ in the entire network that catches a lot of the network sources. no node can answer RREP packets to these flooded RREQ. The attacker in data flooding gains access into the network and organizes routes among all the nodes in the network. Once the routes are set up, the attacker sends a boundless amount of useless data packets into the network which goes straight to all other nodes in the network which causes the system to deny the service or may lead to a system crash.

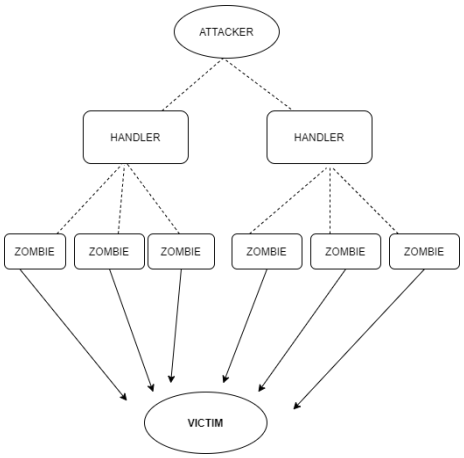


Fig.3 Architecture of DDoS attack [19]

PERFORMANCE METRICS

i. ***Packet delivery rate***: It is derived by the receiver's total packets (at destination) divided by the sender's total packets (from source)[4].

$$PDR = \frac{total\ number\ of\ packets\ recieved}{number\ of\ packets\ sent} \quad (1)$$

ii. ***Routing overhead***: It represents the total number of control packets generated in the network where n is the number of nodes generating a control packet

iii. ***Throughput*:** throughput is an effective transmission rate. it is total number of bits transferred (b_i) over the two devices or networks per unit time (t_i). It depends upon the capacity of the channel (n) and i is sequence number [20].

$$\text{Throughput} = \frac{1}{n}\sum_{i=1}^{n}\frac{b_i}{t_i} \quad (2)$$

iv. ***Average throughput***: It is calculated by the total size of packets received at the destination divided by the difference of stop and start simulation time.

$$\text{Average throughput} = \frac{total\ size\ of\ packets\ received}{difference\ between\ simulation\ start\ and\ stop\ time} \quad (3)$$

v. ***End to End Delay***: In mobile ad hoc, it is defined as the time it takes for a data packet to be transmitted from source to destination.

vi. ***Packet drop*:** it is calculated by subtracting the number of received packets ($P_{recieved}$) at destination from number of packets sent (P_{Sent}) at the source[4].

$$\text{Packet drop} = P_{Sent} - P_{recieved} \quad (4)$$

TABLE II
Comparison of different routing protocol approaches under the DDoS attack

AUTHORS	PROTOCOLS	ATTACK TYPES	METRICS	SIMULATOR	RESULTS	LIMITATION/ FUTURE SCOPE
Maha abdelhhaq, raed alsaqour, mada alaskar (2020)[1]	AODV, ZRP, LAR	Flooding attack	Throughput, end to end delay	NS2	ZRP worked best for both performance metrics	Include performance metrics like jitter , over head routing
Eman Farag Ahmed, Reham Abdellatif Abouhogai, Ahmed Yahya (2014)[6]	DSDV,OLSR, AODV,DSR	Blackhole and greyhole attack	E2E delay, packet drop, packet delivery ratio, routing overhead	NS2	Proactive survives longer in presence of DoS attacks.	-
Tejaskumar Bhatt, Chetan Kotwal, Nirbhay kumar Chaubey(2019) [7]	RIP, OSPF,EIGRP	-	Throughput, delay, utilization, queuing delay , network convergence	OPNET	EIGRP show is better than OSPF and RIP with respect to Delay, throughput, Network Utilization	-
Manju & Mrs Maninder kaur (2016)[8]	AODV, TORA	Flooding attack	delay, network load, packet drop rate, total no. of packets sent, throughput	NS2	AODV performed better than TORA	AODV protocol can be compared with the other candidate protocols used for MANET simulations

Dr. Gorine and Rabia Saleh (2019)[9]	AODV , DSR	Selfish node , blackhole, greyhole attack	Throughput, packet loss, average delay, energy per byte	NS2	Both protocols have been affected by the attacks	Comparison is done for reactive protocols only.
Parvinder Kaur, Dalveer Kaur, Rajiv Mahajan (2017)[10]	AODV, DSR,ZRP	Wormhole attack	Packet Delivery Ratio, Throughput and Packet Loss, jitter	NS2	the performance of AODV is more affected by the attack	comparative analysis and simulation of different routing protocols under various wormhole attacks can be performed
Sherin Hijazi, Mahmoud Moshref , Saleh Al-Sharaeh(2017)[11]	AODV, DSR	Blackhole attack	Throughput,Packet delay, Packet loss	NS2	enhanced AODV achieves the highest accuracy than blackhole AODV, but it's less accurate than normal AODV	aim to revise a new protocol called as "idsAODV" protocol

CONCLUSION

In this work, the systematic review of various research works has been done. The work of various authors has been analysed those evaluated the performance of various routing protocols when facing DoS attacks. The research work shows resistance of reactive, proactive, and hybrid protocols for various types of DoS attacks based upon different parameters like packet delivery rate, throughput, end-to-end delay, packet delay, and packet loss. In the presence of DoS attacks, proactive routing protocols survive longer. This is because in case a route is broken, each node can use pre-established paths as an alternative route. As compare to reactive and hybrid protocols, proactive protocols perform better in terms of packet drop, throughput and overhead. In this paper, the work focused on flat routing protocols of MANET which are suitable for small networks only. For future work, more work should be done for the protocols of larger networks like hierarchical and geographical protocols of MANET and their security. Security is a significant issue in MANET as cybercriminals are discovering ways to steal or tamper with data which is transmitted over wireless nodes in a network.

REFERENCES

[1] Maha Abdelhaq & Raed Alsaqour & Mada Alaskar & Fayza Alotaibi & Rawan Almutlaq & Bushra Alghamdi & Bayan Alhammad & Malak Sehaibani & Donia Moyna,(2020). "*The resistance of routing protocols against DDOS attack in MANET*". International Journal of Electrical and Computer Engineering (IJECE). Vol 10(5).doi: 10.11591/ijece.v10i5. pp4844-4852.

[2] Zuhairi, Megat & Kolade, Ayanwuyi & Yafi, Eiad & Zeng, Cai. (2017). "Performance analysis of black hole attack in MANET". Doi: 10.1145/3022227.3022228.

[3] Philomina S, Ramesh R. "*Secure Routing-Solution to Diminish DOS Attack in AODV Based MANET*". Int J Chem Sci. 2017; vol 15(4):188

[4] S. Abbas, M. Haqdad, M. Z. Khan, H. U. Rehman, A. Khan and A. u. R. Khan, "*Survivability Analysis of MANET Routing Protocols under DOS Attacks*," KSII Transactions on Internet and Information Systems, vol. 14, no. 9, pp. 3639-3662, 2020. DOI: 10.3837/tiis.2020.09.004.

[5] Talwar, Bhavna and Anuj Kumar Gupta. "*Ant Colony based Mobile Ad Hoc Networks Routing Protocols: A Review*." *International Journal of Computer Applications* 49 (2012): pp36-42.

[6] Ahmed, E.F. & Abouhogail, Reham & Yahya, Ahmed. (2014)." *Performance Evaluation of Blackhole Attack on VANET's Routing Protocols*". 8. pp39-54. 10.14257/ijseia.2014.8.9.04.

[7] Bhatt, Tejas. (2019). "*Implementing and Examination of EIGRP OSPF RIP Routing protocol in AMI Network for DDoS attack using OPNET*" IJRTE Research Paper. 10.35940/ijrte.B1490.0982S1119.

[8] Manju, Mrs Maninder kaur (2016). "*A Survey over the Critical Performance Analytical Study of the MANET Routing Protocols (AODV & TORA*)". International Journal of Advance Research, Ideas and Innovations in Technology, vol2 issue (4) www.IJARIIT.com.

[9] Gorine, Adam and Saleh, Rabia (2019) "*Performance Analysis of Routing Protocols in MANET under Malicious Attacks*". International Journal of Network Security and Its Applications, 11 (2). pp. 1-12. doi:10.5121/ijnsa.2019.11201.

[10] Kaur, Parvinder & Kaur, Dalveer & Mahajan, Rajiv. (2017). "*Simulation Based Comparative Study of Routing Protocols Under Wormhole Attack in Manet*". Wireless Personal Communications. 96. pp47–63 doi: 10.1007/s11277-017-4150-2.

[11] Hijazi, Sherin & Moshref, Mahmoud & al-sharaeh, Saleh. (2017). "*Enhanced AODV Protocol for Detection and Prevention of Blackhole Attack in Mobile Ad Hoc Network*". Vol 1 issue (6). pp7535-7547. Doi: 10.24297/ijct.v16i1.5728.

[12] https://www.geeksforgeeks.org/manet-routing-protocols/

[13] Thomas Clausen , Philippe Jacquet. "*Optimized Link State Routing Protocol (OLSR*)," 2003. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc3626.

[14] Kumar, Jaspal & Kulkarni, Muralidhar & Gupta, Daya. (2013). "*Effect of Black Hole Attack on MANET Routing Protocols*". International Journal of Computer Network and Information Security. Vol 5. pp64-72. Doi: 10.5815/ijcnis.2013.05.08.

[15] Harjeet Singh Chhabra, Ashish Mundhra, Yashasya Sharma, 2014, "*Study and comparative Analysis of AODV and DSR Routing Protocol"*, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) ETRASCT – 2014 (Volume 2 – Issue 03), ISSN (Online) : 2278-0181 pp. 255-259

[16] Barve, Amit; Kini, Ashwin; Ekbote, Onkar; Abraham, Jibi (2016). [IEEE 2016 2nd International Conference on Communication, Control & Intelligent *Systems (CCIS)* - Mathura, India (2016.11.18-2016.11.20)] 2016 2nd International Conference on Communication Control and Intelligent Systems (CCIS) – "*Optimization of DSR routing protocol in MANET using passive clustering*",pp23-27. doi:10.1109/CCIntelS.2016.7878193.

[17] "*Destination-Sequenced Distance Vector Routing (DSDV),*" BINUS university graduation program, [Online]. Available: https://mti.binus.ac.id/2014/08/15/destination-sequenced-distance-vector-routing-dsdv/.

[18] Hemant Pareek, Vishal Shrivastva "*Denial of Service Attacks Implementation and Detection Approach for MANET* "International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 2, February 2014 ISSN (Online) : 2278-1021, 2014

[19] Shi Dong Khushnood Abbas And Raj Jain (2019) "A Survey on Distributed Denial of Service (DDoS) Attacks in SDN and Cloud Computing Environments" IEEE vol 7, pp80813 -80828 doi: 10.1109/ACCESS.2019.2922196

[20] Shashi Gurung1,Siddhartha Chauhan (2019) " *Performance analysis of black-hole attack mitigation protocols under gray-hole attacks in MANET"* Wireless Networks vol 25: pp975–988 doi: 10.1007/s11276-017-1639-2

TECHNIQUES OF HANDLING MISSING VALUES IN DATA MINING: A REVIEW

Harmanpreet Singh[*1], Amrit Kaur[#2]
[#]*Department of Computer Science and Engineering, Punjabi University, Patiala*
[1]gillharman1307@gmail.com
[2]amrit.tiet@gmail.com

ABSTRACT— The issue regarding missing data exists from the very beginning of data mining era. In real world it is almost impossible to get a data set free from the missing values. Missing data can directly responsible for lower efficiency and quality of machine learning model which further leads to biased conclusions and invalid decisions. To deal with the inconsistent data there exists numerous data preprocessing techniques such as data cleaning, transformation and reduction etc. Data Cleaning process especially deals with problems related to outliers, missing values and duplicate data. The presence of missing values in data can be from various sources like human error, technical error or from hardware failure. In this paper different types of missing data and techniques to handle the missing values are reviewed and discussed.

KEYWORDS— Missing Values, Imputation techniques, MCAR, MAR, MNAR, Bayesian, KNN, Multiple Imputations, Maximum Likelihood, SVM

INTRODUCTION

Data mining simply refers to mine the data for valuable information. The quality of data used for mining plays a significant role for desirable outcomes. More often the data available for mining process is not in the state of perfection that it would be consider as a sensible fit for the task. The gathered data must go through a refinement process known as preprocessing. Data preprocessing involves data cleaning, integration, transformation and reduction related tasks to improve the overall quality of the data. Although the outcome may depend upon several factors like algorithms used, sampling techniques and attribute selected, but key dependency lies on the fact that how efficiently researchers handle missing values[1]. Missing values are the qualities of attributes related to data that are missing from the data set which sometimes influence the mined information critically. The problem of learning from inconsistent or incomplete data becomes fundamental in the field of the data mining[2]. Missing data in data set presents various threats to data mining that it may significantly contribute for increase in the computational cost, skewed results. Ineffective handling of missing values can reduce the statistical power of the model, cause bias in estimation of parameters or can complicate the study which further leads to invalid conclusions[1], [3].

Effective way to deal with the missing data requires an essential task to observe first i.e., to determine the pattern and type of the missing instances which can be the deciding factor for selection of the technique used to handle the missing data[4]. A basic way to handle missing data is simply ignore the records that contain the missing values but it can produce biased result if the size of sample is limited or a significant quantity of data is missing. To minimize the effect of the loss of data several imputation techniques are used. These techniques can be statistical such as mean mode substitution or can be in the form of machine learning algorithms like KNN or K-means etc. Maximum likelihood and expected maximization are the other some model-based techniques. This article can be viewed as two sections i.e., section one explains types of missing values and section two discuss about the various techniques used to handle the missing data.

MISSING VALUES CATEGORIZATION

According to Rubin[5] three mechanisms are described to explain the relationships of missingness with the values of attributes or variables in data matrix. These are as follows:

Missing at Random (MAR)

The MAR values are those random missing values where missing pattern depend or can be observed via some known variable [6].

Missing Completely at Random (MCAR)

The missing value said to MCAR if the missing data items are independent of both the observable and non-observable variables or parameters that means missing entirely at random i.e., maximum level of randomness [7].

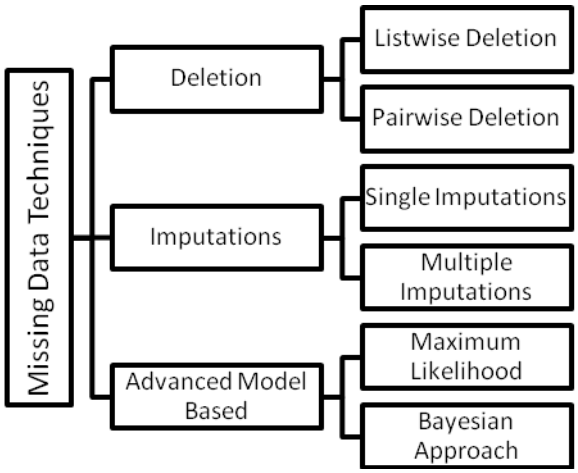
Missing not at Random (MNAR)

This mechanism occurs when values that are missing depend upon the other missing values so that researchers cannot use the available data for approximation of missing data, also described as non-ignorable case [8].

HANDLING MISSING VALUES

The best solution to missing data is to avoid it in the first place and thus to prevent the missing values in data by enhancing the data collection methodology. But it is not as simple as it in theory, for real problems it can comprises the redesign of whole data collection process from scratch or can say the existing system right from the planning to current state of the system.

Missing values can be handled by various methods such as the most common one is deletion or ignoring and the other one is imputing the missing value with the calculated or predicted one. There exist other advanced model-based methods like



multiple imputation, maximum likelihood and Bayesian simulations etc. Missing data handling techniques can be seen as in three groups.

Fig. 9 Missing data techniques

One other class is also mentioned by Rubin & Little i.e., Weighting Procedure[5] used to reduce bias in complete case analysis for nonresponsive and noncoverage[9]. This section includes some popular techniques for handling Missing Values.

Deletion Methods

The simplest way is to delete instances with missing data. This can be exercised mainly via three cases[10] such as complete case analysis (listwise deletion) in which delete all the instances with at least one missing value. Available case analysis (Pairwise deletion) where delete only instances for needed variables out of total available variables. Weighting complete case analysis where model the missingness to reduce the bias for nonresponse variables.

Mean Mode Imputation

In this technique missing values are imputed by their corresponding attribute mean and mode. Mode case mostly used for categorial variables. This method can lead to biased result caused by underestimating the standard error that it considers the imputed values as actual values[1] . Only benefit is the simplicity for implementation.

Regression Imputation

In this different regression models are used to predict missing value. In regression, variables assumed to be in linear relationship with other variables in data set. So that missing values are replaced by using the linear relationship functions. SVM (Support Vector Machine) based imputation is an example of Regression method.

Hot Deck & Cold Deck Imputation

In hot deck imputation missing values (recipient or nonrespondent) are replaced by values predicted from similar distribution (usually called donor or respondent) in dataset. Data is divided into smaller decks or portion to analyze. On the other hand, in cold deck case missing value are replaced by value provided by external source. It can be from previous dataset from same survey [11] [12].

KNN (K-Nearest Neighbor)

KNN is a method based on finding the nearest neighbor in terms of Euclidian distance. In this k number of similar records i.e., nearest neighbors are calculated in respect of Euclidian distance. And missing values are replaced by their respective k-neighbors [13]. The advantages of KNN method are that it can be used both for qualitative and quantitative attributes and also no need to build different predictive models for each attribute[14].

Clustering

Clustering is method to define cluster of values based on some properties. Clusters can be governed by minimize the sum of squares of data points and cluster centroids [[15]. Clustering technique estimate the value using cluster location for that missing instance. Fuzzy mean imputation based on fuzzy logic uses k-mean or c-mean clustering algorithms to find out the respective cluster for missing values[16].

ANN (Artificial Neural Network)

ANN is basis for Artificial intelligence which simulates the functioning of a human brain. In ANN based prediction multi-layer perceptron method can be used to predict the missing values. Mostly feed forward multilayer perceptron method is used. ANN based Model can be trained via estimating the available data points as training set [6] [16].

Bayesian Approach

Bayes theorem basically used to find out conditional probabilities[17]. It is a logical approach to updating the probability of hypotheses in the light of new evidence[18]. Missing data in Bayesian frameworks are considered as random variables that can be sampled by using their corresponding conditional distributions. Posterior distributions are used to make inferences[19].

Maximum Likelihood

In maximum likelihood parameters are estimated which are likely to be in the resulted data. Likelihood for complete data case for all variables and some of the attributes get computed separately and maximized to estimate parameters in ML. ML is considered as an alternative to Posterior draw based on Bayesian Posterior Distributions. As compared to Posterior, ML yields slightly better results, is faster and computationally intensive [20]. Expectation-Maximization (EM) is a type of maximum likelihood method, an iterative way in which values estimated by the maximum likelihood methods are used to impute missing values and likely to create new data sets [3].

Decision Tree

Different decision tree algorithms are used to estimate the missing values like C4.5, a tree-based algorithm uses entropy-based measure known as gain ration. Gain ration is used to predict best value which used to divide the data point into smaller subsets[11]. Another hybrid method known as DIFC also proposed, is based on iterative fuzzy clustering and decision trees. DIFC basically an iterative method to impute missing values which combines both the unsupervised (fuzzy clustering) and supervised learning (decision trees) as one [21].

A BRIEF COMPARATIVE STUDY

A short review regarding study (different techniques of handling missing data published as various articles) is summarized below as a table.

TABLE I
SUMMARY OF PUBLISHED ARTICLES

Ref	Dataset	Technique/Algorithm	Performance Metrics	Description	Remarks
[1]	1. Local health dataset 2. Hair eye color dataset 3. UCI car dataset 4. Kaggle house price dataset	1. MICE-BLR 2. SICE-BLR 3. MICE-CART 4. SICE-CART 5. KNN 6. Amelia 7. FURIA 8. SVM 9. LDA 10. PolyReg 11. PMM	RMSE (Root Mean Square Error)	SICE an Improved method based on MICE (Multivariate Imputation by Chained Equation)	SICE-Categorical shows better performance over MICE, SICE requires slightly higher execution time than MICE
[6]	UNSODA	1. Random Forest (RF) Regression 2. Support Vector (SVR) Regression 3. Artificial Neural Network (ANN) Regression 4. Mean Imputation 5. Multiple Imputation (MI)	1. RMSE (Root Mean Square Error) 2. MLE (Multiple Linear Regression)	Comparative study for various methods for handling missing data	RF and MI methods are best for imputing in UNSODA. Standard error significantly decreased after imputation
[8]	1. Power Plant Data 2. HIV Data 3. Industrial winding Data	1. Expectation Maximization 2. Autoencoder Neural Network	1. Correlation Coefficient 2. Relative Prediction Accuracy	Comparison of Neural Network and Expectation maximization using Genetic Algorithm	Imputation ability is problem dependent. EM performs better in less dependent variable cases

Ref	Dataset	Technique/Algorithm	Performance Metrics	Description	Remarks
[12]	Real Data from Norwegian OMT programme	1. Hot deck imputation 2. Multiple imputation using latent class analysis 3. Multiple imputation using expectation–maximization with bootstrapping 4. Multivariate imputation by chained equations 5. Multiple imputation using multiple correspondence analysis 6. Complete Case Analysis	1. Relative Bias of Regression coefficient 2. Stability of estimates 3. Coverage and width of confidence intervals	Comparison of different imputation methods	CCA performs well for low missingness. MIMCA performed best overall.
[13]	Benchmark data from UCI & Real Data	1. Mean 2. Hot Deck 3. KNN 4. BPCA (Bayesian estimation of PCA) 5. KNBP (Hybrid of KNN & BPCA)	1. RMSE (Root Mean Square Error) 2. MLE (Multiple Linear Regression)	Imputation And Normalization Techniques	KNBP generate lower error rate than other traditional methods
[14]	Student Dataset	1. Listwise Deletion 2. Mean Imputation 3. KNN	C4.5/J48 classification algorithm using confusion matrix	Imputation in multi attribute data set	KNN performs better than other two
[22]	1. Soybean (small) 2. Postoperative patient data 3. Promoters 4. Monks 1 5. Monks 2 6. Monks 3 7. Balance 8. Tic-tac-toe 9. CMC 10. Car 11. Splice 12. Kr-vs-kp 13. LED 14. Nursery 15. Kr-V-K	1. Hot Deck 2. Framework Hot Deck 3. Naïve Bayes 4. Framework Naïve Bayes 5. Polytomous Regression 6. Mean Imputation	Six classifiers are used such as KNN, RIPPER, C4.5, SVM with RBF Kernel, SVM with Polynomial kernel, Naïve Bayes. Zero-one loss method is used for classifier performance. Imputation significance is shown by using t-values.	Impact of imputation of missing values for discrete data on classification error	There is no best imputation method, imputation improvements are classifier dependent
[23]	1. Iris 2. Wine 3. Glass 4. Liver disorder 5. Ionosphere 6. Statlog Shuttle	1. Listwise Deletion 2. Mean 3. Group Mean 4. Predictive Mean 5. KNN 6. K-Mean 7. Hot Deck	RMSE (Root Mean Square Error)	Optimal selection of an imputation method	predictive mean method yielded overall best results & hot-deck imputation the worst.
[24]	1. Breast Cancer Wisconsin 2. Chronic kidney disease 3. Congressional voting records 4. Credit approval 5. Cylinder bands 6. Heart disease-ungarian 7. Hepatitis 8. Horse colic 9. Mammographic mass 10. Ozone level detection	1. BART with MIA (Missingness Incorporated in Attributes) 2. BART with missForest approach 3. Random Forest with missForest approach	Cross Validation	Handling missingness using Bayesian additive regression trees (BART) Model for classification	BART model with MIA approach provides best accuracy and low run-time

Ref	Dataset	Technique/Algorithm	Performance Metrics	Description	Remarks
[25]	1. Breast Cancer Dataset 2. Hepatitis Dataset 3. Lung Cancer Dataset	1. Mean 2. Imputation 3. K-NN 4. Weighted- 5. KNN 6. Fuzzy 7. K-NN	NRMSE (Normalized Root Mean Square Error)	Enhanced Fuzzy K-NN method for Handling Missing Data	Fuzzy K-NN out performs the other methods

DISCUSSION & CONCLUSION

The inevitability of missing data is not deniable in most cases. It can be quite challenging to handle missing values effectively, as it includes the careful study that what type missingness in data present and how different techniques work on the missing data. In short, the choice of method to deal with missing data is crucial in all studies. The method used to treat the data should satisfy these three conditions [26] that there should be no bias in estimation; second is the relationship among variables must be retain and last is minimal the cost (time & complexity). Various techniques existed to handle the missing data, yet there is no universal technique which performs best in all cases [27]. In comparison, KNN is mostly used traditional method for imputation [13]. Paul argues that Maximum likelihood is preferable over the multiple imputation in availability of software cases [28]. Thus, the problem in hand and the data available is significant for the study and overall efficiency of the model.

REFERENCES

[1] S. I. Khan and A. S. M. L. Hoque, "SICE: an improved missing data imputation technique," *Journal of Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00313-w.

[2] M. M. Marek´smieja, Ł. Struski, J. Tabor, B. Zielí, and P. Spurek, "Processing of missing data by neural networks."

[3] H. Kang, "The prevention and handling of the missing data," *Korean Journal of Anesthesiology*, vol. 64, no. 5. pp. 402–406, May 2013. doi: 10.4097/kjae.2013.64.5.402.

[4] M. Soley-Bori, M. Horn, J. Morgan, and K. Min Lee, "Dealing with missing data: Key assumptions and methods for applied analysis," 2013.

[5] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, vol. 3rd edition. 2020.

[6] Y. Fu, H. Liao, and L. Lv, "A comparative study of various methods of handling missing data in unsoda," *Agriculture (Switzerland)*, vol. 11, no. 8, Aug. 2021, doi: 10.3390/agriculture11080727.

[7] S. Rawal, S. C. Gupta, and M. S. Singh, "Predicting Missing Values in a Dataset: Challenges and Approaches," 2017.

[8] F. v Nelwamondo, S. Mohamed, and T. Marwala, "Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques."

[9] G. Kalton and I. Flores-Cervantes, "Weighting Methods," 2003. [Online]. Available: www.asc.org.uk

[10] C. M. Salgado, C. Azevedo, H. Proença, and S. M. Vieira, "Missing data," in *Secondary Analysis of Electronic Health Records*, Springer International Publishing, 2016, pp. 143–162. doi: 10.1007/978-3-319-43742-2_13.

[11] S. Singh and J. Prasad, "Estimation of Missing Values in the Data Mining and Comparison of Imputation Methods," *Mathematical Journal of Interdisciplinary Sciences*, vol. 1, no. 2, pp. 75–90, Mar. 2013, doi: 10.15415/mjis.2013.12015.

[12] M. R. Stavseth, T. Clausen, and J. Røislien, "How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data," *SAGE Open Medicine*, vol. 7, p. 205031211882291, Jan. 2019, doi: 10.1177/2050312118822912.

[13] K. Manimekalai and A. Kavitha, "MISSING VALUE IMPUTATION AND NORMALIZATION TECHNIQUES IN MYOCARDIAL INFARCTION," *ICTACT JOURNAL ON SOFT COMPUTING*, p. 3, 2018, doi: 10.21917/ijsc.2018.0230.

[14] M. Rajan and V. Gimpy, "Missing Value Imputation in Multi Attribute Data Set." [Online]. Available: www.ijcsit.com

[15] S. Jain and M. Kalpana Jain, "Estimation of Missing Attribute Value in Time Series Database in Data Mining," *Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc*, vol. 16, 2016.

[16] A. Puri, S. Mata Vaishno, and M. Gupta, "Review on Missing Value Imputation Techniques in Data Mining," 2017. [Online]. Available: https://www.researchgate.net/publication/329625460

[17] Pavan Vadapalli, "Bayes Theorem in Machine Learning: Introduction, How to Apply & Example," Feb. 04, 2021. https://www.upgrad.com/blog/bayes-theorem-in-machine-learning/ (accessed Oct. 28, 2021).

[18] D. Berrar, "Bayes' theorem and naive bayes classifier," in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, vol. 1–3, Elsevier, 2018, pp. 403–412. doi: 10.1016/B978-0-12-809633-8.20473-1.

[19] Z. Ma and G. Chen, "Bayesian methods for dealing with missing data problems," *Journal of the Korean Statistical Society*, vol. 47, no. 3. Korean Statistical Society, pp. 297–313, Sep. 01, 2018. doi: 10.1016/j.jkss.2018.03.002.

[20] P. T. von Hippel and J. W. Bartlett, "Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws," *Statistical Science*, vol. 36, no. 3, Jul. 2021, doi: 10.1214/20-sts793.

[21] S. Nikfalazar, C. H. Yeh, S. Bedingfield, and H. A. Khorshidi, "Missing data imputation using decision trees and fuzzy clustering with iterative learning," *Knowledge and Information Systems*, vol. 62, no. 6, pp. 2419–2437, Jun. 2020, doi: 10.1007/s10115-019-01427-1.

[22] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, no. 12, pp. 3692–3705, Dec. 2008, doi: 10.1016/j.patcog.2008.05.019.

[23] J. Sim, J. S. Lee, and O. Kwon, "Missing values and optimal selection of an imputation method and classification algorithm to improve the accuracy of ubiquitous computing applications," *Mathematical Problems in Engineering*, vol. 2015, 2015, doi: 10.1155/2015/538613.

[24] K. Mehrabani-Zeinabad, M. Doostfatemeh, and S. M. T. Ayatollahi, "An efficient and effective model to handle missing data in classification," *BioMed Research International*, vol. 2020, 2020, doi: 10.1155/2020/8810143.

[25] R. Naveen Kumar and M. Anand Kumar, "Enhanced Fuzzy K-NN Approach for Handling Missing Values in Medical Data Mining," *Indian Journal of Science and Technology*, vol. 9, no. S1, Dec. 2016, doi: 10.17485/ijst/2016/v9is1/94094.

[26] M. Sharik, U. Zama, and S. Ramasubbareddy, "Missing values analysis techniques in Data mining: Review 1*," *International Journal of Advanced Science and Technology*, vol. 28, no. 15, pp. 377–382, 2019.

[27] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," *Knowledge and Information Systems*, vol. 32, no. 1, pp. 77–108, Jul. 2012, doi: 10.1007/s10115-011-0424-2.

[28] P. D. Allison, "Handling Missing Data by Maximum Likelihood," 2012.

DETAILED REVIEW OF HISTOGRAM EQUALIZATION TECHNIQUES

Komal Sharma[#1], Rakesh Singh[#2]

[#]*Department of Computer Science and Engineering, Punjabi University Patiala*

[1]komalsharma24126@gmail.com

[2]rakesh_ce@pbi.ac.in

ABSTRACT— Nowadays, technology plays a vital role in all spheres of human life such as smartphones, smartwatches, electric cars, medical images, etc. We are on the roads of digital transformation so there is a need of maintaining the quality of digital images which in turn leads to various image enhancement techniques. As we probably are aware, histogram equalization is one of the most often used technique for image enhancement. The primary goal of histogram equalization is that it provides us a uniform histogram. The modern world allows us to store and download digital images so getting a high-quality image becomes a major concern to percept sufficient data from an image.

KEYWORDS— Image Enhancement, Histogram, Various Types of Histograms, Histogram Equalization Mathematically, Contrast Enhancement Techniques

I. INTRODUCTION

Image enhancement is defined as a process that is used to improve the quality of an image as it improves the interpretability of information present in images for human beings and as well as it provides better input for other automated image processing techniques. Contrast is defined as the difference in intensities due to which we are able to distinguish the objects from other objects in an image. Contrast enhancement is defined as a technique which is used to improve the brightness of an image. There are various reasons for an image to have poor contrast. These are: - the incorrect setting of the aperture, poor quality of the imaging device. Due to this, images may not reveal all the relevant details in the captured scene. As a result of this, the image will provide a washed-out and unnatural look.

II. LITERATURE SURVEY

In 2011, Sapana S. Bagade, and Vijaya K. Shandilya explained histogram equalization to enhance the low contrast of an image. This method used three steps; in the first step, histogram formation is performed. Then, in a second step, new intensity values are assigned. After that in the third step, previous intensity values are replaced with the new intensity values.

In 2017, a new method for image enhancement for medical images has been proposed by Akash Gandhamal, Sanjay Talbar, Suhas Gajre, Ahmad Fadzil M. Hani, Dileep Kumar which is local gray level S curve transformation. The contrast of an image is enhanced by enlarging the difference between the highest and lowest gray level.

In 2016, Rashmi Choudhary and Sushopti Gawade proposed Brightness preserving Bi-Histogram Equalization (BBHE) which is used to overcome the drawbacks of histogram equalization. Basically, it decomposes the original image based on its mean value and then applies histogram equalization on each sub-image independently.

In 2013, Tarun Maheshwari presented a fuzzy logic technique for image enhancement in which the image is enhanced by three steps. The first step is image fuzzification. The second step is membership modification and the third step is image defuzzification.

In 1988, John B. Zimmerman, Stephen M. Pizer, Edward V. Staab, J. Randolph Perry, William Mccartney, and Bradley C. Brenton proposed another technique for contrast enhancement is Global Histogram Equalization (GHE) is a simple and effective technique for contrast enhancement and it provides us a resulting image with constant intensity values by using a cumulative density function.

III. HISTOGRAM

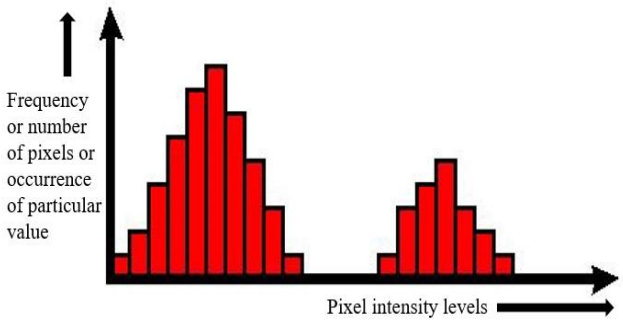


Fig. 1 Histogram in digital image processing

In the context of digital image processing, the histogram is a graph that represents the number of pixels in an image at each different intensity value. For a grayscale image, the range of intensity levels will be from 0 to L-1. The highest grey

level detected in an image is represented by L-1. For an 8-bit grayscale image, there are 256 possible intensities such that the range of intensity values will be from 0 to 255. We can see in the given histogram, here the x-axis represents the pixel intensity levels and the y-axis represents the number of pixels or the occurrence of a particular value in the captured scene.

IV. CONTRAST ENHANCEMENT TECHNIQUES

A. Histogram Equalization

Histogram Equalization is an image processing technique that is used to adjust the contrast of an image by altering the intensity distribution of the histogram. Histogram equalisation, in a nutshell, re-distributes the intensities. It is used to improve the contrast in images in order to provide a better quality of images without the loss of any information. With this technique, we can obtain sufficient data from images with poor contrast to percept that image. This method's primary purpose is to give the cumulative probability function a linear trend. Basically, we use a function named as histeq in MATLAB for histogram equalization. Histeq function produces an output image that has high contrast and pixel values that are evenly or uniformly distributed throughout the intensity range associated with that image. In layman's terms, we can say that this function automatically adjusts the intensity values of pixels such that it can improve the performance of our computer vision and machine learning tasks. If the histogram of an image has many peaks and valleys, then we can use this function to flatten the histogram. We can perform histogram equalization with the help of a function named histeq on a .tif file that is tag image file format.

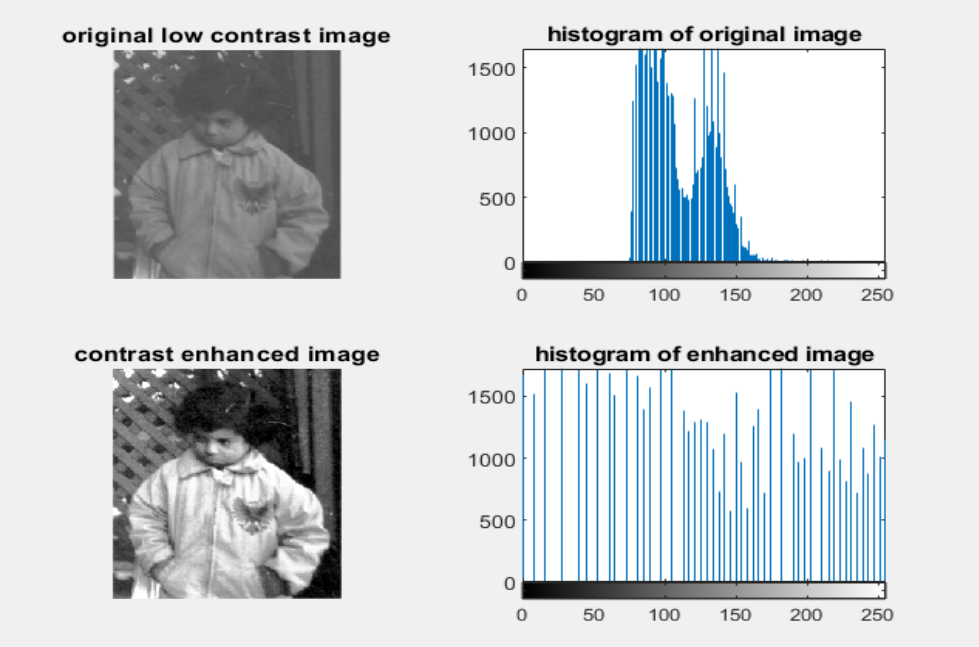
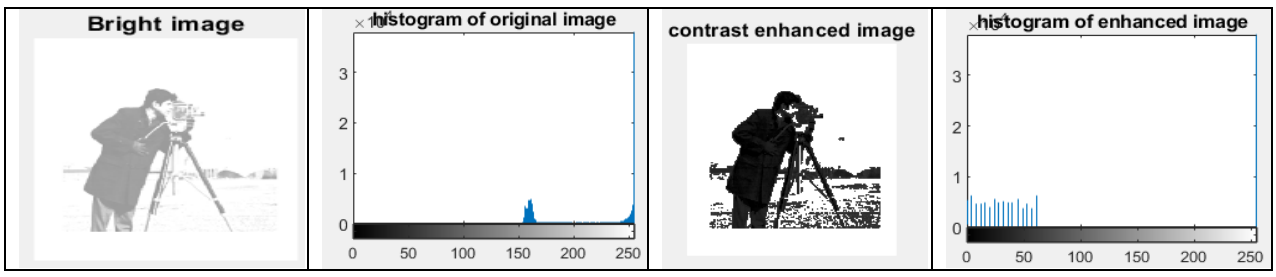


Fig. 2 Experimental result of original low contrast image and contrast enhanced image by using Histogram Equalization

We have used the Histogram Equalization approach in this experiment and we can notice a significant difference between the original image and the high contrast image.

V. VARIOUS TYPES OF HISTOGRAM



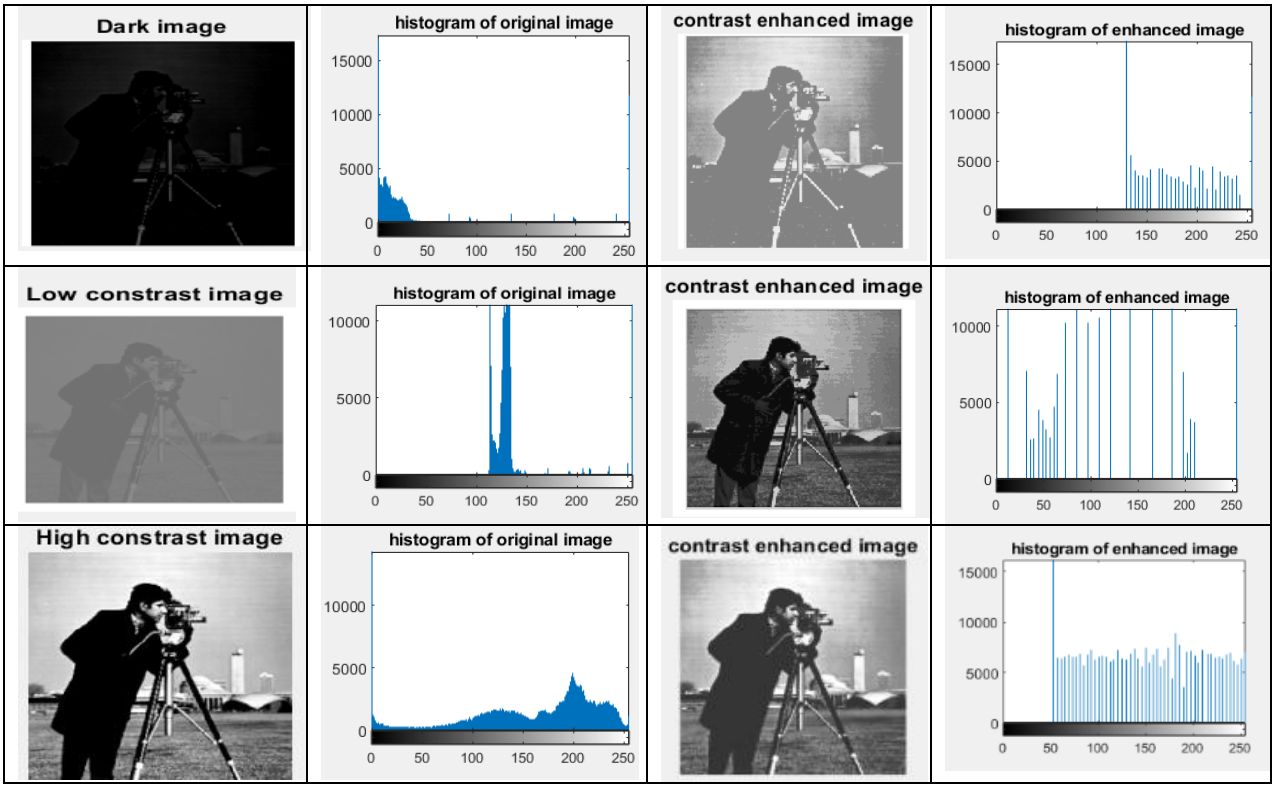


Fig. 3 Shows Various Types of Histograms

For different sorts of images, there are numerous forms of histograms. For dark images, the components of the histogram of an image are concentrated on the low side of the grayscale. For bright images, the components of the histogram are concentrated on the high side of the grayscale. For low contrast images, the components of the histogram are concentrated on the centre and have a narrow shape towards the middle. For high contrast images, the components of the histogram cover a broad range over the grayscale. So we can say that from the histogram only we can get a glimpse of an image like whether it is dark, bright. After histogram equalization, the peaks and valleys of the histogram will be shifted and we can get a uniform histogram such that each pixel will be assigned a new intensity value with respect to its previous intensity level. The histogram equalization is operated on an image by the following steps.

STEP 1. Find the sum of histogram values.
STEP 2. Normalize the histogram.
STEP 3. Multiply step 2 by the highest gray level and round the values.
STEP 4. Map the gray level values from step 2.

$$S_K = T(r_k) = (L-1)\sum_{I=0}^{K} P\,r(r_j)$$

Here j=0,1,2,…..k.

In this equation, S_k represents the output values after histogram equalization. $T(r_k)$ is the transformation on r_k. L-1 is the highest intensity level. $P_r(r_j)$ is the probability of occurence on r_j.

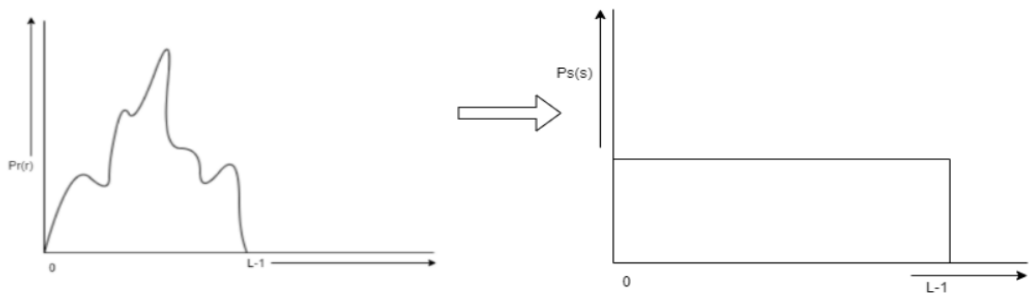


Fig. 4 (a)Arbitrary Function (b) Transformation function

Our main goal is to go from a given distribution (a) to a uniform distribution(b). In order to find this, we need a transformation function that is s=T(r). Now we are going to prove this mathematically, the gray levels in the image can be viewed as random variables taking values in the range [0,1]. Let $p_r(r)$ is the probability density function of input level r and let $p_s(s)$ is the probability density function of s. Let us assume that s = T (r). In practice we don't exactly get a uniform histogram as this function could result in a non-uniform values so we have to approximate these values to make integer values.

$$P_s(s) = P_r(r) \left|\frac{dr}{ds}\right| \quad \text{.....(1)}$$

Here, $P_s(s)$ is the new value which we want to obtain. So the new probability density function is equal to all probabilty

density function of r times the absolute value of the derivative of r with respect to s.

$$s=T(r)=(L-1)\int_0^r P_r(\omega)d\,\omega \quad \text{.....(2)}$$

here (L-1) is the maximum gray level. We are integrating the current probabilty distribution .ω is the dummy variable.
Now we are going to compute the derivative

$$\frac{ds}{dr} = \frac{dT(r)}{dr}$$

Now we are going to substitute the value of T(r) from (2) equation,

$$=(L-1)\frac{d}{dr}\left[\int_0^r P_r(w)\,dw\right]$$

$$=(L-1)\,P_r(r)$$

Now we will substitute the value of $\frac{ds}{dr}$ in (1) equation

$$P_s(s) = P_r(r) \bmod \frac{dr}{ds}$$

$$=P_r(r)\left|\frac{1}{(L-1)P_r(r)}\right|$$

$$=\frac{1}{(L-1)} \quad 0<s<L-1$$

Hence proved.

B. Bi-Histogram Equalization

Basically, the Bi-Histogram Equalization Technique is used to overcome the drawbacks of histogram equalization. it decomposes the original image into two sub-images such that one image has its range value from minimum gray-level to mean value and the other has its range from mean value to maximum gray level. After that both the histograms are normalized independently. The main goal of this technique is to preserve the brightness of an original image.

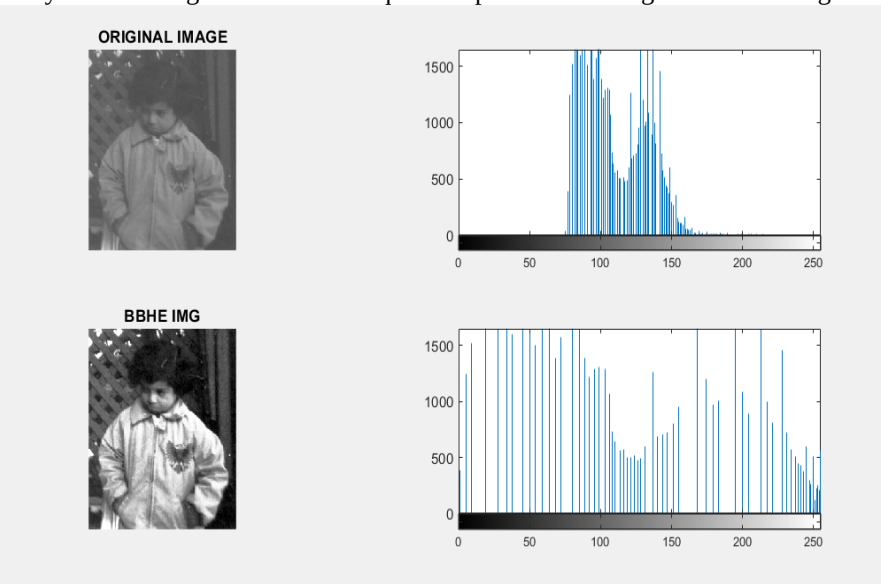


Fig. 5 shows the implementation of BBHE technique

We have used the Bi-Histogram Equalization approach in this experiment and we can notice a significant difference between the original image and the high contrast image.

C. Dynamic Histogram Equalization

Primarily, the Dynamic Histogram Equalization decomposes the input histogram of an image into a number of sub-histograms according to their local minima. After that dynamic histogram equalization is applied and all the sub-histograms are merged.

D. Fuzzy Logic Technique

The Fuzzy Logic Technique is used for image enhancement. Essentially, three stages are used to improve the image. To begin with, fuzzification is defined as the coding of picture data, and defuzzification is defined as the decoding of image data using a membership function.

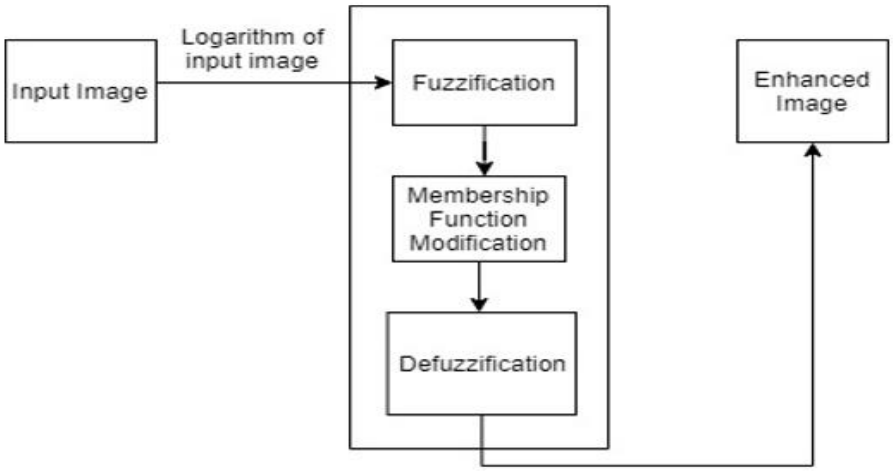


Fig. 6 Fuzzy Logic Technique

E. Adaptive Histogram Equalization

As we know, Adaptive Histogram Equalization is an image processing technique that is used to enhance the contrast of images. Basically, it computes several histograms and each histogram corresponds to a distinct section of an image and redistributes the intensities accordingly.

F. Local gray level –S transformation

Generally, Local gray level S-curve transformation is used for enhancing the contrast of a given image. This approach eliminates the drawbacks of the global method, such as over-enhancing, and produces a good contrast enhancement in medical images. The blocking effects can be decreased to a negligible level by increasing the number of block divisions in the source image. To solve the global transformation's constraints of large disparities between maximum and minimum intensity, a local operation is done, which reduces the intensity difference. Basically, the growth in the number of block divisions above a specific ideal range may have an impact on contrast enhancement.

VI. CONCLUSION

Histogram Equalization is a simple and effective method for image enhancement. As we know the concept of histogram equalization is being used in medical applications such as X-rays, MRIs, CT scan so that doctors are able to diagnose the disease more easily. Medical imaging has a crucial impact as the quality of healthcare is directly concerned with the health of a patient. This technique is used to overcome the linear stretch defect in which an equal number of intensity levels are assigned irrespective of the number of pixels associated with that image. The main objective is to obtain a uniform histogram with high contrast image.

REFERENCES

[1] Akash Gandhamala, S. T. (2017). Local gray level S-curve transformation – A generalized contrast enhancement technique for medical images. *Computers in Biology and Medicine*, 14.

[2] Gonzalez, R. C., & Woods, R. (2007). *Digital Image Processing.*

[3] JOHN B. ZIMMERMAN, S. M. (1988). An Evaluation of the Effectiveness of Adaptive Histogram Equalization for Contrast Enhancement . *IEEE TRANSACTIONS ON MEDICAL IMAGING, VOL. 7. NO. 4*, 9.

[4] Rashmi Choudhary, S. G. (2016). Survey on Image Contrast Enhancement Techniques. *International Journal of Innovative Studies in Sciences and Engineering Technology*, 5.

[5] Shandilya, V. K., & Bagade, S. S. (2011). USE OF HISTOGRAM EQUALIZATION IN IMAGE PROCESSING FOR IMAGE ENHANCEMENT. *International Journal of Software Engineering Research*, 5.

[6] Tarun Mahashwari, A. A. (2013). Image Enhancement Using Fuzzy Technique. *INTERNATIONAL JOURNAL OF RESEARCH REVIEW IN ENGINEERING SCIENCE & TECHNOLOGY*, 4.

KIDNEY ABNORMALITY DETECTION AND SEGMENTATION

Saloni Devi[1], Supreet kaur[2]
Department of Computer Science, Punjabi University, Patiala [1]salonidandyan@gmail.com [2]supreetgill13@gmail.com

ABSTRACT— Chronic kidney disease (CKD) is a systematic, irreparable loss of renal under which the capability of the body to maintain fluids, and electrolyte balance is disturbed which ultimately leads to uremia or azotemia. This is not a specific disorder and is connected with a variety of medical disorders such as diabetes, hypertension, and anemia. Kidney disease are deadly and are considered as the 16th largest cause globally for deaths. Therefore, its early detection is necessary. In this review paper, a brief knowledge about the kidney diseases and other diseases related to it such as kidney stones, polycystic kidney stone (PKD), urinary tract infections etc. is presented. Moreover, different datasets which include AASK (African American Study of Kidney Disease and Hypertension), Atlanta Pediatric Clinics, CRISP (The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease), NiCK Study (Neurocognitive Assessment and Magnetic Resonance Imaging Analysis of Children and Young Adults with Chronic Kidney Disease Study), NKDEP (National Kidney Disease Education Program) datasets are reviewed. Majority of the authors worked on classification and segmentation of diseases. Recently, the authors in [34], utilized the ANN and multi-kernel k-means clustering algorithm for classifying the detecting the CKD disease at early stages. In addition to this, the process of identifying the kidney diseases is also given in this paper. Finally, the literature survey is conducted in which different techniques that were proposed by various researchers to detect CKD effectively are discussed.

KEYWORDS— Ultrasound image; speckle noise; Image restoration; Segmentation; Kidney diagnosis; feature extraction.

I. INTRODUCTION

Kidney disorders have recently emerged as one of the most frequent disorders, and their prevalence is constantly increasing over the world. The kidney is a body organ that helps to regulate different levels such as salts, potassium, and pH inside the body [1]. The kidney functions as a body filter that helps to remove waste material, extra water from the body and with the help of the urine process it also detoxify blood that gets deposited inside the bladder [2].In addition the kidneys play a vital role in producing hormones that regulate RBC count and blood pressure. Moreover, from 2007-2012 several new patients with kidney failure are recorded in growing countries such as Indonesia [3]. Kidney disease is a disorder in which the kidneys stop working at the required level as a result of damaged conditions. Damage can be caused by a variety of factors such as high blood pressure, diabetes, or other illnesses. Other disorders in the body arise because of kidney disease [4]. Kidney disease is further divided into five categories.

A. Chronic Kidney Diseases (CKD)

CKD is a kind of kidney disease in which the kidney function gradually deteriorates over months or years. If CKD gets worse, wastes can build up in the bloodstream causing problems such as anaemia, high blood pressure, bone thinning, nerve damage, and poor nutritional status. Kidney disease can also cause blood vessel and heart problems [5].

B. Kidney stones

The kidney stone is caused due to the development of solid mass caused by the crystallization of minerals and substance of the blood. Renal calculi are another name for kidney stone and they can be developed at any point along the urinary tract. The kidney stones have a crystalline and sharp structure that can grow as large as a ping-pong ball [6].

C. Glomerulonephritis

Glomerulonephritis is a swollen Glomeruli disorder that is responsible for blood filtration within the kidney. It is also termed nephritis and can become severe that require efficient and timely action and treatment. The Glomerulonephritis symptoms can be severe and might last for a long time, it is also known as Bright's disease [7].

D. Polycystic kidney disease (PKD)

PKD is a hereditary condition that is developed by a genetic issue. It also builds cysts within the kidneys. These cysts enlarge the kidneys beyond their normal size and harm the tissue that makes up the kidneys. CKD is caused by PKD and can cause diseases like ESRD (End-Stage Renal Disease) or kidney failure [8].

E. Urinary tract infections (UTI)

UTI is a bacterial illness that can damage any region of the urinary tract. The bladder and urethra are considered the most commonly infected areas. These issues have solutions and do not create other health issues inside the body. The UTI can be categorized into two type's complicated and uncomplicated (patients with normal urinary tracts from the perspective of a functional and structural) [9].

Early detection of these disorders can help to slow the disease's progression and prevent complications or tough scenarios from affecting the patient's health [10]. To diagnose these disorders, the doctors conduct various tests like glomerular filtration rate (girl), kidney biopsy, urine test, blood creatinine test, and ultrasound or computed tomography (ct) scan [11]. Ultrasound is widely accepted because it is a portable, non-invasive, and generally inexpensive method for examining the human body. Furthermore, speckle noise that affects all coherent imaging systems, has a negative impact on

ultrasonographic image quality [12]. Ultrasound can also provide real-time images which can be useful in a variety of procedures. The three main forms of ultrasounds are Amplitude, Brightness, and Motion modulation that are commonly regarded as A, B, and M modulation [13]. An accurate split of the image is needed in ultrasound to detect specific items. Segmentation is a processing method for dividing an area or subject [14]. Image segmentation is utilized in a variety of applications such as processing medical images, image retrieval, processing remote sensing images, and face recognition. During the process of image segmentation, the images are separated into multiple related regions and each pixel of an image is analysed [15]. To remove the images, a variety of texture characteristics are accessible such as texture features, wavelet features, statistical features, GLCM features, region-based features, etc [16]. Following are some generic datasets present for kidney disease detection

I. DIFFERENT TYPES OF DATASETS

A. AASK (African American Study of Kidney Disease and Hypertension).

The AASK is considered double-blinded, randomized, and controlled research among non-diabetic African Americans that suffer from a hypertensive renal illness that was caused due to high chronic blood pressure. The people who participated in this research were randomly allocated to one of the two targeted blood pressure levels and also accept one of the three antihypertensive drugs to analyse certain antihypertensive drugs and see how the level of blood pressure control affects the primary results of hypertensive kidney disease progression as indicated by GFR change. The research's trial phase lasted from 1995 to 2001, followed by a cohort follow-up period from 2002 to 2007. APOL1 markers were genotyped in patients who supplied signed permission for collecting DNA during the experiment.

B. Atlanta Pediatric Clinics

This cross-sectional cohort research was conducted in Atlanta, on a group of children ranging between 8-17 years who lived in Atlanta's metropolitan areas. The children with CKD were chosen from Atlanta's outpatient clinics and the hemodialysis unit at Children's Health Care, while control participants were recruited from two pediatric practices in Atlanta. The CKD group included those who were on dialysis, had a working transplant or had a minimal eGFR of 60 mL/min/1.73 m2.

C. CRISP (The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease).

CRISP is ten-year prospective cohort research that began in March 1999 intending to develop unique imaging schemes by utilizing magnetic resonance imaging (MR imaging) to consistently and correctly assess cyst and renal volume in people with autosomal dominant polycystic kidney disease (ADPKD) in their disease's course.

D. NiCK Study (Neurocognitive Assessment and Magnetic Resonance Imaging Analysis of Children and Young Adults with Chronic Kidney Disease Study).

Nick study is a cross-sectional study that was conducted on a group of children and young adults ranging from 8-25 years. The purpose of this research was to uncover any neurologic anomalies in the population of CKD by comparing the neurological function of children and young adults with CKD to those without CKD. The brain's MRI, Neurocognitive testing, clinical phenotyping, blood flow, and functional connectivity are all part of the study's evaluation. If participants had received kidney transplant or were on dialysis and had egress under 90 mL/min/1.73 m2 for at least 6 months were chosen for the group of CKDS. The participants of this research spoke English as their first language. The participants of the control group were drawn from paediatric practices across the Children's Hospital of Philadelphia system.

E. NKDEP (National Kidney Disease Education Program).

The NKDEP aimed to decrease the mortality and morbidity associated with kidney disease and its consequences. NKDEP aimed to improve early diagnosis of CKD by allowing individuals identification at the highest risk of kidney failure; promoted scientifically proven therapies to reduce the course of kidney disease, and improved the synchronization of Federal reactions to CKD by educating people of the seriousness of kidney failure and the availability of tests, the significance of testing people who are at high risk, and the treatment options available to stop or slow renal failure. Participants from voluntary and professional groups participating in kidney disease programs make up the NKDEP's Coordinating Panel. The "African-Americans Unaware of High Kidney Disease Risk," was published in 2004 on 8th March.

II. PROCESS FOR DETECTING KIDNEY DISEASE

In the process of detecting kidney disease, image acquisition is considered the first step. The Data is first gathered from multiple data sets acquired from CT scans, ultrasound, MRI, X-rays, and other sources [17]. The photos are then pre-processed to improve their quality and eliminate noise. The main goal of the pre-processing procedure is to eliminate extraneous data, remove noise from images, and other factors that may aid in removing information or characteristics which may lead to incorrect segmentation and classification results. The image segmentation step seeks to assess the kidney portion by estimating ROI (Region of Interest) and looking at the anatomical composition of various bodily components. Following the segmentation of an image, features are taken from the segmented region and forwarded to the feature selection step. Feature selection is not required for all systems; nonetheless, the fundamental contribution of this step is to choose the features that are more patterned and can increase classification accuracy. Finally, there is the classifier training and classification step, which involves making decisions about disease class. Many detection characteristics of

kidney disease enhances the decision-making process. Several automatic algorithms for segmenting ultrasound photos have been developed. Artificial Neural Network (ANN) acts as a classifier to detect the impacts of impact of feature selection and pre-processing on basis of classification accuracy and other metrics [18]. To segment ultrasound images several techniques like watershed segmentation, clustering approaches, and k-means clustering are utilized. Some of these approaches are described in the next section [19].

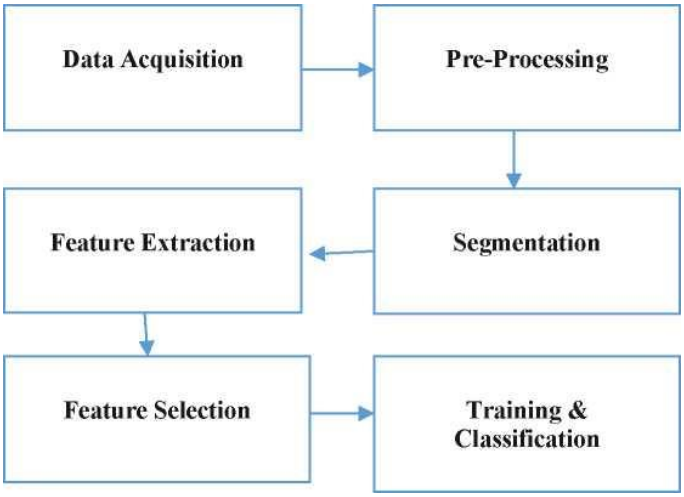


Fig. 1 Illustrates the general block diagram for detecting kidney disease

III. LITERATURE SURVEY

A variety of studies have looked at how ultrasound images can be utilized to identify and segment kidney disease. Among them, some of the works are examined in this paper:

A. *B. V. Ravindra et al. [20],* the authors of this paper used ANN to classify CKD and NCKD. Potassium, sodium, urea, and creatinine were used by authors to determine whether or not the patient had CDK. Datasets comprising a total of n=230 were acquired from a nearby general hospital. For classification authors used BPNN (Back Propagation Neural Network) model, whereas the BPNN classifier's efficiency was assessed by utilizing specificity, classification accuracy, and sensitivity. Clustering was used to determine the valid attribute values in the Datasets at first. The simulation findings suggested that the total classification accuracy for distinguishing CKD from NCKD is 95.3 %.

B. *Chakrapani et al. [21],* by utilizing an ANN with Back Propagation Network (BPN), researchers proposed a novel method to detect CKD. In addition, researchers trained the NN classifier and tested its detection ability on a different dataset. The ANN recognition accuracy was greatly promising when compared to many other types of classifiers such as Classification and Regression tree, K-NN, SVM.

C. *K. Lakhwani et al. [22],* the authors of this paper suggested a method that automatically diagnosis diabetic data by utilizing ANN and datasets of Pima Indians Diabetes. A logistic-activation function is employed for neuron activation in the suggested model, whereas for training algorithms, the Quasi-Newton technique is employed. As a consequence of the cumulative gain plot, the greatest gain score is utilized as a measure of the model's quality.

D. *S. Packirisamy Balamurugan et al. [23],* the researchers of this paper utilized ANN to develop a unique ultrasound kidney disease prediction. The specificity, accuracy, and sensitivity were used to analyze the suggested method's efficiency. The experimental findings demonstrate that the suggested system achieves the highest accuracy of 95.83 % when compared to other systems.

E. *Vinayagam.P et al. [24],* the authors in this paper suggested technology by utilizing DWT (Discrete Wavelet Transform) to pre-process nephrolithiasis in an MRI picture. GLCM is used to retrieve major characteristics. The BPNN (Back Propagation Method of Neural Network) was utilized for classifying a dataset of 20 test data comprising abnormal and normal MRI images of kidney MRI pictures. The BPN classification result was presented on an LCD board that was connected to an Arduino no board. For successful segmentation of kidney stones, the Fuzzy Clustering Mean Algorithm (FCM) is applied.

F. *H. Zhang et al. [25],* to propose a classification model, the authors of this paper focused on pre-processing, feature discovery, and the selection phase of kidneys' ultrasound images. For image pre-processing, researchers used four processes i.e. cropping, rotation, interpolation, and background removal in the proposed methodology to increase the quality of the image and make diagnosis simple and efficient. Researchers utilized GLCM to generate several second-order statistical texture characteristics such as homogeneity, entropy, contrast, dissimilarity, contrast energy, and correlation. At last, by using PCA (Principle of Component Analysis), the obtained characteristics were reduced to the ideal subsets. When the classified image was applied to the ANN, the outcome showed that GLCM was in combination with PCA for feature reduction gave the classification accuracy at a high rate.

G. *P. R. Tabrizi et al. [26],* to construct a unique automated collecting system segmentation approach, the authors of this paper utilized a 3D U-net deep neural network. A 3D ultrasound pictures dataset from 64 cases of hydronephrotic was used to assess the proposed method's performance that revealed 1.29 ± 0.95 mm as an average symmetric surface distance, an average value of HIerror 2.1 ± 2.8 %, and 0.76 ± 0.12 as an average dice similarity coefficient.

H. *Chen et al. [14],* to diagnose kidney disease effectively and efficiently at an early stage, the researchers in this paper proposed AHDCNN (Adaptive Hybridized Deep Convolutional Neural Network). The effectiveness of the classification technique was determined by the data set's function. An algorithm model based on CNN has indeed been designed to by researchers improve the classification system's efficiency by decreasing the dimension of the features. These high-level characteristics aid in the development of a monitored tissue classifier that distinguishes between two tissue types. Through predictive analytics, the experimental phase on the Internet of medical things platform (IoMT) suggested that developments in machine learning offer a successful basis for the identification of intelligent solutions to demonstrate their predictive capabilities far beyond the area of kidney disease.

I. *FuatTurk et al. [27],* the authors of this paper developed a modified hybrid model that combined the best attributes of standard V-Net models. The suggested model outperformed standard imaging models in segmentation, whereas the proposed model has versatile applicability and structure so it can easily incorporate into any system.Fortumor and kidney segmentation, the hybrid V-Net model had average Dice coefficients of 86.5 %and 97.7%, respectively, and can thus be utilized as a realistic option for segmenting soft tissue organs.

J. *M. Akshaya et al. [28],* to detect kidney stones, the author in this paper used BPN. The authors divided the decisionmaking process into two stages i.e. extraction of features and classification of images. To extract features, the authors used the Principal component, and to classify the image authors utilized BPN. The authors used the FCM clustering algorithm to address the segmentation approach. The BPN classifier's efficiency was measured in terms of classification accuracy and training execution. When contrasted to standard neuralnetwork-based approaches, BPN provides precise classification.

K. *Janani Jand Sathyaraj R[29],* proposed a trustworthy machine learning technique that can accurately predictthe onset of CKD. For this, the authors used the data set is available in UCI ML library, however, it somehow it contains a large number of null data that are eliminated by the using KNN Imputation.

L. *H. Alasker et al. [30],* analyzed the performance of various data mining classifiers which include; naive Bayes, Decision table, Back propagation neural network, KNN, decision trees and one rule classifiers. These classifiers were implemented on the datasets which include information about the kidney disease and on the basis of this information the classifiers detected the disease.

M. *Rashed-Al-Mahfuz et al.* [31], suggested a kidney disease detection method that utilized the selective key pathological categories for detecting the clinical test parameters that help to improve the accuracy of CKD detection. The proposed method was beneficial as it could save a lot of time and money during diagnostic screening.

N. *Elhoseny et al. [32],* the authors of this paper developed a new and unique method for detecting the CKD in humans that was based on the Density Feature selection (DFS) and Ant colony optimization (ACO), as D-ACO. The proposed D-ACO method proven out to be effectively removing the redundant data by using DFS before ACO classifies it.

O. *Khaled Mohamad Al-Mustafa. [33],* different classifiers like KNN, decision table, decision tree, SGD (stochastic gradient descent), j48 and Naive Bayes were implemented to suggest a new model for detecting the CKD in humans on the basis of the selected features.

TABLE 1
LITERATURE COMPARISON FOR DIFFERENT KIDNEY DISEASE DETECTION SYSTEMS

YOP	Authors Name	Work done	Strength	Weakness
2018	B. V. Ravindra et al. [20]	The scholars used ANN to classify CKD and NCKD and datasets of Pima Indians Diabetes.	ANN can produce results even with insufficient knowledge	Hardware dependent and unknown duration of network.
2019	Chakrapani et al. [21]	By utilizing an ANN with BPN, the authors proposed a novel method to detect CKD	Faster in identifying disease.	Sensitive to noise and depends on input data.
2020	K. Lakhwani et al. [22]	The researchers suggested a method that automatically diagnosis diabetic data by utilizing ANN	Distributed storage and trains machine quickly.	It is difficult to detect the problems in ANN network

2020	S. Packirisamy Balamurugan et al. [23]	Utilized ANN to develop a unique ultrasound kidney disease prediction.	Fault tolerant and effective	It is difficult to detect the problems in ANN network
2019	Vinayagam.P et al. [24]	The authors suggested technology by utilizing DWT to pre-process nephrolithiasis in an MRI picture	Best compression ratio	Complex computations
2019	P. R. Tabrizi et al. [26]	To construct a unique automated collecting system segmentation approach, the authors of this paper utilized a 3D U-net deep neural network.	Efficient and effective in detecting CKD.	It needs vast amount of data for training.
2020	G. Chen et al. [14]	To diagnose kidney disease effectively and efficiently at an early stage, the researchers in this paper proposed AHDCNN.	Selects only important features for detecting the CKD with high accuracy.	It requires large datasets for training.
2020	FuatTurk et al. [27]	The authors of this paper developed a modified hybrid model that combined the best attributes of standard V-Net models.	-	-
2020	M. Akshaya et al. [28]	To detect kidney stones, the author in this paper used BPN.	Fasterin recognition	Sensitivity towards noisy data
2021	Janani J and Sathyaraj R [29],	proposed a trustworthy machine learning technique that can accurately predict the onset of CKD.	Fast and efficient method for detecting CKD	Error possibility is high.
2017	H. Alasker et al. [30],	Implemented various data mining classifiers on datasets for detecting the diseases.	Easy to understand and accurate.	Costly and time consuming
2021	M. Rashed-Al-Mahfuz et al. [31],	Suggested kidney disease detection method that utilized selective key pathological categories to improve the accuracy of CKD detection	-	-
2019	Elhoseny et al. [32],	developed a method for detecting the CKD, that was based on Density Feature selection (DFS) and Ant colony optimization (ACO), namely as D-ACO.	It can detect noisy data and provide accurate results and decreases complexity of the model	Slow convergence speed.
2021	Khaled Mohamad Al-Mustafa. [33],	Implemented various classifiers and suggested a new model for detecting CKD on the basis of the selected features.	Easy to understand and accurate.	Costly and time consuming

IV. CONCLUSIONS

CKD or chronic kidney diseases is considered as one of deadliest disease which is caused by the disorder in renal. Therefore, identification and treatment of this disease at early stage is crucial in order to avoid human loss. in this review article, various methods and techniques that were proposed by experts in the past are reviewed and after reviewng it is observed that most of the authors used the deep neural networks like ann, cnn, bpn etc.

In their work to detect the kidney diseases at early stages . It was also analyzed that these methods were providing efficient results that enhanced their performance. Furthermore, it is also analyzed that very few works hase been done on pre-processing, feature extraction and classification modules which are the key factors in detecting the kidney diseases. In addition to this, utilization of dataset also plays a crucial role in detecting CKD and very few works have been done in this case. Therefore, it is observed that the existing approaches are providing good results but there is still a scope of improvement in these models in order to make them more accurate and convenient.

REFERENCES

[1] Dirks, J., Remuzzi, G., Horton, S., Schieppati, A,. *"Disease Control Priorities in Developing Countries",* New York: Oxford University Press. pp. 695-706; 2006.

[2] Maurya, R. Wable, R. Shinde, S. John, R. Jadhav and R. Dakshayani, *"(Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), pp. 1-4, 2019.*

[3] M. Ahmad, V. Tundjungsari, D. Widianti, P. Amalia and U. A. Rachmawati, *"Diagnostic decision support system of chronic kidney disease using support vector machine,"* 2017 Second International Conference on Informatics and Computing (ICIC), pp. 1-4, 2017.

[4] Ngo, L. Y., Meng, O. L., Leong, G. B., and Guat, L. D. "All *renal replacement therapy in Malaysia".* The 19th Report of the Malaysian Dialysis and Transplant Registry, 2011.

[5] Maurya, R. Wable, R. Shinde, S. John, R. Jadhav and R. Dakshayani, "*Chronic Kidney Disease Prediction and Recommendation of Suitable Diet Plan by using Machine Learning,"* 2019 International Conference on Nascent Technologies in Engineering (ICNTE), pp. 1-4, 2019.

[6] T. Shah and S. Kadge, *"Analysis and Identification of Renal Calculi in Computed Tomography Images*," 2019 International Conference on Nascent Technologies in Engineering (ICNTE), pp. 1-4, 2019.

[7] X. Yao, X. Wang, Y. Karaca, J. Xie and S. Wang, *"Glomerulus Classification via an Improved GoogLeNet*," in IEEE Access, vol. 8, pp. 176916-176923, 2020.

[8] American Kidney Fund Fighting On All Fronts, Polycystic kidney disease, updated on October 16, 2020, https://www.kidneyfund.org/kidney- disease/other-kidney-conditions/polycystic-kidney-disease.html.

[9] E. I. Papageorgiou, C. Papadimitriou and S. Karkanis, *"Management of uncomplicated urinary tract infections using fuzzy cognitive maps,"* 2009 9th International Conference on Information Technology and Applications in Biomedicine, pp. 1-4, 2009.

[10] Kaur, Nikita, H. Sadawarti and J. Singla, *"A Comprehensive Review of Medical Expert Systems for Diagnosis of Chronic Kidney Diseases*," 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1008-1013, 2019.

[11] Frank O'Brien, "Overview of Kidney Filtering Disorders", Reviewed on Mar 2020, https://www.msdmanuals.com/en-in/home/kidney-and-urinary-tract-disorders/kidney-filtering-disorders/overview-of-kidney-filtering-disorders.

[12] H. Wang, C. Wu, J. Chi, X. Yu and Q. Hu, *"Speckle Noise Removal in Ultrasound Images with Stationary Wavelet Transform and Canny Operator,"* 2019 Chinese Control Conference (CCC), 2019, pp. 7822-7827, 2020.

[13] N. A. Zulkiflli et al., *"Ultrasound Tomography Hardware System for Multiphase Flow Imaging,"* 2019 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), pp. 264-268, 2019.

[14] G. Chen et al., *"Prediction of Chronic Kidney Disease Using Adaptive Hybridized Deep Convolutional Neural Network on the Internet of Medical Things Platform*," in IEEE Access, vol. 8, pp. 100497-100508, 2020.

[15] Liao, H. Lu, X. Xu and Q. Gao, *"Image Segmentation Based on Deep Learning Features,"* 2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI), pp. 296-301, 2019.

[16] Bagri, Neelima & Johari, Punit, "A *Comparative Study on Feature Extraction using Texture and Shape for Content Based Image Retrieval,"* International Journal of Advanced Science and Technology, pp. 41-52, 2015.

[17] Shengfeng Liu, Yi Wang, Xin Yang, Baiying Lei, Li Liu, Shawn Xiang Li, Dong Ni, Tianfu Wang, " *Deep Learning in Medical Ultrasound Analysis: A Review,"* Engineering, vol. 5, pp. 261-275, 2019.

[18] Rahman, Akizur & Muniyandi, Ravie, *"An Enhancement in Cancer Classification Accuracy Using a Two-Step Feature Selection Method Based on Artificial Neural Networks with 15 Neurons,"* Symmetry, 2020.

[19] Nithya, A. & Appathurai, Ahilan & Venkatadri, N. & Ramji, D. & Palagan, C., *"Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images,"* Measurement, 2019.

[20] RAVINDRA, N. SRIRAAM AND M. GEETHA, *"CHRONIC KIDNEY DISEASE DETECTION USING BACK PROPAGATION NEURAL NETWORK CLASSIFIER,"* 2018 INTERNATIONAL CONFERENCE ON COMMUNICATION, COMPUTING AND INTERNET OF THINGS (IC3IOT), PP. 65-68, 2018.

[21] Chakrapani et al., *"Detection of Chronic Kidney Disease Using Artificial Neural Network,"* International Journal of Applied Engineering Research, vol. 14, 2019.

[22] K. Lakhwani, S. Bhargava, K. K. Hiran, M. M. Bundele and D. Somwanshi, "*Prediction of the Onset of Diabetes Using Artificial Neural Network and Pima Indians Diabetes Dataset*," 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), pp. 1-6, 2020.

[23] S. Packirisamy Balamurugan et al., "*A novel method for predicting kidney diseases using optimal artificial neural network in ultrasound images,"* Int. J. Intelligent Enterprise, vol. 7, 2020.

[24] Vinayagam.P, Sreemathi.M, Jeevitha K, Sandhya S, *"KIDNEY STONE DETECTION USING NEURAL NETWORK,* " International Journal of Applied Engineering Research, vol.14,2019.

[25] H. Zhang, C. Hung, W. C. Chu, P. Chiu and C. Y. Tang, *"Chronic Kidney Disease Survival Prediction with Artificial Neural Networks,"* 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1351-1356, 2018.

[26] P. R. Tabrizi et al., *"Automatic Segmentation of The Renal Collecting System in 3D Pediatric Ultrasound to Assess the Severity of Hydronephrosis,"* 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 1717-1720, 2019.

[27] Fuat Turk et al. *"Kidney and Renal Tumor Segmentation Using a Hybrid V-Net-BasedModel,"* mathematics, pp.1-17, 2020.

[28] M. Akshaya, R. Nithushaa, N. S. M. Raja and S. Padmapriya, *"Kidney Stone Detection Using Neural Networks*," 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), pp. 1-4, 2020.

[29] Janani J a ,SathyarajR, *"Diagnosing Chronic Kidney Disease Using Hybrid Machine Learning Techniques',* Turkish Journal of Computer and Mathematics Education Vol.12 No.13 (2021), 6383 - 6390.

[30] H. Alasker, S. Alharkan, W. Alharkan, A. Zaki and L. S. Riza, "*Detection of kidney disease using various intelligent classifiers*," 2017 3rd International Conference on Science in Information Technology (ICSITech), 2017, pp. 681-684.

[31] M. Rashed-Al-Mahfuz, A. Haque, A. Azad, S. A. Alyami, J. M. W. Quinn and M. A. Moni, "*Clinically Applicable Machine Learning Approaches to Identify Attributes of Chronic Kidney Disease (CKD) for Use in Low-Cost Diagnostic Screening*," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 9, pp. 1-11, 2021, Art no. 4900511, doi: 10.1109/JTEHM.2021.3073629.

[32] Elhoseny, M., Shankar, K. &Uthayakumar, J. "*Intelligent Diagnostic Prediction and Classification System for Chronic Kidney Disease*" Sci Rep **9,** 9583 (2019). https://doi.org/10.1038/s41598-019-46074-2.

[33] Khaled Mohamad Almustafa,Prediction of chronic kidney disease using different classification algorithms,Informatics in Medicine Unlocked,Volume 24,2021.

[34] A.Nithya, AhilanAppathurai, N. Venkatadri, D.R. Ramji, C. Anna Palagan, "*Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound* images",Measurement,Volume149,2020.

A COMPLETE MOBILE BASED GURMUKHI OCR SYSTEM

Ravneet Kaur[1], Dharam Veer Sharma[b]
[a]Research Scholar, Department of Computer Science, Punjabi University, Patiala, India [b]Professor, Department of Computer Science, Punjabi University, Patiala, India. *"mail2ravneetkaur@smail.com, -dveer72@ftotmail.com*

Abstract—— OCR is technology in the field of computer through which textual information contained in the image is converted into editable text format. This technology is used to convert scanned documents, PDF files, or images into an editable soft copy that can be searched, reproduced, and transported with ease. In comparison to desktop OCR systems a number of additional challenges are to be considered while implementing mobile based system. Conventional desktop OCR systems assume input data to be high resolution images obtained from flatbed scanners. Whereas, mobile phone cameras produce lower fidelity images that suffer from defects such as noise, uneven lighting or perspective projections. In this paper a complete mobile based OCR system is presented.

KEYWORDS:—OCR, Pre Processing, Layout Analysis, Tesseract, Post Processing, Mobile OCR.

1. INTRODUCTION

Optical Character Recognition (OCR) technology in the field of computer is used to convert scanned documents, PDF files, or images into editable soft copies. Earlier the approach to character recognition was restricted to desktop scanner. The usability of such system was limited as they are non-portable because of large size. With the advent of technology and portable computer devices such as PDA, mobile phone, iPhone etc. new trends of research have emerged, where mobile phones are the most commonly used electronic device, eliminating the need for bulky devices like scanners, desktops and laptops. Mobile based OCR system allow user to process text segments anywhere anytime.

The analysis of documents captured with mobile devices is in demand as they can capture images of any kind of document including very thick books, historical pages too fragile to touch, and text in scenes. Though mobile OCR packages are available in market with very high degree of accuracy in recognizing printed text in European and other scripts, but for Indian scripts only few applications are available with fair accuracy, and particularly for Gurmukhi very few commercial mobile applications are available. Some document processing surveys include [1], [2], [4] and [6]. In particular, attempts have also been made for recognizing Punjabi text in scanned images [3-5]. Doermann et al. [6] presented a survey of application domains, technical challenges and solutions for recognizing documents captured by digital cameras. Attempts have been made for developing mobile OCR applications. Kaur et al. [7-8] presented a state-of-the-art survey and review of available mobile based OCR systems.

In this paper a complete system architecture for mobile based Gurmukhi OCR system is presented. Proposed system enables users to convert different document images captured from newspapers, reports, books, journals, magazines, government files or letters into machine editable soft copies.

2. GURMUKHI SCRIPT CHARACTER SET

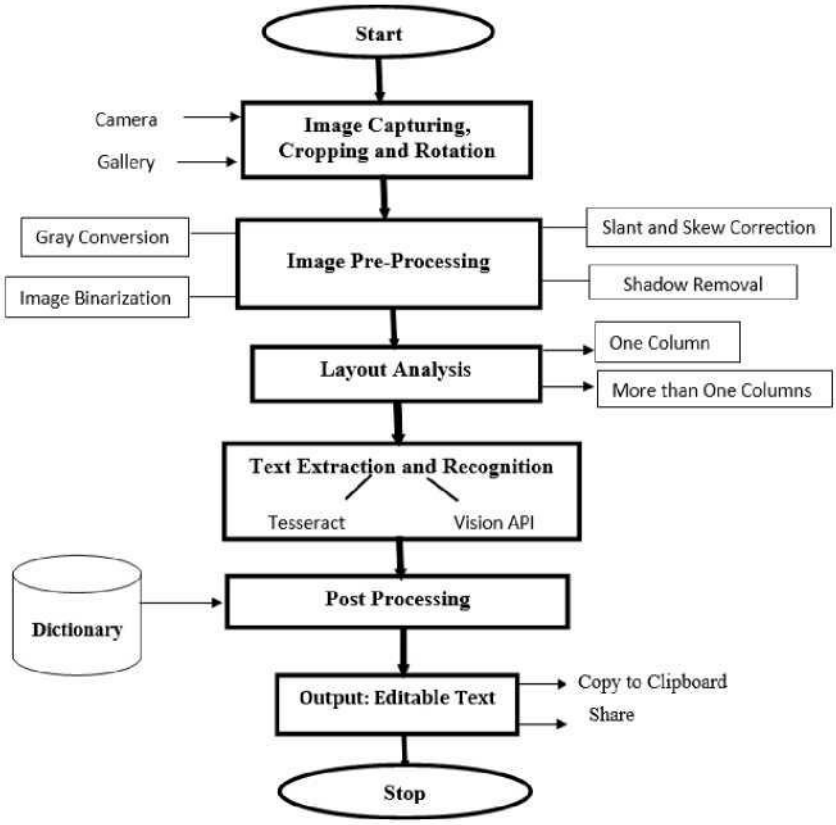
Punjabi language is one of the popular languages of India. Script used to write Punjabi language is Gurmukhi. Among all world languages Punjabi is the 10th most spoken language. The Unicode character set of Punjabi language is shown in Table 1. The alphabet consists of 35 consonants, 06 additional consonants, 09 independent vowels, 12 dependent vowels, and 3 half characters. Various properties of Gurmukhi script are discussed by Lehal [9].

Table 1: Character Set of Gurmukhi Script

Consonants	ੳ ਅ ੲ ਸ ਹ ਕ ਖ ਗ ਘ ਙ ਜ ਝ ਞ ਟ ਠ ਡ ਢ ਣ ਤ ਥ ਦ ਧ ਨ ਪ ਫ ਬ ਭ ਮ ਯ ਰ ਲ ਵ ੜ
Additional consonants	ਸ਼ ਖ਼ ਫ਼ ਗ਼ ਜ਼ ਲ਼
Independent vowels	ਓ ਉ ਊ ਆ ਐ ਔ ਇ ਈ ਏ
Gurmukhi numerals	0 1 2 3 4 5 6 9 7 8 9
Dependent vowel signs	ਾ ਿ ੀ ੁ ੂ ੇ ੈ ੋ ੌ ਂ ੱ ੍
Other signs	ੑ ਃ ੰ ੱ

3. System Architecture

The System perform following steps to convert document images into editable data (Figure 1).



Fitgure 1: System Architecture

4. IMAGE PRE-PROCESSING AND LAYOUT ANALYSIS

4.1 Image Pre-Processing of Input Image

The quality of the input image/document plays vital role in character recognition. Better quality of the source image significantly improves the accuracy of OCR. Mobile phone cameras produce lower fidelity images that suffer from defects such as noise, uneven lighting, poor focus and skewness. Therefore, in case of mobile based OCR systems pre-processing plays a major role in optimizing the image for character recognition. Various challenges faced by smartphone devices OCR are discussed by [10]. The pre-processing includes image gray conversion, binarization, shadow removal, skew correction and removal of noise. A Complete algorithm for preprocessing of Image is discussed by Kaur et al. [11].

4.2 Layout Analysis of Images

The goal of layout analysis is to extract the area of interest from the image. Layout analysis or page segmentation separate the document image into regions such as images, text, tables and drawings. In optical character recognition the region of interest is text region. At this stage the input image is partitioned into several regions, usually rectangular, each of which is referred to as a block. Input document image and its structural layout is shown in Figure 2.

<table>
<tr><td colspan="2">Heading</td></tr>
<tr><td rowspan="2">Text Region</td><td>Text Region</td></tr>
<tr><td>Image</td></tr>
</table>

(a) (b)

Figure 2: (a) Original Document Image. (b) Structural Layout of Document

A top-down image segmentation technique is presented to segment document image into sub parts such as tables, text blocks, and pictures. Run length smearing algorithm (RSLA algorithm) and recursive top-down decomposition techniques are used for layout analysis. The threshold values in the proposed technique are calculated based upon the geometric structure of image using histograms. Horizontal and vertical projection profiles are used in recursive image segmentation technique. The proposed technique recursively decomposes document image into hierarchy of similar regions.

5. TEXT EXTRACTION AND RECOGNITION

Text extraction and recognition is the important step in any optical character recognition system. For text extraction and recognition tesseract and Cloud vision API's are used. The vision OCR API is paid and is faster than tesseract in term of time and accuracy, while tesseract is completely free and an open-source library written in C++ [12]. A detailed comparison of vision API and tesseract OCR engine is presented by Kaur et al. [13]. An Android OCR application for Punjabi language is developed by integrating tesseract. Further, tesseract engine is powered by an image processing library named Leptonica. Tesseract together with leptonica is broadly useful for image processing and image analysis applications [14-16].

6. POST PROCESSING

Post Processor is integral part of any Optical Character Recognition (OCR) as it is applied to the recognition results of OCR System. The objective of post processing is to correct the errors in OCR result and improve the recognition rate. Post-processing plays a major role in optimizing the recognition results by correcting the errors. In this research a dictionary and minimum edit distance based post-processing system is presented to improve the accuracy of mobile based Gurmukhi OCR system. The idea for post processor is derived from algorithms used by Lehal, Zhuang and Sharma [18-19]. The post processor system consists of following steps to improve the recognition rate by correcting the spelling errors in the OCR output.

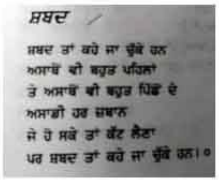
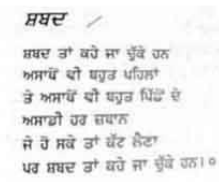
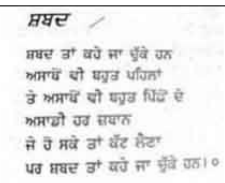
- Use of Punjabi corpus for statistical analysis of Punjabi language and checking the spelling of words.
- Design of consonant subset of visual similar characters.
- Dictionary using AVL tree, hash map, and array list.
- Result validation using minimum edit distance, word frequency list and Punjabi grammar rules.

In the post processing algorithm semantic lexicon and statistical language model are combined. Candidate distance information is used to reduce the size of the search space. For the implementation of post processor Punjabi corpus with list of words is required. The accuracy of post processor depends upon divergent words in corpus and size of corpus.

7. RESULT AND DISCUSSION

The OCR results, result after pre-processing and results after post processing are presented in this section. The recognition accuracy is improved after performing image pre-processing steps. In pre-processing, work is presented on increasing the image brightness without overflow. It results in cleaning the background of the image and removal of shadow from the image. The results before and after pre-processing are shown in Table 2

Table 2 Results before and after pre processing Image Number Image Results

Sample Image 2 (34 words)	Before Pre-Processing (accuracy 58.82 %)		ਸ਼ਅ਼ਬਦ ਆ ਜ ਸ਼ਬਦ ਤਾਂ ਕਹੇ ਜਾ ਚੁੱਕੇ ਅਸਾਥੋਂ ਵੀ ਬਹੁਤ ! ਤੇ ਅਸਾਥੋਂ ਵੀ ਬ ਅਸਾਡੀ ਹਰ ਜੇ ਹੋ ਸਕੇ ਤਾਂ ਪਰ ਸ਼ਬਦ ਤਾਂ (Incorrect 14 words)
	After Pre-Processing (accuracy 91.18 %)		ਲਦ ," ਸ਼ਬਦ ਤਾਂ ਕਹੇ ਜਾ ਚੁੱਕੇ ਹਨ ਅਸਾਥੋਂ ਵੀ ਬਹੁਤ ਪਹਿਲਾਂ ਤੇ ਅਸਾਥੋਂ ਵੀ ਬਹੁਤ ਪਿੱਛੋਂ ਦੇ ਅਸਾਡੀ ਹਰ ਜ਼ਅ਼ਬਾਨ ਜੇ ਹੋ ਸਕੇ ਤਾਂ ਕੱਟ ਲੈਣਾ ਪਰ ਸ਼ਬਦ ਤਾਂ ਕਹੇ ਜਾਂ ਚੁੱਕੇ ਹਨ ; (Incorrect 03 words)
	After Post-Processing (accuracy 97.057 %)		ਲਦ ," ਸ਼ਬਦ ਤਾਂ ਕਹੇ ਜਾ ਚੁੱਕੇ ਹਨ ਅਸਾਥੋਂ ਵੀ ਬਹੁਤ ਪਹਿਲਾਂ ਤੇ ਅਸਾਥੋਂ ਵੀ ਬਹੁਤ ਪਿੱਛੋਂ ਦੇ ਅਸਾਡੀ ਹਰ ਜ਼ੁਬਾਨ ਜੇ ਹੋ ਸਕੇ ਤਾਂ ਕੱਟ ਲੈਣਾ ਪਰ ਸ਼ਬਦ ਤਾਂ ਕਹੇ ਜਾਂ ਚੁੱਕੇ ਹਨ ; (Incorrect 01 words)

The post processor algorithm applied on the output of 40 camera captured documents with 2488 words. With pre-processing of image and tesseract training process an improvement of 25.13% from 66.67% to 91.80% is obtained in the tesseract output. The recognition accuracy of the OCR without post processing was 91.80% which is increased to 95.44% by applying post processing to the recognised text. Figure 3 highlight results before and after post processing.

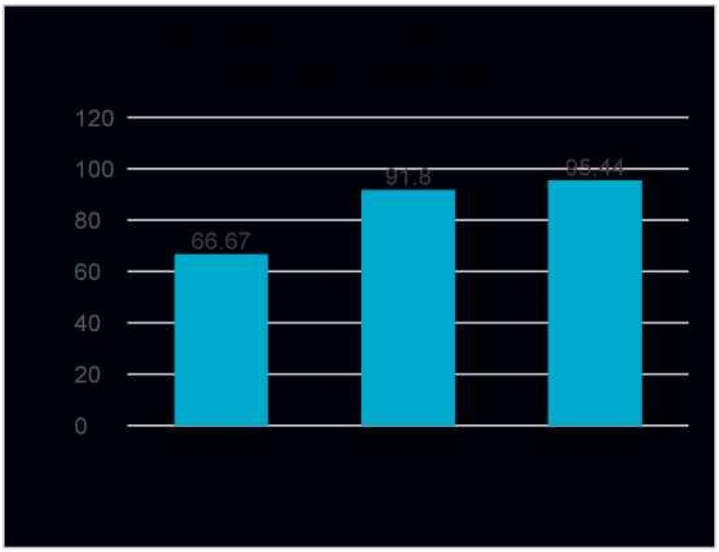


Figure 3: Result analysis pre-processing and post processing algorithms

In our current work, a complete system for text recognition in mobile devices is presented. Algorithms for preprocessing and layout analysis of the text images are discussed. At first, text image is digitized, then preprocessed and afterwards the text is extracted and recognized from the image. The proposed pre-processing steps are implemented with the help of powerful OpenCV library. Furthermore, a post processor system is developed to correct the spelling errors in the OCR output to improve the recognition rate of OCR. With minor modifications the presented Mobile OCR framework can be used for other European and Indian scripts.

References

[1] Y. Y. Tang and C .Y. Suen, “Document Structure: A Survey”, IEEE International Conference on Document Image Analysis and Recognition, pp. 99-102, October 1993.

[2] Y. Y. Tang, S. W. Lee, and C. Y. Suen, "Automatic Document Processing: A Survey" Pattern Recognition, Vol. 29, Issue 12, pp. 1931-1952, December 1996.

[3] G. S. Lehal and C. Singh, "A Gurmukhi Script Recognition System", Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, Vol. 2, pp. 557-560, September 2000.

[4] M. K. Jindal, R. K. Sharma and G. S. Lehal, "A Study of Different Kinds of Degradation in Printed Gurmukhi Script", Proceedings of the International Conference on Computing: Theory and Applications (ICCTA'07), pp. 538-544, 2007.

[5] G. S. Lehal, "A Complete Machine-Printed Gurmukhi OCR System", Advances in Pattern Recognition, Springer, pp. 43-71, 2009.

[6] D. Doermann, J. Liang and H. Li, ''Progress in Camera-Based Document Image Analysis", IEEE International Conference on Document Analysis and Recognition (ICDAR'03), pp. 606-616, 2003.

[7] R. Kaur, D. Sharma, "Mobile Based OCR Systems: State-of-the-art Survey for Indian Scripts" in International Journal of Computer Sciences and Engineering (IJCSE), impact Factor 3.022, ISSN: 2347-2693, Volume 7, Issue 4, pp. 501-505, April - 2019. https://doi.org/10.26438/iicse/v7i4.457461

[8] R. Kaur, "Text Recognition Applications for Mobile Devices", International Journal of Global Research in Computer Science (JGRCS), impact Factor 1.2, ISSN: 2229-371X, Volume 9, Number 4, pp. 20-24, April 2018.

[9] G. S. Lehal G S, "A Complete Machine - Printed Gurmukhi OCR System", Advances in Pattern Recognition, Springer, pp. 43-71, 2009.

[10] N. Sourvanos and G. Tsatiris, "Challenges in Input Preprocessing for Mobile OCR Applications: A Realistic Testing Scenario", IEEE 9th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1-5, 2018, https://doi.org/10.1109/IISA.2018.8633688.

[11] R. Kaur, D. Sharma, "Pre-processing for Improving the Recognition of Mobile based Gurmukhi Text Recognition System", IEEE 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, ISBN:978-1-7281-8594-1, pp. 1-5,: https://doi.org/10.1109/ICCCNT51525.2021.9579867.

[12] Allessio B and Valeria H, "Camera Keyboard: A Novel Interaction Technique for Text Entry Through Smartphone Cameras", IEEE Access, Vol 7, pp. 167982-996, 2019.

[13] R. Kaur, D. Sharma, "Punjabi Text Recognition System for Portable Devices: A Comparative Performance Analysis of Cloud Vision API with Tesseract", Journal of Computer Science and Engineering (JCSE), ISSN 2721-0251, Vol. 2, No. 2, August 2021, pp. 104-111, http://dx.doi.org/10.36596/jcse.v2i2.195

[14] Smith R, "An overview of the Tesseract OCR Engine", 9th International Conference on Document Analysis and Recognition (ICDAR 2007) IEEE, Curitiba, Brazil, pp. 629-633, 2007.

[15] S Badla, "Improving the efficiency of Tesseract OCR Engine. Master's projects, 2014. https://doi.org/10.31979/etd.5avd-kf2 .

[16] Robby G A, Tandra A, Susanto I, Harefa J, Chowanda A, "Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application", Procedia Computer Science, Volume 157, pp. 499-505, ISSN 1877-0509, 2019. https://doi.org/10.1016/j.procs.2019.09.006.

[17] G S Lehal and C Singh , "A Post Processor for Gurmukhi OCR", Sadhna vol 27, part 1, pp. 99-111, 2002.

[18] Z Zhuang and X. Zhu, " An OCR Post-processing Approach Based on Multi-knowledge", Knowledge- Based Intelligent Information and Engineering Systems. KES 2005. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, vol 3681, pp. 346-352, 2005, https://doi.org/ 10.1007/11552413 50.

[19] D. Sharma, G. S. Lehal, "Shape encoded post processing of Gurmukhi OCR", IEEE 10th International Conference on Document Analysis and Recognition, pp. 788-792, 2009, 10.1109/ICDAR.2009.180.

A COMPARATIVE ANALYSIS OF DEEPFAKE DETECTION TECHNIQUES

Ramandeep Kaur [#1], Navdeep Kanwal [*2]
[1,2]*Department of Computer Science & Engineering, Punjabi University Patiala*
[1]deep.91325@gmail.com
[2]navdeepkanwal@gmail.com

ABSTRACT— Deepfake is one of the cornerstones that has obtained a remarkable attention by the media. Deepfakes are manipulated media in which a person in an existing video or picture is convincingly replaced by a computer generated face. Social media is one of the main areas where massive propagation of deepfake takes place. Deepfakes have garnered wide attention for their uses in fake news, fake terrorism events, financial fraud, and hoaxes. Therefore, there is an urgent need for automated tools capable of detecting false media content and avoiding the spread of dangerous false information. For this reason, this review paper presents a comparative analysis of various deepfake detection methods. This paper also examines the use of optical flow based CNN for cross-forgery scenario.

KEYWORDS— Digital forensics, Deepfake manipulations, Deep learning, Neural Networks

INTRODUCTION

Forged multimedia has become a central problem in the past few years, especially after the advent of the so called deepfake, i.e., fake media created with the use of powerful deep learning tools. Photographs, audios and videos are frequently used as evidence in different types of investigations to resolve legal cases since they are considered to be reputable sources. But the phenomena deepfake potentially made these sources of evidences unreliable [1]. Korshunov et al. [2] states that Deepfake is a tampering or manipulation strategy that allows the user to swap the face of one person with another in a digital image and video. The very same technology, however, can be used for malicious purposes, like creating fake porn videos to blackmail people [3], or building fake news campaigns to manipulate the public opinion [4].

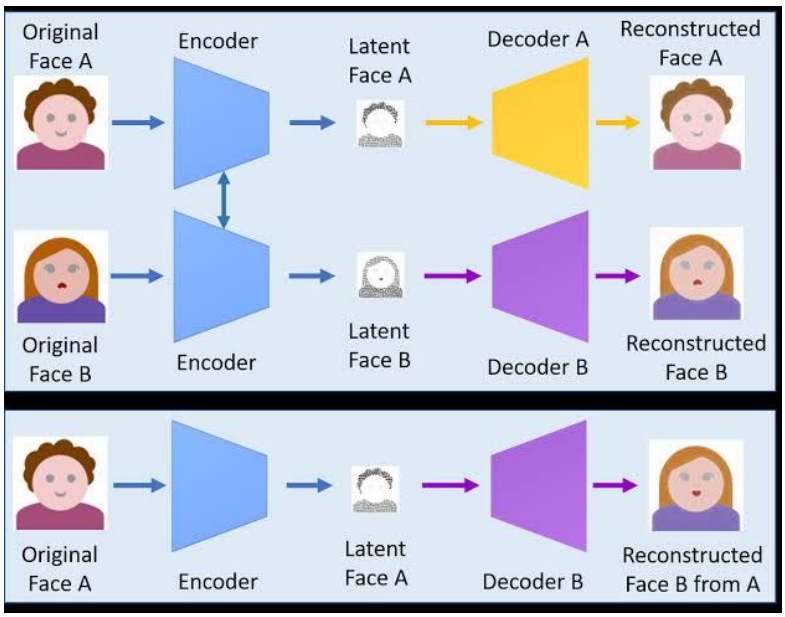


Fig. 1: Deepfake Construction

Applications of Deepfake

While deepfake technology appears to have a negative impact in the majority of cases, it has the potential to transform the multimedia and content development industries. Some well-known applications of deepfake are discussed in the subsequent sections.

a) Multimedia industries

As stated in [5], the game, advertisement, and film industries would greatly benefit from deepfake technology. This is due to fact that through deepfake, actor's dialogues or expressions could synthetically be replaced, which would not only ease the work of editing and save time but can also reduce the cost of production. Multimedia firms can redub the advertisements and films into multiple languages using the phenomena of deepfake. Mouth motions can be completely synchronised with a foreign language because it is based on artificial intelligence [6]. Deepfake is also widely used in the gaming sector. To ensure realistic voice narration, game characters' mouth motions are synchronized with the voice of the actors. As mentioned by Gardiner [5], deepfakes are used in virtual and augmented reality applications.

b) Digital reconstruction and public safety

To reconstruct a crime scene is both a science and an art. It necessitates evidence as well as inductive and deductive reasoning. Synthetic media created by artificial intelligence can aid in the reconstruction of a crime scene [7], [8]. A team of civil detectives also used mobile phone videos to build a virtual crime scene. It relied on autopsy reports and surveillance footage

c) Other social applications

Deepfakes can also be utilized in a range of social applications, such as remote instruction, speech therapy, virtual or personalized digital assistants, and real-time language translation, as mentioned in [6], [5]. Furthermore, deepfake has the potential to be used in medical technology. Project Revoice was created in collaboration with the Lyrebird team [5] for this objective. Project Revoice is an application of deepfake where sufferers of amyotrophic lateral sclerosis (ALS) repossess their distinctive voice, though in computerized form, even after they have lost their capability to speak. Based on these findings, it may be predicted that in the near future, using deepfake, full-fledged digital avatars will be available as vehicles for self-expression for individuals who need them most.

Layout of the manuscript

The current study focuses on the application areas and ongoing difficulties with deepfake detection methods, as well as potential solutions. Fig. 2 depicts the organisation of each part of the manuscript, as well as the discussion of all of its subsections.

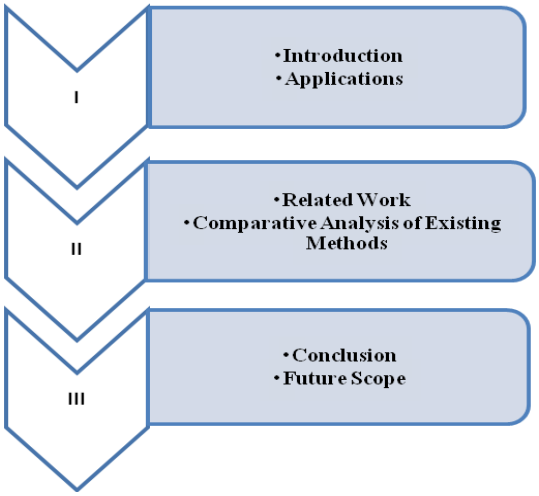


Fig. 2: Paper Organization

LITERATURE SURVEY

This paper presents a comparative analysis of different deepfake detection techniques, their strengths and limitations. An overview of existing techniques is also described in this section.

Related Work

D. Güera et al. [9] discussed a temporal based process that is used to detect deepfake videos automatically. The frame-level features that are extracted using Convolutional Neural Network (CNN) are than used as an input to Recurrent Neural Network (RNN). With the help of these features RNN learns to detect if a video has been tempered or not. The presented technique is evaluated against a huge amount of deepfake videos taken from HOHA dataset [10] and multiple video hosting websites.

E. Sabir et al. [11] proposed a technique that combine the Recurrent Convolutional model and face alignment approach to detect face manipulations in videos, where the input is a sequence of video frames. The first step of the discussed technique detects crops and aligns faces on a sequence of frames. Recurrent convolutional model is utilized in the second step to detect manipulation.

I. Amerini et al. [12] developed a technique that used Optical Flow (OF) fields to exploit possible inter-frame dissimilarities unlike existing state-of-the-art methods. VGG16 CNN [13] Classifier takes cropped OF fields as an input and using sigmoid function detect the deepfake videos. The main emphasis is given on same forgery scenario. Preliminary results obtained on FaceForensics++ [14] dataset highlight very promising performances. The dataset contains total 1000 videos, 720 of the videos are used for training, 120 for validation phase and 120 for testing.

Roberto et al. [15] developed a technique to detect synthetic videos generated by using different kinds of manipulations. The main focus of the technique is on cross-forgery scenario. In first phase of pipeline, video frames are processed to estimate the OF fields that are than cropped according to a dlib face detector. After that cropped OF fields are passed as an input to a CNN ResNet50 whose final fully connected layer is represented by one output unit followed by a sigmoid

activation used for the binary classification of each frame stating if such a frame is tampered or original. The authors provide the future direction to other researchers to evaluate the reliability of the proposed method by testing it against more reference datasets.

Dennis Siegel et al. [16] presented an alternative that detect deepfake with the help of features handcrafted by domain specialists. The key advantage of hand-crafted features over learnt features is their interpretability, as well as the implications for decision plausibility validation. To build DeepFake detection method, three sets of hand-crafted characteristics and three distinct fusion algorithms are discussed.

Yipin Zhou et al. [17] presented a deepfake detection technique, which involves modelling visual and audio modalities together. This effort is imperative since we have no means of knowing whether the video or audio has been altered in practise. Using learned intrinsic synchronisation between video and audio enhanced both video and audio based deepfake detection and whole sequence prediction, according to the technique discussed. The model's generalisation to previously unknown deepfake categories is aided by the learned synchronisation patterns.

Table 1: Comparative analysis of deepfake detection techniques

Reference	Title	Technique	Research Findings
D. Güera et al. 2018[9]	Deepfake Video Detection Using Recurrent Neural Networks	RNN	Frame level features are used as an input to RNN to detect any changes
E. Sabir et al. 2019[11]	Recurrent Convolutional strategies for face manipulation detection in videos	Recurrent Convolutional model	Combination of face alignment and recurrent convolutional model is used to detect any changes
I. Amerini et al. 2019[12]	Deepfake Video Detection through Optical Flow based CNN	OF (Optical Flow) based CNN	Optical Flow field based method is used to detect inter-frame dissimilarities
Roberto et al. 2021 [15]	Optical Flow based CNN for detection of unlearnt deepfake manipulations	OF based ResNet50 CNN	Optical Flow field based method is used to detect cross-forgery
Dennis Siegel et al. 2021 [16]	Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features	Hand-Crafted Features	Hand-Crafted features are used to detect synthetic media rather than neural networks
Yipin Zhou et al. 2021 [17]	Joint Audio-Visual Deepfake Detection	Synchronization pattern between audio and video	Learnt intrinsic synchronization pattern is used to detect unseen deepfake categories.

CONCLUSION

A convincing deepfake may be intended to give false information and forged news, such as a politician giving a speech or making a statement, usually necessitates painstaking alterations of both the video and audio channels. Several deepfake detection strategies have been developed by different researchers. Research findings highlight the strengths and limitations of numerous deepfake detection techniques. In this paper, comparative analysis shows that each strategy uses different techniques for feature extraction and detection of deepfake media. In future work, we can plan to evaluate the reliability of various deepfake detection strategies against datasets.

REFERENCES

[1] M. Koopman, A. M. Rodriguez, and Z Geradts, Detection of Deepfake Video Manipulation, Conference: IMVIP, 2018.

[2] P. Korshunov and S Marcel, Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685.

[3] De Ruiter A. The Distinct Wrong of Deepfakes. Philosophy & Technology. 2021 Jun 10:1-22.

[4] D. Harris, Deepfakes: False Pornography Is Here and the Law Can not Protect You, Duke L. & Tech. Rev, vol. 17, pp. 99–99, 2018.

[5] N. Gardiner, Facial re-enactment, speech synthesis and the rise of the Deepfake. Edith Cowan University, Theses 2019.

[6] R. Chawla, Deepfakes: How a pervert shook the world, International Journal of Advance Research and Development, Vol 4, 2019.

[7] N Kaur, N Kanwal, Review And Analysis of Image Forgery Detection Technique for Digital Images, International Journal of Advanced Research in Computer Science . May/Jun2017, Vol. 8 Issue 5, p2700-2706. 7p.

[8] Kaur, C. deep, & Kanwal, N. (2019). An Analysis of Image Forgery Detection Techniques. Statistics, Optimization & Information Computing, 7(2), 486-500.

[9] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. in: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pages 1–6, 2018.

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, Learning realistic human actions from movies. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8, June 2008.

[11] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos, in: 2019 IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 4396-4405.

[12] I. Amerini, L. Galteri, R. Caldelli, A. Del Bimbo, Deepfake Video detection through optical flow based CNN, in: The IEEE International Conference on Computer Vision (ICCV) Workshops, 2019,pp.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556, 09 2014.

[14] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: IEEE/CVF International Conference on Computer Vision (ICCV), nov 2019.

[15] R. Caldelli, L. Galteri, I. Amerini, A. Del Bimbo, Optical Flow based CNN for detection of unlearnt deepfake manipulations , Pattern Recognition Letters 146 (2021), pp. 31-37.

[16] D. Siegel, C. Kraetzer, S. Seidlitz, J. Dittmann, Media Forensics Considerations on DeepFake Detection with Hand-Crafted Features, MDPI Journal of Imaging 2021, 7, 108.

[17] Yipin Zhou, Ser-Nam Lim, Joint Audio-Visual Deepfake Detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 14800-14809

AN EXTENDED ENCRYPTION ARCHITECTURE TO ENHANCE DATA SECURITY IN TERMS OFQUERIES AND CONTENT AT CLOUD SERVER

Sheenam Malhotra [#1], Williamjeet Singh [*2]
[#1]*Research Scholar, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Punjabi University, Patiala, Punjab. India,*
[2]*Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Punjabi University, Patiala, Punjab. India*
[1]sheenam.malhotra@gmail.com
[2]williamjeet@gmail.com

ABSTRACT— Cloud computing has evolved as an emerging architecture to store and process data in the modern time frame. Security has always been an area where the researchers have put their mind to provide safe environment for the users. This article enhances the security architecture by providing a selection mechanism among the available encryption algorithms namely RSA, NTRU and AES. The proposed algorithm architecture reduces the computation complexity of the storage and search of the data from the cloud server. In order to rank the data, the proposed algorithm uses Artificial Neural Network in order to apprehend the relevance of the data to the query. There are two segments of the proposed algorithm namely the storage and the retrieval section. The evaluation of the result has been done utilizing the standard data set Kaggle based on the retrieval complexity and submission complexity of the data.

KEYWORDS— Artificial Neural Network, encryption, computation complexity, cloud computing, secure cloud

INTRODUCTION

Increasingly, cloud computing is attracting academic and IT industry interest due to its prominence as the main deployment platform of distributed applications, especially for huge data management. The public cloud provides end-users with the ability to store and retrieve personal information such as passwords, e-mails, personal health data records, connected financial information, and government papers [1]. The cloud architecture allows for resource sharing and elastic pay-as-you-go policies that adjust to changing demand. One of the main principles that help the cloud paradigm work better is resource sharing. Data communication comprises two stages, namely storage and retrieval, to provide the user with precisely what they want.

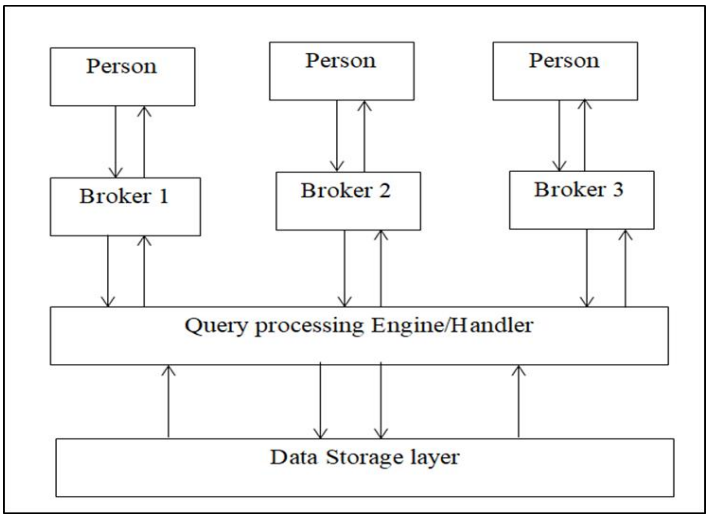


Fig. 10 cloud storage architecture

To distribute resources across many cloud users, cloud architecture provides an appropriate communication route [2]. Tech giants such As Amazon, Google, Apple, and others are already making use of the best cloud advantages by controlling storage rules and improving retrieval strategies to respond quickly. Figure 1 shows the three levels of a cloud network architecture, from the user layer through the query handler and finally to the storage layer [3].

The query processing engine in the cloud receives the request either directly from the user or through a broker. The processing engine sifts through the data to find what's significant and then applies the appropriate processing [4]. In this case, the data is sent to the file storage layer. The file storage layer or the broker provides the information if it is needed for processing. Process results are returned to the user; hence the processing element also serves as a responder. The issue of security has been raised by several researchers depending of the type of searches made into the server. Security-enhanced encryption (SEC) is difficult and resource-intensive, and its design must be effective while handling encrypted information. SE allows you to do searches based on a single keyword or a set of keywords. To use SE, the sender must provide the plain text to be ciphered by the process engine. There is additionally a token bit representing a key, in addition to the plain text.

There are several security algorithms in the cloud architecture and they use different encryption algorithms for the encryption of the data such as RSA, AES and NTRU which are described briefly as follows.

KK. RSA algorithm

RSA stands for Rivest, Shamir, Adleman. [5] they invented public-key encryption, which is a public-key cryptosystem for secure data transfer. It's a common encryption technique for sending sensitive data, particularly when doing so over the internet. The RSA algorithm makes use of the difficulties of factoring huge integers. The RSA technique is based on the idea that factoring really large numbers is time-consuming. As a result, deducing an RSA key would require a long time and a lot of computing resources. The RSA algorithm is an asymmetric encryption algorithm because it uses two keys: a public and a private key. The public key is shared with everyone, while the private key is kept secret. One of the integers in the public key is a multiplication of two huge prime numbers.

LL. NTRU Algorithm

NTRU stands for Nth degree truncated polynomial Ring Units. [6] It is an open-source public-key cryptosystem that encrypts and decrypts data using lattice-based encryption. It is made up of two algorithms: NTRU-Encrypt for encryption and NTRU-Sign for digital signatures. It is immune to assaults utilizing Shor's algorithm, unlike other prominent public-key cryptosystems. NTRU conducts expensive private key operations significantly quicker than RSA with similar cryptographic strength. The time it takes to do an RSA private operation grows as the cube of the key size grows, but the time it takes to perform an NTRU action grows quadratically.

MM. AES Algorithm

AES stands for Advanced Encryption Standard [7]. It is a symmetrical block cipher that turns plain text into ciphertext in blocks of 128 bits utilizing keys of 128, 192, and 256 bits. The Rijndael algorithm is another name for this method. The AES algorithm is a global standard because it is deemed safe. To generate ciphertext, the AES algorithm employs a substitution-permutation (SP) network with many rounds. The number of rounds is determined by the key size. Ten rounds are dictated by a 128-bit key size, 12 rounds by a 192-bit key size, and 14 rounds by a 256-bit key size. Each of these rounds requires around the key, but since the method only accepts one key, it must be enlarged to get keys for each round, including round [8].

The contribution of the proposed algorithm embeds an encryption selection algorithm which relies on the type of data which is submitted to the server. This reduces the computation complexity and is illustrated in the result section

RELATED WORK

Researchers' contributions to the area of SE may be roughly classified into two categories: (i) single-keyword searchable encryption, and (ii) multi-keyword searchable encryption. When SE was first introduced, they used the cryptographic technique to overcome the search difficulties that appeared in encrypted data. [9] Created a mechanism employing the "Bloom Filter" to allow safe indexing over encrypted data as time passed, and the encrypted index became popular as a result.

[10] Addressed the challenge of using Multi-Keyword Ranked Search (MKRS) to encrypt cloud data with comparable queries. The researchers focused primarily on two issues: (i) improving search results using MKRS, and (ii) supporting comparable quarries using synonym-based search. [11] introduced a multi-query approach to alleviating the cost associated with keyword repository growth. This is accomplished by weighing the importance of each search term and the historical performance of query results.

[12] Addressed the multi-keyword search issue presented by a lot of cloud data consumers. Data users with verified data were able to search reliably, simply, and effectively across numerous cloud data users using this study, unlike earlier research. [9] Have provided for the first time an efficient encryption technique that supports MKRS and parallel search. To improve the efficiency of the search, a tree-based index model has also been proposed, which makes use of the Vector Space Model.

A Multi-Keyword Synonym-based Fuzzy Ranked Search strategy was developed by researchers [13] to address the issue of users forgetting keywords during the document retrieval process. By combining a synonym pool with keywords, this may be accomplished. At the same time, an index update is also supplied. [14] Have created a methodology for protecting cloud computing privacy from outsourced data. To train the categorization model, several public keys belonging to various data owners have been obtained. According to [15] under the twin cloud model based on secret sharing and homomorphic cryptosystem, a privacy preservation technique based on KNN architecture was presented for use. Lightweight building blocks were reinvented to offer comparable security while also being more efficient. [16] Originally developed an Attribute-Based Keyword Search (ABKS) to address the issue of terms harbouring internal attackers. [17]have found a solution to the issue of long calculation times in attribute-based multi-keyword search (ABMKS). To optimize search performance and improve system dependability, the encrypted keyword indexes were combined into a single element regardless of the number of main keywords in a document [18].

[19] To provide dual security for ciphertext and trapdoor in cloud data do not meet the author's expectations. In practice, the recommended technique is insecure. A viable cloud model and deterrent idea were introduced for safe and efficient file sharing. To keep communication costs low, the model requires the cloud server to operate unlawfully. A search technique

based on the Mapping Set Matching (MSMR) was presented by [20] for searching ciphertext using multiple keywords. The job entails coordinating document searches across several cloud services. To save computation time, this data is shared with the public cloud, which uses it to find documents that satisfy the search criteria.

PROPOSED WORK

The proposed work is divided into two segments. The first segment is for data submission and the second segment is for the content retrieval from the cloud server. The data is already categorized into subsequent group based on the similarity attained between the dataset elements. In order to segregate the data documents, the proposed work has utilized k-means clustering algorithm. In order to select the encryption mechanism, the proposed work uses the complexity finder architecture provided by Sheenam et al. [21]. In order to choose the encryption algorithm, cosine similarity has been earlier used as the key evaluator between the data elements that are to be stored on the cloud server. The dual threshold rule is applied to generate two threshold ranges namely C1, C2 and C3 and an addition bit is added to the encrypted part in order to detail the type of encryption that has been applied. When it comes to retrieval of the document, ANN based search and ranking has been imposed.

The work can be explained in terms of pseudo code as follows.

Algorithm Content Storage and Retrieval

Input: Data, Data Category

1. Take user credentials
   a) Input Id
   b) Password
   c) Captcha
2. If User Information matches to Cloud Base information
   a) Select the option to Save or Retrieve
   b) If_1 Option. Result== Save

i. Apply complexity finder [21]
ii. Choose encryption algorithm based on complexity finder
iii. Append Encryption Verification Code (EVC)

EVC= 1 If RSA Applied
EVC=2. If AES Applied
EVC=3 If NTRU Applied

iv. Check storage space against the user account
v. If_2 Storage Requirements are satisfied

Store data to server
Else_2. // Else part of the second if
Ask user to upgrade storage
End If_2. // Ending the if statement

c) Else_1. // If the user has to retrieve the content

i. Take user query
ii. Apply complexity finder [21]
iii. Apply EVC to encrypted query
iv. Transfer query to database server
v. Decrypt query by de-ciphering EVC
vi. Run back propagation simulation for content retrieval
vii. Calculate HAC Index by as follows
viii. HAC encrypted index is generated by the following equation

ix. $$HAC = \frac{\|ROT\|}{\|CROT\|} \quad (1)$$

x. $$\|ROT\| = \frac{ROT}{\sqrt{\sum_{i=1}^{K}(ROT_L)^2}} \quad (2)$$

xi. $$\|CROT\| = \frac{CROT}{\sqrt{\sum_{i=1}^{K}(CROT_L)^2}} \quad (3)$$

xii. Arrange the documents as per the most matched HAC index when propagated through neural networks.

3. Provide data to user
4. If_4 User. Session Expires
   Log Out User
5. End Algorithm

Where, stands for the ratio of occurrences of terms up to L evaluations, stands for the ratio of occurrences of terms up to L evaluations.

The algorithm initially validates the users based on the registered Id and password of the user. Along with the user-id and the password, a captcha is also generated and integrated to the authentication process. If the user is authenticated, the based on the user's accessibility, the user selects its desired operation viz. storage or retrieval. If the client chooses to store the data, the proposed algorithm uses complexity finder mechanism [21]. An additional bit in order to reduce the overall complexity, is appended at the last of the encrypted bit pattern. If the user's account, meets the storage requirements, the database engine saves the data in encrypted form at the cloud server. If the user opts to retrieve content, a neural engine is applied to train the system based on the generated HAC index as explained in the earlier part of the algorithm. The neural engine propagates the HAC value against the stored and propagated HAC values and most similar propagated contents are placed at the top of the retrieved document set.

RESULTS AND DISCUSSION

The evaluation of the results has been based on the following parameters.

a) Computation Cost Ratio (CCR)

It is the ratio of the computation complexity of the proposed selection and retrieval algorithm to the individual encryption algorithm computation complexity. The computation complexity is measured in seconds.

$$CCR = \frac{Computation\ Complexity\ of\ retrieval\ proposed}{(Computation\ Complexity\ of\ individual\ algorithm)} \quad (4)$$

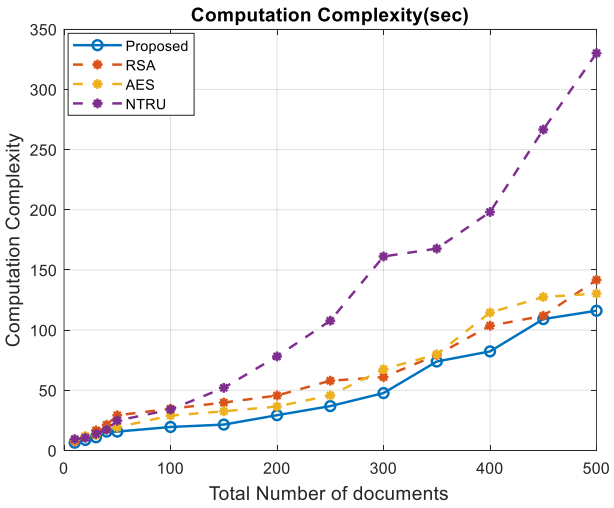


Fig. 11 Computation Complexity

From the Fig. 2 it is observed that the proposed scheme associated with minimal computational complexity than all the other mentioned encryption algorithms, and this following statement is supported by the bar plot of CCR (Fig..3).

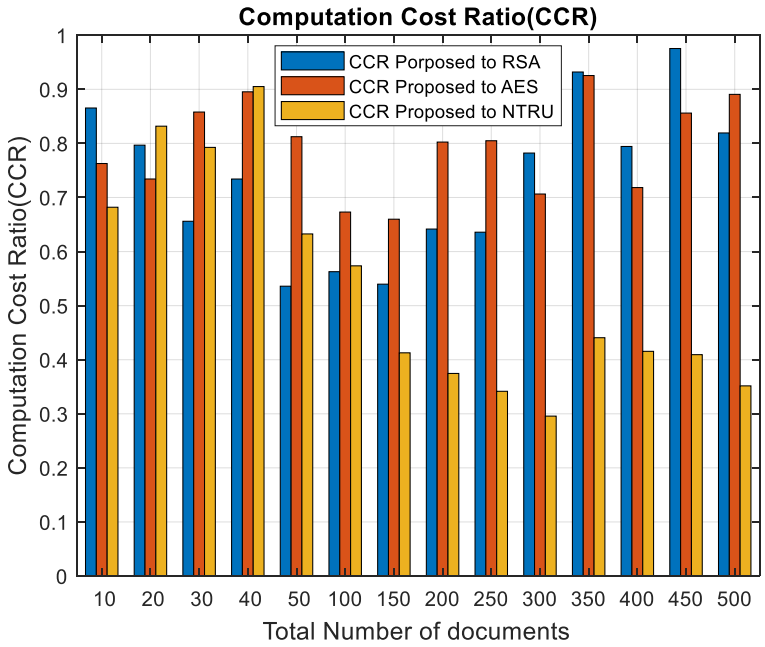


Fig. 12 Computation Cost Ratio

As stated, the fact of computational complexity reduction is observed by the parameter computation cost ratio, which was framed in the above Fig. 3.

TABLE XIV

PERCENTAGE OF TIME CONSUMPTION FOR CERTAIN NUMBER OF DOCUMENTS

Total Number of documents	CCR proposed to RSA	CCR Proposed to AES	CCR Proposed to NTRU
10	0.865448527	0.762811162	0.681937214
20	0.796733397	0.734171341	0.831844028
30	0.655985143	0.857960538	0.792593777
40	0.734139848	0.89536299	0.905081761
50	0.535914211	0.812374174	0.632555966
100	0.562758362	0.673040758	0.573597814
150	0.539672648	0.659788827	0.412711243
200	0.641577143	0.802441838	0.374594649
250	0.635815683	0.804910982	0.341585053
300	0.782127666	0.706390681	0.295686205
350	0.932004451	0.925497463	0.440642878
400	0.794304316	0.718317862	0.415473109
450	0.975457045	0.856061612	0.409340087
500	0.819280732	0.890674069	0.351618155

The above table depicts the percentage of time that a particular algorithm consumes when compared them with the proposed algorithm for a certain number of documents. If observed on an average, the algorithm RSA consumes 73% of computational time when compared with the proposed scheme. The algorithm AES consumes 79% and NTRU consumes 53% of computational time when compared with the proposed scheme.

b) Make Span

It is total time consumed in order to store and retrieve a set of documents.

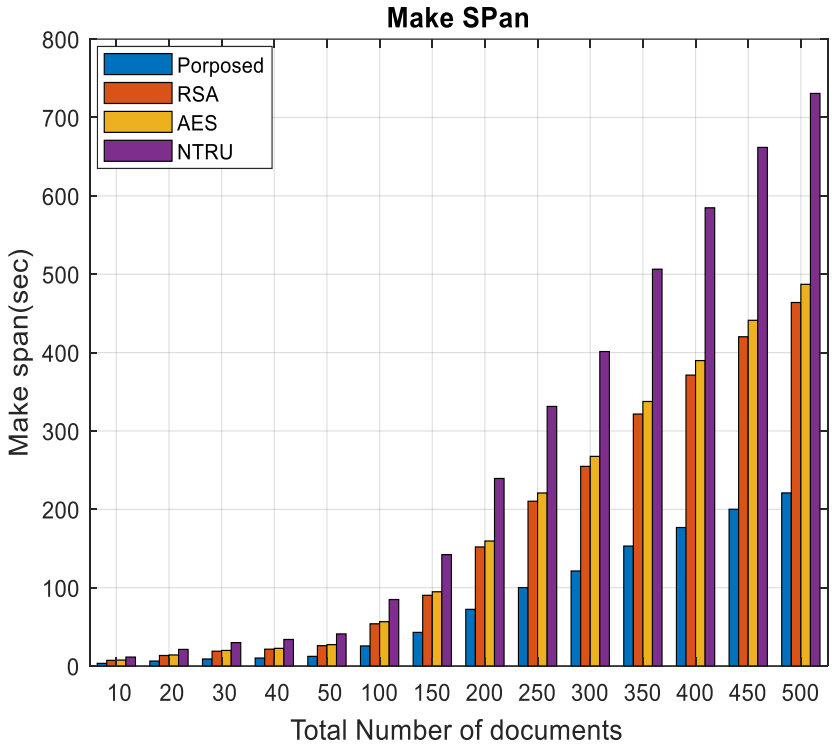


Fig. 13 Make span

From the above Fig. 4 it is clear that the proposed scheme associated with minimal make span which supports the fact that the proposed scheme is comparatively consumes less time for document storage and retrieval.

TABLE XV

TIME ELAPSE (IN SECONDS) FOR CERTAIN NUMBER OF DOCUMENTS

Total Number of documents	Proposed	RSA	AES	NTRU
10	3.892308678	8.173848	8.582541	12.87381
20	6.180431818	12.97891	13.62785	20.44178
30	8.99119954	18.88152	19.82559	29.73839
40	11.09590298	23.3014	24.46647	36.6997
50	12.36350553	25.96336	27.26153	40.89229
100	25.04098231	52.58606	55.21537	82.82305
150	42.92502628	90.14256	94.64968	141.9745
200	72.36553188	151.9676	159.566	239.349
250	100.7826721	211.6436	222.2258	333.3387
300	121.6821345	255.5325	268.3091	402.4637
350	153.3101138	321.9512	338.0488	507.0732
400	176.5412206	370.7366	389.2734	583.9101
450	200.3310515	420.6952	441.73	662.595
500	220.4048989	462.8503	485.9928	728.9892

The above table depicts the time elapse for document storage and retrieval of the chosen algorithms for certain number of documents. The proposed scheme has an average of 82 seconds of time elapse. The algorithm RSA have 173 seconds, AES have 181 seconds, NTRU have 272 seconds of time elapse. Hence it is observed that the proposed scheme has the least time elapse when compared with other algorithms.

c) Relevance Percentage

The relevance percentage is calculated by taking the ratio of the term frequency of the query and the average keyword term frequency of the retrieved document.

$$Relevance\ Percentage = \frac{\frac{\sum_{i=1}^{p} Tf_{query}}{p}}{\frac{\sum_{j=1}^{l} Tf_{stored_{document}}}{l}} \quad (5)$$

where p is total number of keywords in the query and l is total number of keywords in the stored document.

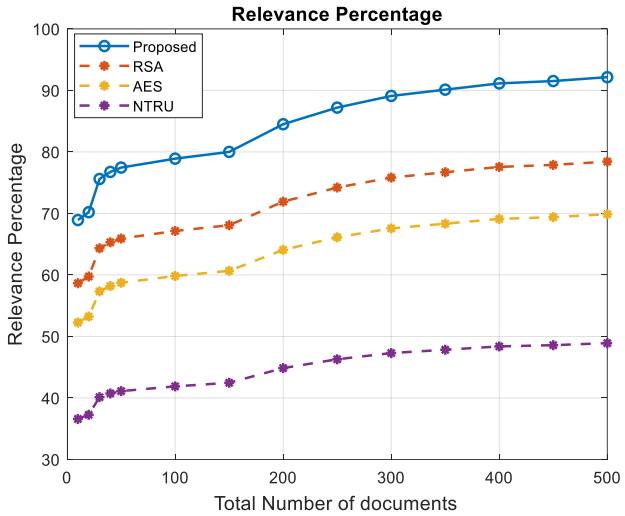


Fig. 14 Relevance Percentage

From the above figure it is clear that the relance percentage of the proposed scheme is comparatively higher that the other schemes. This property enhances the efficiency of document retrieval procedure.

TABLE XVI

PERCENTAGE OF RELEVANCE FOR CERTAIN NUMBER OF DOCUMENTS

Total Number of documents	Proposed	RSA	AES	NTRU
10	68.92	58.65092	52.25797	36.58058
20	70.2	59.7402	53.22852	37.25996
30	75.6	64.3356	57.32302	40.12611
40	76.74	65.30574	58.18741	40.73119
50	77.45	65.90995	58.72577	41.10804
100	78.9	67.1439	59.82521	41.87765
150	80	68.08	60.65928	42.4615
200	84.5	71.9095	64.07136	44.84996
250	87.2	74.2072	66.11862	46.28303
300	89.1	75.8241	67.55927	47.29149
350	90.12	76.69212	68.33268	47.83288
400	91.14	77.56014	69.10608	48.37426
450	91.52	77.88352	69.39422	48.57595
500	92.15	78.41965	69.87191	48.91034

The above table displays the relevance percentage of the keywords for a particular algorithm for a certain number of documents. As the number of documents gets increased, the relevance percentage of the keywords also gets increased. If observed on average, the proposed scheme has 82 relevance percentage. The algorithms RSA have 70%, AES have 62%, and NTRU have 43%. Hence it is observable that the proposed scheme has high relevance percentage for keywords when compared to other algorithms.

CONCLUSIONS

The proposed work and all the other mentioned encryption algorithms are simulated using MATLAB 2016a.In this paper an encryption selection algorithm is proposed to reduce the computation complexity, whose effectiveness is verified by comparing the CCR, Make span and relevance percentages of the proposed scheme with the encryption algorithms such as RSA.AES and NTRU for various number of documents. And the outcome of thestudy supports the efficiency of the proposed state of art.

REFERENCES

[1] N. Subramanian and A. Jeyaraj, "Recent security challenges in cloud computing," *Comput. Electr. Eng.*, vol. 71, no. July 2017, pp. 28–42, 2018, doi: 10.1016/j.compeleceng.2018.06.006.

[2] P. R. Kumar, P. H. Raj, and P. Jelciana, "Exploring Data Security Issues and Solutions in Cloud Computing," *Procedia Comput. Sci.*, vol. 125, no. 2009, pp. 691–697, 2018, doi: 10.1016/j.procs.2017.12.089.

[3] C. H. V. N. U. Bharathi Murthy, M. L. Shri, S. Kadry, and S. Lim, "Blockchain based cloud computing: Architecture and research challenges," *IEEE Access*, vol. 8, pp. 205190–205205, 2020, doi: 10.1109/ACCESS.2020.3036812.

[4] L. Ziouche, S. Ben Meskina, M. Khalgui, L. Kahloul, and Z. Li, "Smart grid rebuilding based on cloud computing architecture," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 2019-Octob, no. April 2021, pp. 2259–2266, 2019, doi: 10.1109/SMC.2019.8914432.

[5] O. G. Abood and S. K. Guirguis, "A Survey on Cryptography Algorithms," *Int. J. Sci. Res. Publ.*, vol. 8, no. 7, 2018, doi: 10.29322/ijsrp.8.7.2018.p7978.

[6] O. M. Guillen, T. Poppelmann, J. M. Bermudo Mera, E. F. Bongenaar, G. Sigl, and J. Sepulveda, "Towards post-quantum security for IoT endpoints with NTRU," *Proc. 2017 Des. Autom. Test Eur. DATE 2017*, pp. 698–703, 2017, doi: 10.23919/DATE.2017.7927079.

[7] P. Chittibabu, M. Kannan, C. Priya, S. Vaishnavisree, and R. Scholar, "a Comparative Analysis of Des, Aes and Rsa Crypt Algorithms for Network Security in Cloud Computing," *J. Emerg. Technol. Innov. Res.*, vol. 6, no. 3, 2019, [Online]. Available: http://doi.one/10.1729/Journal.19997.

[8] N. Hemala, "A Survey on Encryption Algorithms – RSA , DES , AES , And SIT Algorithm," vol. 5, no. 4, pp. 61–65, 2018, doi: 10.30750/ijrast.5410.

[9] S. Kumar and H. Kim, "Packet rate adaptation protocol based on bloom filter for hidden node avoidance in vehicular ad-hoc networks," *IEEE Access*, vol. 7, pp. 137446–137460, 2019, doi: 10.1109/ACCESS.2019.2942971.

[10] C. F. Wu, Y. W. Ti, S. Y. Kuo, and C. M. Yu, "Benchmarking Dynamic Searchable Symmetric Encryption with Search Pattern Hiding," *Proc. - 2019 Int. Conf. Intell. Comput. Its Emerg. Appl. ICEA 2019*, pp. 65–69, 2019, doi: 10.1109/ICEA.2019.8858302.

[11] K. M. Mohan, "Multi-Keyword Similarity Search Using Asymmetric Encryption," vol. 25, no. 4, pp. 11423–11438, 2021.

[12] A. F. Fard and M. M. Ardakani, "Researchable encryption in cloud databases: a survey," 2020, [Online]. Available: www.japer.in.

[13] T. Ahmed, Ashrafunnessa, and J. Rahman, "Development of a visual inspection programme for cervical cancer prevention in Bangladesh," *Reprod. Health Matters*, vol. 16, no. 32, pp. 78–85, 2008, doi: 10.1016/S0968-8080(08)32419-7.

[14] W. Jiang, H. Li, G. Xu, M. Wen, G. Dong, and X. Lin, "PTAS: Privacy-preserving Thin-client Authentication Scheme in blockchain-based PKI," *Futur. Gener. Comput. Syst.*, vol. 96, pp. 185–195, 2019, doi: 10.1016/j.future.2019.01.026.

[15] A. Mohiyuddin, A. R. Javed, C. Chakraborty, M. Rizwan, M. Shabbir, and J. Nebhen, "Secure Cloud Storage for Medical IoT Data using Adaptive Neuro-Fuzzy Inference System," *Int. J. Fuzzy Syst.*, no. June, 2021, doi: 10.1007/s40815-021-01104-y.

[16] Y. Zhou, S. Zheng, and L. Wang, "Privacy-preserving and efficient public key encryption with keyword search based on cp-abe in cloud," *Cryptography*, vol. 4, no. 4, pp. 1–14, 2020, doi: 10.3390/cryptography4040028.

[17] X. Liu, T. Lu, X. He, X. Yang, and S. Niu, "Verifiable Attribute-Based Keyword Search over Encrypted Cloud Data Supporting Data Deduplication," *IEEE Access*, vol. 8, pp. 52062–52074, 2020, doi: 10.1109/ACCESS.2020.2980627.

[18] Y. Cui, F. Gao, Y. Shi, W. Yin, E. Panaousis, and K. Liang, "An Efficient Attribute-Based Multi-Keyword Search Scheme in Encrypted Keyword Generation," *IEEE Access*, vol. 8, pp. 99024–99036, 2020, doi: 10.1109/ACCESS.2020.2996940.

[19] J. Sun, L. Ren, S. Wang, and X. Yao, "Multi-Keyword Searchable and Data Verifiable Attribute-Based Encryption Scheme for Cloud Storage," *IEEE Access*, vol. 7, pp. 66655–66667, 2019, doi: 10.1109/ACCESS.2019.2917772.

[20] C. Mao, Q. Wang, K. Feng, S. Yang, and L. Bao, "Research on the Mechanism of Multi-resource Mapping Matching and Transmission of Road System Based on Meme Theory," *J. Phys. Conf. Ser.*, vol. 1838, no. 1, 2021, doi: 10.1088/1742-6596/1838/1/012048.

[21] S. Malhotra, W. Singh, "An Optimized Solution for Ranking Based On Data Complexity", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, vol. 8, no. 11, September 2019

ROLE OF ISRO'S KU-BAND BASED SCATSAT-1 IN AGRICULTURE APPLICATIONS

Ravneet Kaur[a]; Raman Maini[b;] Reet Kamal Tiwari[c]; Sartajvir Singh[d]
[a,b]Department of Computer Science Engineering, Punjabi University, Patiala, Punjab 147 002.
[a]Apex Institute of Technology – Department of Computer Science and Technology, Chandigarh University, Gharuan, Punjab 140 407.
[c,d]Department of Civil Engineering, Indian Institute of Technology, Ropar, Punjab, India 140 001.
[d]Chitkara University School of Engineering and Technology, Chitkara University, Himachal Pradesh, India 174 103.
[a]ravneet.e11361@cumail.in;
[b]*research.raman@gmail.com*;
[c]*reetkamal@iitrpr.ac.in*;
[d]sartajvir.singh@chitkarauniversity.edu.in
Funding
This work is sponsored by SERB-DST under Grant no. TAR/2019/000354.

ABSTRACT:— In the present work, the applicability of the Indian Space Research Organization's (ISRO) Scatterometer Satellite (SCATSAT-1) is explored in the field of agriculture. The SCATSAT-1 provides the backscattered measurements at a frequency of 13.5 GHz (Ku-band). The backscattered coefficient (, sigma nought) is sensitive to the roughness of the surface and soil moisture contents which highlights the utilization of SCATSAT-1 in agriculture applications. The global level coverage and daily availability make the SCATSAT-1 more competent towards the management of agricultural land and of paddy crops to meet the requirements of the future. In existing work, numerous studies were conducted to estimate the agricultural land using SMAP (Soil Moisture Active Passive), and optical data-based NDVI (Normalized Difference Vegetation Index). However, the applicability of SCATSAT-1 has also been explored in various applications such as paddy crop, jute crop and crop yield estimation with the help of multivariate analysis models, statistical models, and classification methods. This work briefly explains the applications of Ku-band-based SCATSAT-1 in agricultural land, the current status of models developed for agriculture applications and future requirements. This study helps the researcher to identify the applicability of SCATSAT-1 in mapping of different types of land cover.

Keywords: SCATSAT-1, SMAP, MODIS, NDVI, backscattered coefficient

I. INTRODUCTION

Scatterometer plays an important role in global level monitoring of winds over the ocean surface, sea-ice over the polar region and essential land cover components (i.e., cryosphere parameters, vegetation dynamics and hydrological parameters) [25,26,29]. Since its first launch in 1978, many spaceborne scatterometers have been launched by various agencies such NASA, ESA, I,SRO, and CNS [39]. In literature, the applications of scatterometers have been in many scientific domains, but it is continuously enhanced due to the development and improvement in the resolution of scatterometers [24]. The scatterometer applications can be categorized into four domains (a) cryosphere, (b) oceanography, (c) agriculture, (d) hydrology. The cryospheric applications include the estimation of sea-ice extent [17,32,36], ice-bergs [2], snow cover [17,32,36] and snow water equivalent [29]. The oceanography application includes the sea winds [3], tropical cyclone monitoring [12], ocean dynamics [15,31] and sea surface temperature [7,50]. The agriculture applications include the paddy crop estimation [30,47], crop phenology [6], and soil moisture estimation [5,14,50]. The hydrological applications include water level estimation and forecasting of droughts [11].

Here, we focused on the applications of scatterometer satellites (SCATSAT-1), especially in agriculture. The SCATSAT-1 as an active microwave device was launched on 26th September 2016 by the ISRO from Satish Dhawan Space Centre, Sriharikota, India. As an active microwave sensor, it sends electromagnetic waves to the surface and gets the signals back after diverting from the objects. It provides numerous advantages such as dairy-based data delivery, all-weather monitoring, global coverage, and is freely available to researchers. Therefore, the application coverage is high as compared to optical datasets. Moreover, the SCATSAT-1 data available in sigma- nought (σ°), gamma nought (γ°) and brightness temperature (Tb). All these parameters are available at HH and VV polarizations to highlight the different properties of the earth surface. Moreover, the SCATSAT-1 products are available at four different levels at a different resolution to enhance their application range. The detailed information can be found in various studies [15,22,39].

The applicability of SCATSAT-1 has already been tested in a variety of various scientific domains such as cryosphere, oceanography, agriculture, hydrology. But here, we will cover the brief utilizations of SCATSAT-1 in agriculture applications only. This introduction section is followed by the background of SCATSAT-1. Afterwards, the applications have been explored in the agricultural land. At last, the conclusion has been drawn.

II. TECHNICAL SPECIFICATION AND PRODUCTS OF SCATSAT-1

Generally, the SCATSAT-1 facilitates to capture of the direction and velocity of the wind over the land and ocean surface. A brief overview of SCATSAT-1 products and technical specifications is shown in Fig 1 and Table 1, respectively. There are four level of SCATSAT-1 data products available that can be categorized as:

(a) Level-1B (L1B) (σ° coefficient with ~6×30 km resolution) for nominal winds;

(b) L2A (σ° coefficient with 25×25 km resolution) for cyclone ad weather;
(c) L2B (Wind velocity with 50×50 km resolution) for weather forecasting;
(d) L3S (σ° with 50×50 km resolution);
(e) L3W (wind velocity with 25×25 km resolution);
(f) L3IC (ice cover with 25×25 km resolution);
(g) L3BT (brightness temperature with 25×25 km resolution) for forecasting services to respond the hazardous situations;
(h) L4_India (σ°, γ°, and Tb with 2×2 / 6.25×6.25 km resolution);
(i) L4_North_Polar (σ°, γ°, and Tb with 2×2 / 6.25×6.25 km resolution);
(j) L4_South_Polar (σ°, γ°, and Tb with 2×2 / 6.25×6.25 km resolution);
(k) L4_Full_Globe (σ°, γ°, and Tb with 2×2 / 6.25×6.25 km resolution) for land surface applications;
(l) L4AW (analysed winds with 25×25 km resolution);
(m) L4AW_625 (AW with 6.25 km resolution);
(n) L4HA (high-resolution (HR) AW with 6.25×6.25 km resolution);
(o) L4HW (HR winds with 6.25×6.25 km resolution), and
(p) L4UI (upwelling index with 25×25 km resolution) for retrieval of wind vector, weather forecasting monitoring, cyclone prediction.

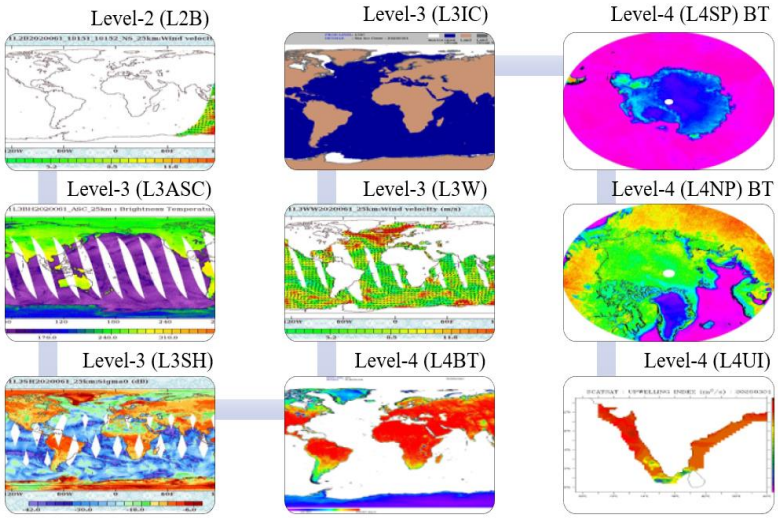


Fig. 1 Representation of different SCATSAT-1 products (Source: MOSDAC).

Table 1. Technical specifications of SCATSAT-1

S.no.	*Parameters*	*Value*
1	Altitude	723 km
2	Orbit	Polar (Sun-synchronous)
3	Operating frequency	13.5 GHz
4	Polarization modes	HH/VV
5	Swath width	1400−1800 km
6	Scanning radius	700−920 km
7	Inclination angle	98°
8	Scanning rate	20.5 rpm
9	Gain	39 dBi
10	Wind speed range	3–30 m/s with accuracy of 10%
11	Antenna Diameter	1m
12	Pulse repetition frequency	193 Hz
13	Transmit Pulse Width	1.35ms
14	Wind Direction	0 to 360 degrees
15	Orbital period	99.19 minutes
16	No. of orbits per day	14+1/2

In earlier studies, different methods and models have been developed to analyse and validate the different SCATSAT-1 products. The velocity of wind can be measured by Level-1 products [20]. A natural disaster like cyclones can be identified by the Level-2 products [4,12]. The weather forecast and the identification of natural hazards can be identified by the Level-3 products [13,15,16,19]. Moreover, the Level-4 products are used in the monitoring of land cover, atmospheric, cryosphere, rice crop phenology and predictions, retrieval of wind vector, weather forecasting monitoring, and cyclone prediction [6,10,11,19,21,31,34,46]

IV SCATSAT-1 IN AGRICULTURE

With the launch of SCATSAT-1, numerous applications have been demonstrated to validate the utilization of SCATSAT-1 in agriculture applications. In the agriculture field, a variety of surface attributes access the scatterometer backscattered signal such as roughness of the surface, the vegetation area, water contents in vegetation, and soil moisture [25]. The most

demanding task is to estimate the vegetation surface using a dataset of scatterometer through backscatter signal [23]. The backscattering signal strength increases with decreases in incidence signal which is generally dependent on surface roughness. Moreover, the incidence angle has an impact on the backscattered signal, which is especially important for vegetation research [25]. It is also observed that the coarseness on the soil surface is less comparable to vegetation which highly varies with the change in season. However, the Level-4 products of SCATSAT-1 are more suitable for the understanding of the vegetation dynamics and soil properties [6,46]. The level-4 products have been generated using various pre-processing techniques as described in different studies [20,38,42]. In agriculture, the major applications involve soil moisture, paddy crop phenology, crop yield estimation, detection of jute crop and leaf area index (LAI).

The Ku-band SCATSAT-1 can deliver the soil moisture measurements on daily basis under all weather conditions. There are several applications of soil moisture products such as flood forecasting, drought monitoring, soil erosion and crop/vegetation forecast. Soil moisture is one of the most important elements in plant growth and climate, and it can be used as a key indicator in agricultural productivity and forecasting. In the past decade, the microwave sensors are used to retrieve and analyse soil moisture using theoretical, empirical, and semi-empirical methods [21], including some improved models [9,27,48]. Moreover, the paddy crop is mostly found in South Asian countries like India, Bangladesh, Bhutan, Nepal, Sri Lanka, and Pakistan which helps to improve the national economy. As per Indian weather, in monsoon rains, rice production is cultivated, with ~60% of rice assured irrigation. India has its own operational programme i.e., National Rice Crop Monitoring that utilizes the RISAT-1, Sentinel-1 and C-band SAR for paddy crop estimation.

Reference [47] reported the monitoring and forecasting of paddy crop yield using multiple regression models based on SH/SV ratio and yield collected from Crop Cutting Experiment (CCE). They have conducted this experiment over six countries. As an outcome, they have shown the possibility of national or state level yield forecasting with the help of the SCATSAT-1 dataset. Reference [10] also shown the potential of scatterometer in the estimation of paddy crop phenology using SCATSAT-1 Scatterometer data and Sentinel-1 dataset at three different stages of rice crop using Hierarchical decision rule-based classification. As result, it has been observed that different stages are well-matched with the reference dataset. Moreover, Reference [30] also estimated the various crop phenological parameters using different satellite-based datasets. This study involves the utilization of statistical methods. The outcomes confirm the feasibility of SCATSAT-1 in the identification of crop phenological stages effectively. Reference [46] demonstrated the utilization of SCATSAT-1 in jute yield estimation at both polarization modes using statistical models over West Bengal and Assam, India. The outcomes have shown the effectiveness of sigma-nought in the estimation of the jute fibre yield at district and state levels one month before the harvest.

Singh et al., (2019) demonstrated a methodology to generate LAI using SCATSAT-1. The LAI is one of the most important factors for understanding plant biophysical processes. It's also an important part of soil moisture modelling (Singh et al., 2019b). In previous literature, various datasets, i.e., ERS-1/ERS-2, RADARSAT-1, ENVISAT-ASAR (Advanced Synthetic Aperture Radar) (during the period 2002–2012), RADARSAT-2 (2007), and Sentinel-1 (2014) have been utilised for vegetation analysis. Recently, two models are developed [43] i.e., the water cloud model (WCM) [1,28] to estimate or recover the LAI via SCATSAT-1 cross-section (σ°). It has been observed that both models provide satisfactory outcomes but the impact of [28] is certainly better as compared to the WCM model. Further, it is expected that SCATSAT-1 can also be explored in different crop applications with the utilization of multispectral dataset [40].

III. CONCLUSION

This study provides a brief overview regarding the applications of SCATSAT-1 in the agriculture field. From this present work, it has been analysed that SCATSAT-1 is one of the competent scatterometers in terms of improvement in resolution, global coverage, and products availability. Moreover, it encourages the researchers to be utilized at a global level on daily basis under the all-weather condition in further exploration of the agricultural land cover region. Nevertheless, due to the coarse resolution, the applicability is limited over the specific land cover regions and generally, suitable for the global level, national level or state level studies. However, further research activities include the fusion of SCATSAT-1 with the help of daily-based MODIS datasets or products that can solve multispectral information problems. It is also expected that the future mission or products of scatterometer will also improve the spatial resolution, noise reduction, and add more features. The full potential of SCATSAT-1 is yet to be explored in different crop types with the association of advanced models such as deep learning models.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous referees and Editor for their constructive suggestions. They also would like to thank the Meteorological & Oceanographic Satellite Data Archival Centre MOSDAC), Indian Space Research Organization (ISRO), Government of India, for providing the SCATSAT-1 data. The authors also would like to thank the SERB, DST for research fellowship.

REFERENCES

[1] Attema, E.P.W., Ulaby, F.T., 1978. Vegetation is modeled as a water cloud. Radio Science 13, 357–364. doi:10.1029/RS013i002p00357

[2] Ballantyne, J., Long, D.G., 2002. A multidecadal study of the number of Antarctic icebergs using scatterometer data. International Geoscience and Remote Sensing Symposium (IGARSS) 5, 3029–3031. doi:10.1109/igarss.2002.1026859

[3] Bartsch, A., 2010. Ten years of Sea Winds on QuikSCAT for snow applications. Remote Sensing 2, 1142–1156. doi:10.3390/rs2041142

[4] Bhowmick, S.A., Cotton, J., Fore, A., Kumar, R., Payan, C., Rodríguez, E., Sharma, A., Stiles, B., Stoffelen, A., Verhoef, A., 2019. An assessment of the performance of ISRO's SCATSAT-1 Scatterometer. Current Science 117, 959–972. doi:10.18520/cs/v117/i6/959-972

[5] Brocca, L., Ciabatta, L., Moramarco, T., Ponziani, F., Berni, N., Wagner, W., 2016. Use of Satellite Soil Moisture Products for the Operational Mitigation of Landslides Risk in Central Italy. In: Satellite Soil Moisture Retrieval. Elsevier, pp. 231–247. doi:10.1016/B978-0-12-803388-3.00012-7

[6] Chaube, N.R., Chaurasia, S., Tripathy, R., Pandey, D.K., Misra, A., Bhattacharya, B.K., Chauhan, P., Yarakulla, K., Bairagi, G.D., Srivastava, P.K., Teheliani, P., Ray, S.S., 2019. Crop phenology and soil moisture applications of SCATSAT-1. Current Science 117, 1022–1031. doi:10.18520/cs/v117/i6/1022-1031

[7] Chelton, D.B., Schlax, M.G., Samelson, R.M., 2007. Summertime Coupling between Sea Surface Temperature and Wind Stress in the California Current System. Journal of Physical Oceanography 37, 495–517. doi:10.1175/JPO3025.1

[8] Dubois, P.C., Zyl, J. van, Engman, T., 1995. Measuring soil moisture with imaging radars. IEEE Transactions on Geoscience and Remote Sensing 33, 915–926. doi:10.1109/36.406677

[9] Fung, A.K., Li, Z., Chen, K.S., 1992. Backscattering from a randomly rough dielectric surface. IEEE Transactions on Geoscience and Remote Sensing 30, 356–369. doi:10.1109/36.134085

[10] Gaur, P., Tahlani, P., Tripathy, R., Bhattacharya, B.K., Ray, S.S., 2019. Identification of rice crop phenology using Scatsat-1 Ku-band scatterometer in Punjab and Haryana. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives 42, 549–555. doi:10.5194/isprs-archives-XLII-3-W6-549-2019

[11] Gupta, P.K., Pradhan, R., Singh, R.P., Misra, A., 2019. Scatterometry for land hydrology science and its applications. Current Science 117, 1014–1021. doi:10.18520/cs/v117/i6/1014-1021

[12] Jaiswal, N., Kumar, P., Kishtawal, C.M., 2019. SCATSAT-1 wind products for tropical cyclone monitoring, prediction and surface wind structure analysis. Current Science 117, 983–992. doi:10.18520/cs/v117/i6/983-992

[13] Johny, C.J., Singh, S.K., Prasad, V.S., 2019. Validation and Impact of SCATSAT-1 Scatterometer Winds. Pure and Applied Geophysics 176, 2659–2678. doi:10.1007/s00024-019-02096-5

[14] Kerr, Y.H., Waldteufel, P., Richaume, P., Wigneron, J.P., Ferrazzoli, P., Mahmoodi, A., Bitar, A. Al, Cabot, F., Gruhier, C., Juglea, S.E., Leroux, D., Mialon, A., Delwart, S., 2012. The SMOS Soil Moisture Retrieval Algorithm. IEEE Transactions on Geoscience and Remote Sensing 50, 1384–1403. doi:10.1109/TGRS.2012.2184548

[15] Kumar, P., Gairola, R., 2019. Impact of SCATSat-1 Retrieved Wind Vectors on Short Range WRF Model Predictions over the South-Asia Region Key Points, Journal of Geophysical Research : Atmospheres. doi:10.1029/2019JD030642

[16] Kumar, R., Bhowmick, S.A., Chakraborty, A., Sharma, A., Sharma, S., Seemanth, M., Gupta, M., Chakraborty, P., Modi, J., Misra, T., 2019. Post-launch calibration-validation and data quality evaluation of SCATSAT-1. Current Science 117, 973–982. doi:10.18520/cs/v117/i6/973-982

[17] Long, D.G., 2017. Polar Applications of Spaceborne Scatterometers. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10, 2307–2320. doi:10.1109/JSTARS.2016.2629418

[18] Long, D.G., Drinkwater, M.R., 1999. Cryosphere Applications of NSCAT Data. IEEE Transactions on Geoscience and Remote Sensing 37, 1662–1670. doi:10.1109/36.763285

[19] Mandal, S., Sil, S., Shee, A., Swain, D., Pandey, P.C., 2018. Comparative Analysis of SCATSat-1 Gridded Winds with Buoys, ASCAT, and ECMWF Winds in the Bay of Bengal. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 11, 845–851. doi:10.1109/JSTARS.2018.2798621

[20] Mankad, D., Sikhakolli, R., Kakkar, P., Saquib, Q., Agrawal, K.M., Gurjar, S., Jain, D.K., Ramanujam, V.M., Thapliyal, P., 2019. SCATSAT-1 Scatterometer data processing. Current Science 117, 950–958. doi:10.18520/cs/v117/i6/950-958

[21] Maurya, A.K., Murugan, D., Singh, D., 2021. An approach for soil moisture estimation using urban and vegetation fraction cover from coarse resolution Scatsat-1 data. Advances in Space Research 68, 1329–1340. doi:10.1016/j.asr.2021.03.022

[22] Misra, T., Chakraborty, P., Lad, C., Gupta, P., Rao, J., Upadhyay, G., Vinay Kumar, S., Saravana Kumar, B., Gangele, S., Sinha, S., Tolani, H., Vithani, V.K., Raman, B.S., Rao, C.V.N., Dave, D.B., Jyoti, R., Desai, N.M., 2019. SCATSAT-1 Scatterometer: An improved successor of OSCAT. Current Science 117, 941–949. doi:10.18520/cs/v117/i6/941-949

[23] Mladenova, I., Lakshmi, V., Walker, J.P., Long, D.G., Jeu, R. De, 2009. An assessment of QuikSCAT Ku-band scatterometer data for soil moisture sensitivity. IEEE Geoscience and Remote Sensing Letters 6, 640–643.

[24] Murugan, D., Maurya, A.K., Garg, A., Singh, D., 2019. A Framework for High-Resolution Soil Moisture Extraction Using SCATSAT-1 Scatterometer Data. IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India) 4602. doi:10.1080/02564602.2019.1575293

[25] Naeimi, V., Wagner, W., 2010. C-band Scatterometers and Their Applications. In: Geoscience and Remote Sensing New Achievements. InTech, pp. 229–246. doi:10.5772/9102

[26] Nghiem, S. V., Tsai, W.Y., 2001. Global snow cover monitoring with spaceborne Ku-band scatterometer. IEEE Transactions on Geoscience and Remote Sensing 39, 2118–2134. doi:10.1109/36.957275

[27] Oh, Y., Sarabandi, K., Ulaby, F.T., 1992. An empirical model and an inversion technique for radar scattering from bare soil surfaces. IEEE Transactions on Geoscience and Remote Sensing 30, 370–381. doi:10.1109/36.134086

[28] Oveisgharan, S., Haddad, Z., Turk, J., Rodriguez, E., Li, L., 2018. Soil moisture and vegetation water content retrieval using QuikSCAT data. Remote Sensing 10, 636. doi:https://doi.org/10.3390/rs10040636

[29] Oza, S.R., Bothale, R. V., Ram Rajak, D., Jayaprasad, P., Maity, S., Thakur, P.K., Tripathi, N., Chouksey, A., Bahuguna, I.M., 2019. Assessment of cryospheric parameters over the Himalaya and Antarctic regions using SCATSAT-1 enhanced resolution data. Current Science 117, 1002–1013. doi:10.18520/cs/v117/i6/1002-1013

[30] Palakuru, M., Yarrakula, K., Chaube, N.R., Sk, K.B., Satyaji Rao, Y.R., 2019. Identification of paddy crop phenological parameters using dual polarized SCATSAT-1 (ISRO, India) scatterometer data. Environmental Science and Pollution Research 26, 1565–1575. doi:10.1007/s11356-018-3692-5

[31] Ratheesh, S., Chaudhary, A., Agarwal, N., Sharma, R., 2019. Role of ocean dynamics on mesoscale and sub-mesoscale variability of Ekman pumping for the Bay of Bengal using SCATSAT-1 forced ocean model simulations. Current Science 117, 993–1001. doi:10.18520/cs/v117/i6/993-1001

[32] Remund, Q.P., Long, D.G., 2014. A Decade of QuikSCAT Scatterometer Sea Ice Extent Data. IEEE Transactions on Geoscience and Remote Sensing 52, 4281–4290. doi:10.1109/TGRS.2013.2281056

[33] Singh, S., Tiwari, R. K., Sood V., 2021. Cloud removal for satellite image using fusion of SCATSAT-1and MODIS data. In: 3rd Conference of the Arabian Journal of Geosciences.

[34] Singh, R.K., Singh, K.N., Maisnam, M., P., J., Maity, S., 2018. Antarctic Sea Ice Extent from ISRO's SCATSAT-1 Using PCA and An Unsupervised Classification. Proceedings 2, 340. doi:10.3390/ecrs-2-05153

[35] Singh, R.K., Singh, K.N., Maisnam, M., P, J., Maity, S., 2019a. Observing Larsen C ice-shelf using ISRO's SCATSAT-1 data. Polar Science 19, 57–68. doi:10.1016/j.polar.2018.12.007

[36] Singh, S., Tiwari, R.K., 2021. Detection of Cryospheric Parameters with Artificial Neural Network over Antarctic Region using Ku-Band based ISRO's SCATSAT-1 data. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. IEEE, pp. 435–438. doi:10.1109/IGARSS47720.2021.9555088

[37] Singh, S., Tiwari, R.K., Sood, V., 2020a. Estimation and Validation of Enhanced Resolution Brightness Temperature Products of SCATSAT-1. In: 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA). IEEE, pp. 758–762. doi:10.1109/ICCCA49541.2020.9250718

[38] Singh, S., Tiwari, R.K., Gusain, H.S., Sood, V., 2020b. Potential applications of SCATSAT-1 satellite sensor: A systematic review. IEEE Sensors Journal 20. doi:10.1109/JSEN.2020.3002720

[39] Singh, S., Tiwari, R.K., Gusain, H.S., Sood, V., 2020c. Potential Applications of SCATSAT-1 Satellite Sensor: A Systematic Review. IEEE Sensors Journal 20, 12459–12471. doi:10.1109/JSEN.2020.3002720

[40] Singh, S., Tiwari, R.K., Sood, V., Prashar, S., 2021a. Fusion of SCATSAT-1 and optical data for cloud-free imaging and its applications in classification. Arabian Journal of Geosciences 14, 1978. doi:10.1007/s12517-021-08359-7

[41] Singh, S., Tiwari, R.K., Sood, V., Gusain, H.S., Prashar, S., 2021b. Image-Fusion of Ku-band based SCATSAT-1 and MODIS data for Cloud-free Change Detection over Western Himalayas. IEEE Transactions on Geoscience and Remote Sensing 1–13. doi:10.1109/TGRS.2021.3123392

[42] Singh, S., Tiwari, R.K., Sood, V., Prashar, S., 2021c. Unsupervised Snow Cover Classification Using Dual-Polarized SCATSAT-1 Satellite Data BT - Soft Computing and Signal Processing. In: Reddy, V.S., Prasad, V.K., Wang, J., Reddy, K.T. V (Eds.), . Springer Singapore, Singapore, pp. 627–635. doi:https://doi.org/10.1007/978-981-33-6912-2_57

[43] Singh, U., Srivastava, P.K., Pandey, D.K., Chaurasia, S., Gupta, D.K., Chaudhary, S.K., Prasad, R., Raghubanshi, A.S., 2019b. ScatSat-1 Leaf Area Index Product: Models Comparison, Development, and Validation Over Cropland. IEEE Geoscience and Remote Sensing Letters PP, 1–5. doi:10.1109/lgrs.2019.2927468

[44] Singh, U.S., Singh, R.K., 2020. Application of maximum-likelihood classification for segregation between Arctic multi-year ice and first-year ice using SCATSAT-1 data. Remote Sensing Applications: Society and Environment 18, 100310. doi:10.1016/j.rsase.2020.100310

[45] Sood, V., Gusain, H.S., Gupta, S., Singh, S., Kaur, S., 2020. Evaluation of SCATSAT-1 data for snow cover area mapping over a part of Western Himalayas. Advances in Space Research 66, 2556–2567. doi:10.1016/j.asr.2020.08.017

[46] Tripathy, R., Bhattacharya, B.K., 2021. Exploring Use of KU-Band Scatterometer Data from SCATSAT-1 for Crop Monitoring in India, a Case Study for Jute Crop. In: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. pp. 431–434.

[47] Tripathy, R., Bhattacharya, B.K., Tahlani, P., Gaur, P., Ray, S.S., 2019. Rice grain yield estimation over some Asian countries using ISRO's SCATSAT-1 Ku-band scatterometer data. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives 42, 257–262. doi:10.5194/isprs-archives-XLII-3-W6-257-2019

[48] Ulaby, F.T., Moore, R.K., Fung, A.K., 1982. Microwave remote sensing: Active and passive. Volume 2-Radar remote sensing and surface scattering and emission theory.

[49] Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa-Saldaña, J., Rosnay, P. de, Jann, A., Schneider, S., Komma, J., Kubu, G., Brugger, K., Aubrecht, C., Züger, J., Gangkofner, U., Kienberger, S., Brocca, L., Wang, Y., Blöschl, G., Eitzinger, J., Steinnocher, K., 2013. The ASCAT Soil Moisture Product: A Review of its Specifications, Validation Results, and Emerging Applications. Meteorologische Zeitschrift 22, 5–33. doi:10.1127/0941-2948/2013/0399

[50] Wang, N.-Y., Vesecky, J.F., 1999. Sea surface temperature estimation using active/passive microwave remote sensing. In: IEEE 1999 International Geoscience and Remote Sensing Symposium. IGARSS'99 (Cat. No. 99CH36293). pp. 971–973.

REVIEW OF DIFFERENT TECHNIQUES TO PREDICT HEART DISEASE WITH ML ALGORITHMS

Savia [#1], Harpreet Kaur[#2]
Department of Computer Science and Engineering, Punjabi University, Patiala., Department of Computer Science and Engineering, Punjabi University, Patiala.
singlasavia37@gmail.com
harpreet.ce@pbi.ac.in

ABSTRACT:– Heart disease is a serious disease affecting a huge number of people all around the world. In today's healthcare industry, heart disease is a serious issue. Machine learning is playing a vital role to predict number of diseases such as heart disease, tumor detection etc. There exist numerous machine learning algorithms to predict Heart Disease. Some state-of-the-art machine learning algorithms based on literature have been compared to predict heart disease like support vector machine, Naïve bayes, Artificial neural network and Random forest. These algorithms give good parameters with accuracy, sensitivity, specificity, precision.

KEYWORDS: —Heart disease prediction, support vector machine, artificial neural network, naïve bayes, and Random forest.

I. INTRODUCTION

A large number of hospitals and health care centres have sprung up as a result of increased awareness and technology in the field of health care. However, in undeveloped nations, offering higher health care at an affordable cost remains a challenge. Despite the fact that many countries have taken concrete steps to provide healthcare, the supply of these services to the poor and needy remains a question mark. Unimaginable services, such as the separation of twin births and the discovery of new treatments for terrible diseases, are occurring in another world. However, cases with poor clinical diagnosis and treatment continue to be reported. According to a WHO research, heart attacks and strokes cause for 17 million deaths worldwide. Work overload, mental stress, and a number of other problems contribute to heart disease deaths in many countries. In general, it is discovered to be the leading cause of death in adults. Diagnosis is a difficult and important activity that must be completed correctly and quickly. The majority of the time, clinical judgments are made based on the doctor's knowledge and experience. All doctors do not have the same level of experience or expertise. Information systems, Decision Support Systems, and Image and Scan Processing Systems are all available in hospitals, although not all hospitals have them, and their applications are limited. It may have happened so often that you or somebody yours need specialists help right away, in any case, they are not accessible because of some reasons. This system which is used for health prediction is an end user support system online consultation project. This system take care of different symptoms related with those the infection/disease. The system permits client to share their symptoms and issues. Here we utilize some data mining strategies to figure the most accurate disease that could be related with patient's symptoms. In the event that the system can't give reasonable outcomes, it illuminates the user about the kind regarding sickness is or disorder it feels client's symptoms are related with [9].

Data mining has played an essential role in heart disease research in recent decades. A notable and powerful technique in the study of heart disease classification is to identify hidden medical information from the different expression between healthy and heart disease persons in the existing clinical data. Patients' treatment is based on the classification of their heart disease. Statistics and machine learning are two key methodologies that have been used to identify the condition of heart disease based on clinical data expression. Data Mining mentions the variety of techniques that recognize the purpose of information, decision-making knowledge in the database. In the industry of healthcare vast amount of information is rack up that unfortunately is not “mined” to find the unseen data i.e., supportive in decision making. The search for connections and global patterns that exist in massive databases but are hidden within large amounts of data is referred to as data mining. Data mining, which is an important part of Knowledge Discovery, is the process of converting data into knowledge to improve decision-making. Data cleansing, data integration, data selection, data mining pattern detection, and knowledge presentation are all steps in the Knowledge Discovery process. A variety of expertise have utilized various classifiers e.g., Naïve Bayes, Support Machine Vector, Random Forest. By using dataset with some attributes such as age, sex, blood pressure, blood sugar, and a variety of other factors can all be used to predict the probability of a patient suffering heart disease. It can be served as training tool to train nurses and medical students to diagnose patients with heart disease [14].

II. RELATED WORK

Various analysts are doing investigation in clinical information examination to anticipate different sicknesses by utilizing diverse machine learning algorithms. This part examines about study in cardiac arrest from different authors.

E. Anupriya et al. (2010) implemented a prediction system using some of the ML algorithms like naïve bayes, decision trees, classification with clustering to predict the heart disease with different accuracy of 96.5%, 99.2%and 88.3%[1].

AH Chen et al. (2011) has enlarge the heart disease prediction system by using Artificial Neural Network to fulfil it with good prediction probability with Accuracy of 80%, sensitivity is 85% and specificity is 70%[2].

Mrs. G. Subbalakshmi (2011) has worked on the system where naïve bayes algorithm is used for the appropriate results [3].

As here Chaitrali S. Dangare (2012) studied on the algorithms of ML i.e., naïve bayes, decision trees, neural networks that gives approximate results for accuracy 94.44%, 96.66%, 99.62% which shows neural network have better results than other algorithms [4].

Dhanashree S. Medhekar et al. (2013) implemented Cleveland's Dataset that data frame with 303 observation with naïve bayes of 88.96% accuracy [5].

R. Chitra et al. (2013) developed a heart disease prediction system to predict heart disease using support machine vector and cascaded neural network to perform with good accuracy, sensitivity and specificity [6].

Purushottam et al. (2016) implemented efficient search for diagnosis of heart disease comparing association rules with Support Vector Machine [7].

Karthikeyan Harimoorthy et al. (2019) implement a multi-disease prediction model for heart prediction random forest, decision tree to perform with good accuracy sensitivity and specificity [8].

V. Jackins et al. (2020) developed an Intelligent Heart Disease Prediction System to predict the heart disease using these classifiers naïve bayes and Random Forest to perform with good accuracy of 97% and 60% [9].

N.Shabaz Ali et al (2020) worked on a smart health care system by using some of the ml algorithms svm, random forest, naïve bayes, knn with good accuracy of 92%, 97%, 82% and 92% [10].

Yuanyuan pan et al. (2020) implement an enhanced deep learning for heart prediction ann, dnn, rnn to perform with accuracy for 50 attributes [11].

Mohammed Jawwad Ali Junaid (2020) used naïve bayes, ann, svm algorithms of machine learning to detect heart disease for good results i.e., accuracy, specificity, sensitivity [12].

Our approach would be another search for efficient diagnosis.

III. DATA SET

A. Data Source

There are numerous state-of-the-art datasets are available in literature as show in table1 various datasets are used in history for the heart disease prediction system.

TABLE I
RECORD OF DATASET FROM LITERATURE REVIEW.

Data set	Year/Reference	No. of records	No. of attributes
Sellapan	2010[1]	909	13 attributes are used but it is reduced to 6 genetic search
Statlog	I. 2012[4] II. 2020[18]	270	I.(13 input attributes for appropriate result) II. (79 raw attributes although only 13 attributes are used)
UCI	2011[2]	303	14 attributes
	2013[6]	270	76 attributes but references to use 13,8 symbolic ad 6 numeric attributes
Framingham	2020[9]	-	13 attributes are used
Cleveland	2011[3]	-	15 attributes
	2012[4]	573	13input attributes for appropriate result two important attributes are added
	2013[5]	303	14 attributes
	2016[7]	303	75 attributes using only 14 of them
	2020[13]	303	76 attributes are included but only 13 attributes are studied
	2020[14]	303	75 attributes using only 14 of them
	2021[16]	303	14 features are used where 8 are categorical and 6 are numeric
	2019[17]	303	76 attributes are provided but only 14 attributes are used
	2020[18]	270	13 attributes

In this study, the Cleveland Heart Disease dataset is used for testing purposes. There were 303 instances and 75 attributes when this informational collection was planned, but all published studies refer to using a subset of 14 of them [14].

B. Features Description of Dataset Referred In Literature.

In Table2 the description of a dataset from Cleveland's dataset for the most part authors used it.

TABLE II
SHOW THE 14 CLINICAL FEATURE AND THEIR DESCRIPTION

S. No	Feature Name	Feature Code	Description
1	Age	AGE	Age in Years
2	Sex	SEX	Male = 1, Female = 0
3	Chest Pain	CPT	Atypical angina=1
			Typical angina=2
			Asymptomatic = 3
			Non-Anginal Pain = 4
4	Resting Blood Pressure	RBP	Mm hg,hospitalized
5	Serum Cholesterol	SCH	In mg/dl
6	fastingbloodsugar > 120mg/dl	FBS	fastingbloodsugar > 120mg/dl
			(T=1)
7	Resting Electrocardiographic	RES	Normal=0
			ST T =1
			Hypertrophy = 2
8	Maximum Heart Rate	MHR	-
9	Exercise Induced Angina	EIA	Yes = 1
			No = 0
10	Old Peak=ST depression induced by exercise relative to rest	OPK	-
11	The slope of the Peak Exercise ST Segment	PES	Up Sloping=1
			Flat=2
			Down Sloping=3
12	Number of major vessels (0-3) coloured by fluoroscopy	VCA	
13	Thallium Scan	THA	Normal = 3
			Fixed defect=6
			Reverseible defect=7
14	Label	LB	Heart disease patient=1
			Healthy=0

IV. STATE-OF-THE-ART ALGORITHMS USED FOR HEART PREDICTION SYSTEM

Data Mining is a process of discovering analysing different data patterns from large raw datasets. The goal of data mining is to extract the relevant information from comprehensive dataset. This is in the form of bundle of package such as machine learning, statistics and database system. All these factors determine the efficiency in Knowledge Discovery in database process. KDD consist of various process such as data cleaning, data selection, data integration, data transformation, data pattern searching and finally knowledge representation. The data mining technique that mainly used are Association rule, Clustering, Classification, regression etc.

- The association rule can be used to establish relationship between two variable.
- The clustering is a process of grouping the structures based on similarity between them.
- The classification is assigning items in collection to target datasets.
- The regression tries to estimate the various mode to find the relation between data with least error.
- There are two methodologies of machine learning is used in the data mining process.
  They are Supervised learning and unsupervised learning.

Supervised learning: In supervised learning the system trains itself by the given input and learn to generate the result.

Unsupervised learning: In unsupervised learning the hidden structure and relation among the dataset is found out.

In medical industry, disease prediction is done by data mining along with machine learning. There are different classification models are used i.e., Decision trees, Artificial neural networks, Support vector machines and K-nearest neighbours. Many other algorithms such as Naive Bayes classifier, linear regression, Logistic Regression are also used.

A. *Support Vector Machine Algorithm (SVM)*

Support Vector Machine (SVM) is a supervised machine-learning algorithm, which can be applied for both classification and regression challenges. Even so, it is for the most part it is used in classification problems. In that matter algorithm, we plot every information thing as a point in n-dimensional space with the worth of each feature being the worth of a specific coordinate. The purpose of the SVM algorithm is to construct decision boundary that can segregate n-dimensional space into classes so we can undoubtedly put the new information point in the right classification. The decision boundary is known as hyper plane. SVM picks the outrageous focuses that assistance in making the hyper plane .These extreme cases are called as support vectors. Multiple decision boundaries can be used to segregate the classes in n-dimensional space, however, we need to discover the best decision boundary that assists with ordering the data points are known as hyper plane .The data points or vectors that are the closest to the hyper plane and which affect the position of the hyper plane are termed as support vector as shown in figure 1. [17]. SVM can solve both linear and non-linear problems as in figure1 it draws a hyper plane and by the use of hyper plane it separates the mixed data into classes as shown in figure 1

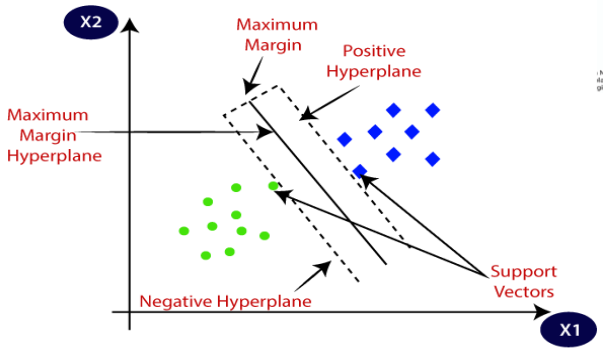


Figure. 1 Overview of Support Vector Machine [17]

In Figure 2 the flow chart of support vector machine shows how some random mixed data is classified into four classes by using support machine vector algorithm.

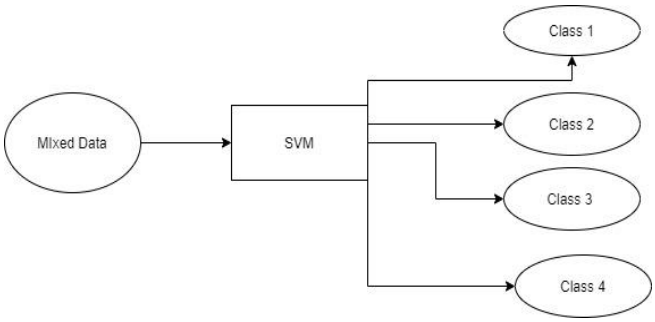


Figure. 2 Flowchart of Support Vector Machine [23]

B. Navies Bayes Algorithm

Naïve Bayes algorithm is supervised learning algorithm. This is used for solving classification problems and is based on Bayes Theorem. Classification algorithm helps to make the fast machine learning model which males a quick predictions that it is a probabilistic classifier on the basis of the probability of an object it predicts[16]. The Naïve Bayes Algorithm is comprised of the words Naïve and Bayes i.e.

Naive:-This assumes that the occurrence of a certain feature is independent of the occurrence of other features. Example if the identification of a fruit based on the taste colour, shape and taste then orange, spherical and citrus fruit is an orange. However, the features are individually contributes to identify they are independent.

Bayes:-this is based on the principle bayes' theorem.

Bayes Theorem: - It is also known as bayes' rule, it determines the hypothesis with prior knowledge which depends on the conditional probability. The formula is used:-

P (A/B) = (P (B/A) P (A))/P (B)

Where P (A/B) is posterior probability, P (B/A) is likelihood probability, P (A) is prior probability, P (B) is marginal probability [16].

The thought behind Naïve Bayes calculation is the posterior probability of a data instance ti in a class cj of the data model. The posterior probability P (ti|cj) is the possibility of that ti can be labeled cj. P (ti|cj) can be calculated by multiplying all probabilities of all attributes of the data instance in the data model:

$$P(t_i/c_j)=\prod_{k=1}^{p} P(x_{ik}\ /c_j)$$

P= no. of attributes in each data instance. The posterior probability is for classes and the highest probability will be instances label as shown in Figure3.

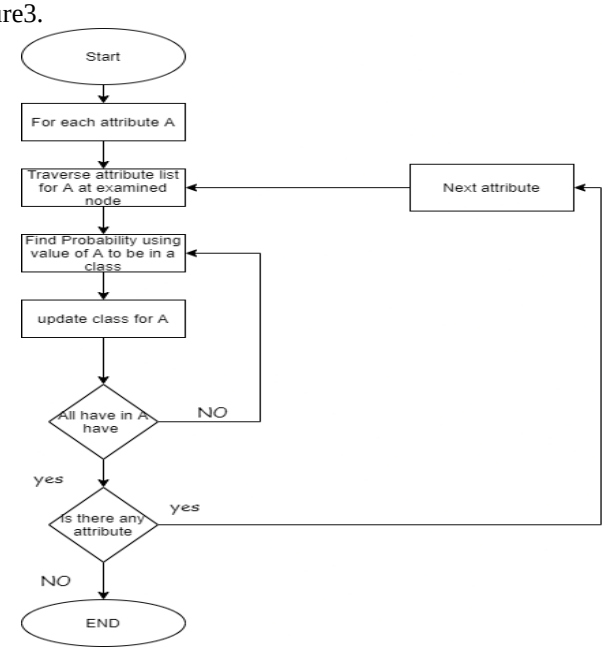


Figure 3: Flowchart of Naive Bayes [25]

C. Artificial Neural Network Algorithm (ANN)

Artificial Neural Network is the term that is derived from the neural networks in biology that develop the human brain structure as the neurons in human are connected to each other the neurons in artificial neural networks also are connected in the form of various network layers. There is relationship between biological and artificial neural network as the terms shown in table3 represents inputs, nodes, weights and output that are playing the role in the biological neural networks as dendrites, cell nucleus, synapse, axon.

TABLE. III
RELATIONSHIP BETWEEN BIOLOGICAL AND ARTIFICIAL NEURAL NETWORK. [15]

Biological Neural Network	Artificial Neural Network
Dendrites	Input
Cell Nucleus	Nodes
Synapse	Weights
Axon	Output

As the Figure4 shows that the nodes, inputs, weight works to get an appropriate output. Also ANN is consist of three layers:

- Input Layer
- Hidden Layer
- Output Layer
- Input Layer: As its name it has accepts inputs in many different formats provided by programmer.
- Hidden Layer: It is in between input layer and output layer. It perform the calculations to find the patterns.
- Output Layer: Using the hidden layer input layer goes through a series of transformations which results in output.

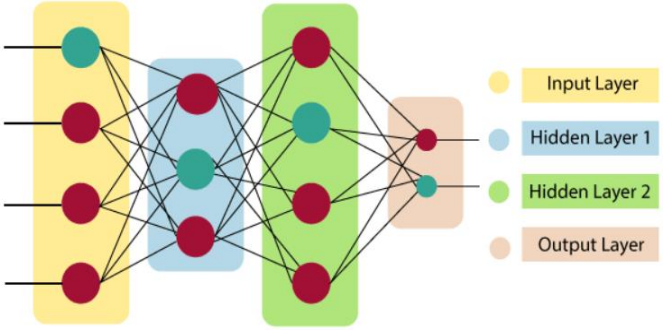


Figure. 4 Artificial Neural Network [15]

1) Working Of Artificial Neural Network: The best way to represent ANN is by a weighted directed graph. The ANN receives the input signal from external source in the form of patterns and these inputs are assigned by the notation x (n) [15]. The input is fed to input layer to each neuron in it, neurons of one layer are connected to neuron of the next layer through channels and these channels are assigned as weights, the input are multiple to the corresponding weights and the sum is sent as input to neurons in the hidden layer. Each of these neurons is associated with a numerical value called bias. This value is passed through a threshold function called activated function. It activates neurons and the activated neuron transmits the data to neurons to the next layer over the channels. The neuron with highest value determines the output as shown in Figure 5.

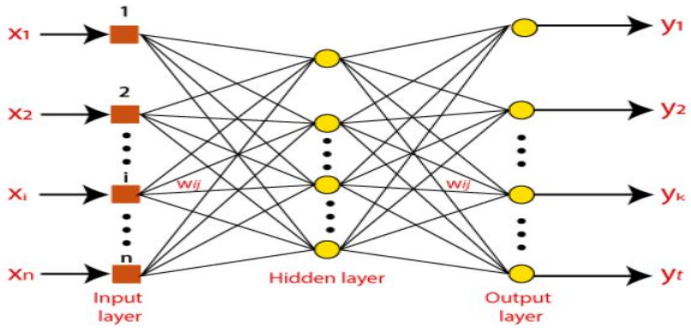


Figure. 5 Working of Artificial Neural Network [15]

D. Random Forest Algorithm

Random Forest is well known machine learning algorithm that is used with supervised learning technique. This algorithm solves the problems of both classification and regression. Random Forest

is a classifier that uses a number of decision trees on different subsets of a dataset and averages the results to increase the dataset's predictive accuracy. Instead of relying on a single decision tree, the random forest takes the prediction from each tree and predicts the final output based on the majority votes of predictions. The higher the number of trees in the forest, the better the accuracy and control of the issue of over fitting . Here is a limitation of this algorithm as it can be used for both classification and regression, though it is better for classification as compare to regression [18].

1) Uses of Random Forest: The uses of random forest are shown below

- This algorithm use less time for training.
- Random Forest can predict output with high accuracy even with for huge data set it runs efficiently.
- Accuracy can be maintained even with huge amount of missing data.

2) Working of Random Forest: Random forest creates the combination of n decision trees and make predictions of every tree. There are three components in decision tree i.e. decision nodes, leaf nodes, and root node. It divides a training dataset into branches, which is divided into different branches This series continues until the leaf node is reached, at which moment it becomes impossible to divide because it is the final node. The nodes represent characteristics that are used to predict the outcome. Figure 6 shows a link between these decision nodes to the leave.

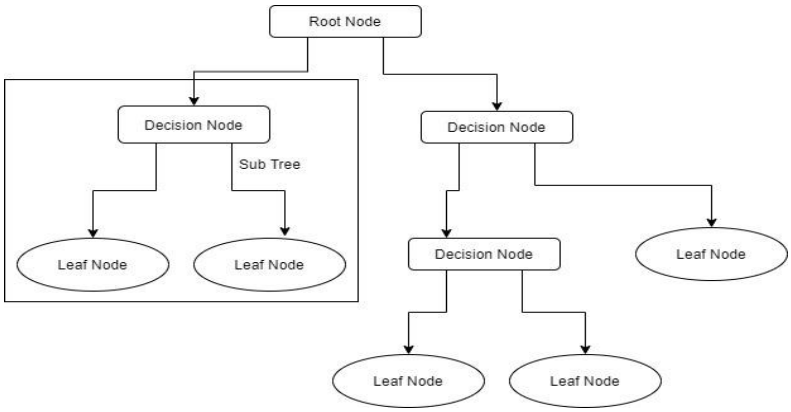


Figure. 6 Working of Random Forest [24]

V. FINDING FROM LITERATURE

The following table4 shows different parameters, datasets and techniques which are computed by different researchers to predict the heart disease and in table5 the values of parameters are shown by techniques and disease.

TABLE IV
REVIEW OF DIFFERENT PARAMETERS IN LITERATURE TO PREDICT HEART DISEASE.

Ref. No.	Author	Title	Datasets	Techniques used	Disease	Parameters
1.	E.Anupriya et al.	Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,2010 (RESEARCHGATE)	13 attribute dataset by Sellapan et al (2008)	Naïve Bayes, Decision Tress	Heart Disease	Accuracy
2.	AH Chen et al.	HDPS: Heart Disease Prediction System,2011(IEEE)	UCI	Artificial neural network (ANN)	Heart Disease	Accuracy, Sensitivity, Specificity
3.	Mrs.G.Subbalakshmi	Decision Support in Heart Disease Prediction System using Naive Bayes,2011(IJCSE)	Cleveland Heart Disease database	Naive Bayes	Heart Disease	It is particularly suited when the dimensionality of the inputs is high. so, it can often outperform more sophisticated classification methods
4.	Chaitrali S. Dangare	Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, 2012 (IJCA)	Cleveland 's database	Naïve Bayes, Decision Tress, Neural Networks	Heart Disease	Accuracy
5.	Dhanashree S. Medhekar et al.	Heart Disease Prediction System using Naive Bayes,2013(IJERSTE)	Cleveland's Dataset	Naïve Bayes	Heart Disease	Accuracy
6.	R. Chitra et al.	Heart Disease Prediction System Using Supervised Learning Classifier,2013 (BONFRING)	UCI	Cascaded Neural Network SVM	Heart Disease	Accuracy, Sensitivity, Specificity
7.	Purushottam et al.	Efficient Heart Disease Prediction System,2016 (ELSEVIER)	Data collected from UCI, Cleveland Clinic Foundation	Support Vector Machine (SVM)	Heart Disease	Accuracy
8.	Karthikeyan Harimo orthy et al.	Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system, 2019(SPRINGER)	-	Random Forest, Decision Tree	Heart Disease	Accuracy, Sensitivity, Specificity
9.	V. Jackins et al.	AI-based smart prediction of clinical disease using random forest classifer and Naive Bayes, 2020 (SPRINGER)	Framingham 's Dataset	Naïve Bayes,	Diabetes, Heart Disease, Cancer	Accuracy
10.	N. Shabaz Ali et al.	Prediction of Diseases in Smart Health Care System using Machine Learning, 2020(IJRTE)	-	SVM Neural networks, Random forest, Naive Bayes, K-Nearest Neighbour(KNN)	Mainly Use for Heart Disease but also use for other diseases	Accuracy
11.	Yuanyuan pan et al.	Enhanced Deep Learning Assisted Convolutional Neural Network for Heart Disease Prediction on the Internet of Medical Things Platform, 2020(IEEE)		ANN, DNN, Recurrent neural network(RNN)	Heart Disease	Accuracy
12.	Mohammed Jawwad Ali Junaid	Data Science And Its Application In Heart Disease Prediction,2020(ICIEM)		Naïve Bayes ANN SVM	Heart Disease	Accuracy, Specificity, Sensitivity
13.	Syed Arslan Ali et al.	An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzo–Tompa and Stacked Genetic Algorithm,2020(IEEE)	Cleveland heart disease dataset is used	Artificial Neural Network (ANN) Deep Neural Network (DNN)	Heart Disease	Sensitivity, Specificity, Precision, Accuracy
14.	Jian Ping Li et al.	Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,2020(IEEE)	Cleveland heart disease dataset 2016	K-Nearest Neighbours (KNN) Artificial Neural network(ANN)	Heart Disease	Accuracy Specificity, Sensitivity

19.	C.Sowmiya et al.	A hybrid approach for mortality prediction or heart patients using ACO-HKNN,2020 (SPRINGER)	Cleveland Dataset	SVM Naïve Bayes KNN Decision Tree Hybrid KNN	Heart Disease	Accuracy, Precision
20.	Pooja Rani et al.	A decision support system for heart disease prediction based upon machine learning,2021 (SPRINGER)	Cleveland Dataset	Naive Bayes SVM Logistic regression Random Forest	Heart Disease	Accuracy, Sensitivity, Specificity, Precision
21.	Senthilkumar mohan et al.	Effective heart disease prediction using hybrid machine learning techniques, 2019 (IEEE)	Cleveland Dataset	Decision Trees SVM Random Forest Naïve Bayes	Heart Disease	Accuracy, Precision, Sensitivity, Specificity
22.	Norma Latif Fitriyani et al.	HDPM: An Effective heart disease predication model for a clinical decision support system, 2020 (IEEE)	Statlog and Cleveland's datasets	Naïve Bayes SVM Decision Tree Random Forest	Heart Disease	Accuracy, Precision

The Table 5 shows the values of the parameters from the table4 according to the techniques and disease there are some parameters other than heart disease.

TABLE V
PARAMETRIC VALUES OF LITERATURE REVIEW

Ref. No.	Disease	Technique	Accuracy	Precision	Sensitivity	Specificity
1.	-	Naïve Bayes	96.5%	-	-	-
		Decision Tree	99.2%	-	-	-
2	-	ANN	80%	-	85%	70%
3.	-	Naïve Bayes	-	-	-	-
4.	-	Naïve Bayes	94.44%	-	-	-
		Decision Tree	96.66%	-	-	-
		Neural Network	99.62%	-	-	-
5.	-	Naïve Bayes	88.96%	-	-	-
6.	-	Naïve Bayes	88.96%	-	-	-
		SVM	82%	-	85%	77.5%
7.	-	SVM	70.59%	-	-	-
8.	-	Random Forest	82%	-	80.5%	83.3%
		Decision Tree	73%	-	71.1%	74.5%
9.	Diabetes	-	92%	-	-	-
	Heart	-	97%	-	-	-
	Disease	-	94%	-	-	-
	Cancer					
10.	-	SVM	92%	-	-	-
	-	Neural Network	97%	-	-	-
	-	Random Forest	94%	-	-	-
	-	Naïve Bayes	82%	-	-	-
	-	KNN	92%	-	-	-
11.	-	ANN	88.5%	-	-	-
		DNN	90.2%	-	-	-
		RNN	97.2%	-	-	-
12.	-	Naïve Bayes	82.97%	-	76.10%	76.27%
		ANN	85.30%	-	81.75%	77.73%
		SVM	86.12%	-	84.87%	79.21%
13	-	ANN	83%	82.7%	85%	81.5%
		DNN	88.1%	89.8%	86%	90%
14	-	KNN	69%	-	64%	70%
		ANN	60%	-	0%	100%
19	-	SVM	97.4%	96.9%	-	-
		Naïve bayes	96.21%	94.8%	-	-
		KNN	97.30%	94.3%	-	-
		Decision Tree	96%	95.2%	-	-
		Hybrid KNN	99.02%	97.1%	-	-
20	-	Naïve Bayes	84.79%	85.49%	80.57%	88.41%
		SVM Logistic Regression	79.50%	79.38%	74.82%	83.53%
		Random Forest	83.8%	84.61%	79.13%	87.80%
			83.83%	85.71%	77.69%	89.02%
21	-	Decision Tree	85%	86%	98.8%	0%
		SVM	86.1%	87.1%	100%	10%
		Random Forest	75.8%	90.5%	98.8%	60%
		Naïve Bayes	98.40%	98.57%	79.8%	

VI. RESULT AND DISCUSSION

There are various parameters on the basis of which different algorithms have been compared. Different parameters used in literature to predict heart disease are:

- Accuracy- Accuracy is how close a measured value is to the real or the true value.
- Sensitivity- Sensitivity is defined as the probability of correctly identifying some condition or disease state.
- Specificity- Specificity describes the characteristic of a test in terms of how well the test correctly identifies true negatives or those who do not have the predicted condition.
- Precision– Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

It has been observed from literature review that artificial neural network is the most accurate with 99.62% accuracy[4] and the specificity of ANN is also the highest that is 100%[14], support vector machine has also performed best in case of sensitivity (100% performance from[21]) and Precision (96.9% performance from[19]).It has been depicted in figure7, figure 8, figure 9 and figure 10.In figure7, figure8, figure9, figure10 i.e. NB is Naïve bayes, ANN is Artificial neural network, SVM Support vector machine, DT Decision tress, RF is Random forest, DNN Deep neural network, KNN k-nearest neighbour.

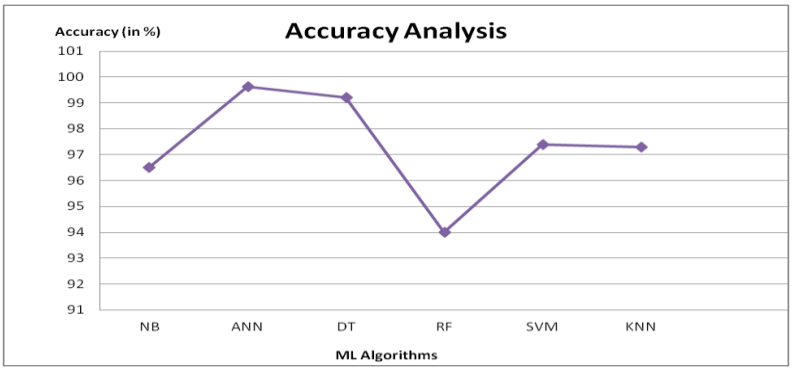


Figure. 7 Accuracy Analysis

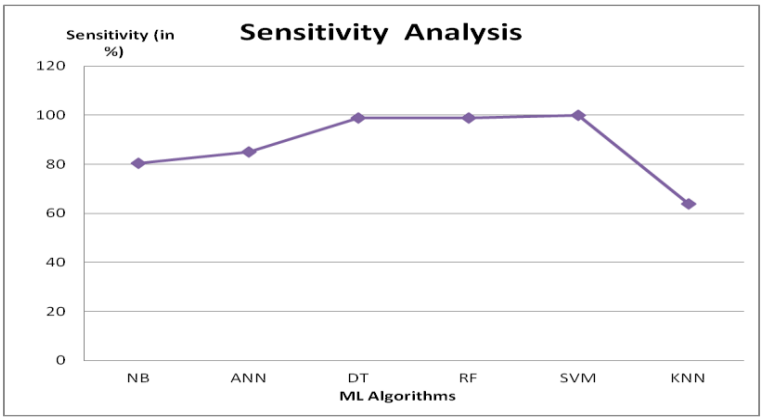


Figure. 8 Sensitivity Analysis

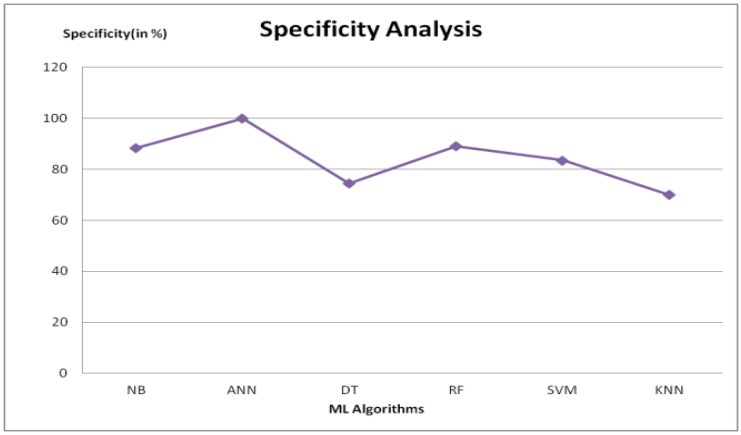


Figure.9 Specificity Analysis

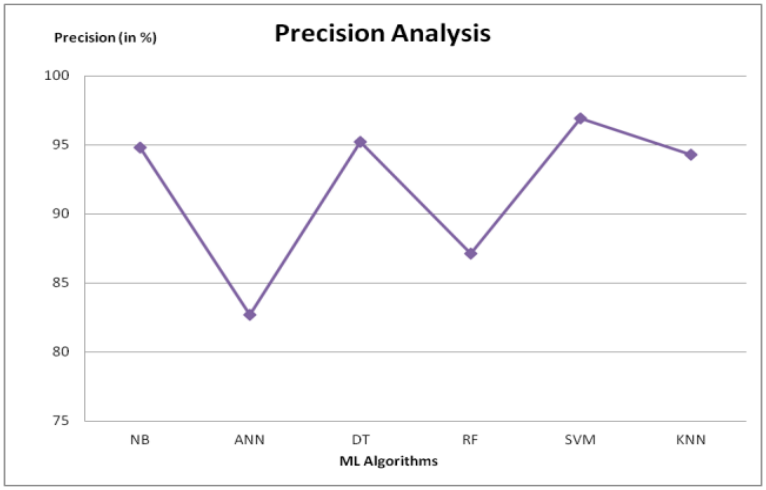


Figure.10 Precision Analysis

VII. CONCLUSION

From this study, it has been concluded that Maximum authors tested the algorithms on the "Cleveland" heart disease dataset and this dataset is used in the study available on UCI machine learning repository. The classifier artificial neural network and support vector machine has good results as compared to other classifiers as these have 100% specificity and sensitivity. In future we can use other machine learning techniques to improve the results and we can also use different datasets with more features.

REFERENCES

[1] M. Anbarasi et. al. *Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm***.** .International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376; ISSN: 0975-5462.

[2] AH Chen, SY Huang, PS Hong, CH Cheng, EJ Lin; HDPS: *Heart Disease Prediction System*; ISSN 0276-6574; Computing in Cardiology 2011;38:557-560.

[3] G.Subbalakshmi et al. **(**2011); *Decision Support in Heart Disease Prediction System using Naive Bayes;*Indian Journal of Computer Science and Engineering (IJCSE); Vol. 2 No. 2 Apr-May 2011 ;ISSN : 0976-5166.

[4] Chaitrali S. Dangare, Sulabha S. Apte, *Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques*; International Journal of Computer Applications (0975 – 888); Volume 47– No.10, June 2012

[5] Dhanashree S. Medhekar[1], Mayur P. Bote[2], Shruti D. Deshmukh[3] [2013]; *Heart Disease Prediction System using Naive Bayes* ; International Journal of Enhanced Research in Science Technology & Engineering; Vol. 2 Issue 3, March.-2013 ISSN NO: 2319-7463.

[6] R. Chitra and Dr.V. Seenivasagam; *Heart Disease Prediction System Using Supervised Learning Classifier***,** Bonfring International Journal of Software Engineering and Soft Computing, Vol. 3, No. 1, March 2013;DOI:10.9756/BIJSESC.4336

[7] Purushottam, ; Saxena, Kanak; Sharma, Richa (2016). *Efficient Heart Disease Prediction System. Procedia Computer Science, 85(), 962–969.* doi:10.1016/j.procs.2016.05.288 Karthikeyan Harimoorthy[1] Menakadevi Thangavelu[2] ; *Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system*[2019];DOI: 10.1007/s12652-019-01652-0.

[8] Harimoorthy, Karthikeyan; Thangavelu, Menakadevi (2020). *Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. Journal of Ambient Intelligence and Humanized Computing, (), –.* doi:10.1007/s12652-019-01652-0

[9] Jackins, V., Vimal, S., Kaliappan, M., & Lee, M. Y. (2020). *AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing, 77(5), 5198–5219.* doi:10.1007/s11227-020-03481-x

[10] N. Shabaz Ali;G. Divya;(2020) *Prediction of Diseases in Smart Health Care System using Machine learning*; International Journal of Recent Technology and Engineering (IJRTE) DOI:10.35940/ijrte.E6482.018520.

[11] Pan, Yuanyuan; Fu, Minghuan; Cheng, Biao; Tao, Xuefei; Guo, Jing (2020). *Enhanced Deep learning assisted Convolutional Neural Network for Heart Disease Prediction on the internet of medical things platform. IEEE Access, (), 1–1.* doi:10.1109/ACCESS.2020.3026214

[12] Ali, Syed Arslan; Raza, Basit; Malik, Ahmad Kamran; Shahid, Ahmad Raza; Faheem, Muhammad; Alquhayz, Hani; Kumar, Yogan Jaya (2020). *An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzoâ€ "Tompa and Stacked Genetic Algorithm. IEEE Access, 8(), 65947–65958.* doi:10.1109/ACCESS.2020.2985646

[13] Li, Jian Ping; Haq, Amin Ul; Din, Salah Ud; Khan, Jalaluddin; Khan, Asif; Saboor, Abdus (2020). *Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access, 8(), 107562–107582.* doi:10.1109/ACCESS.2020.3001149

[14] Java point; Artificial neural network: https://www.javatpoint.com/artificial-neural-network

[15] Java point; Naive bayes: https://www.javatpoint.com/machine-learning-naive-bayes-classifier

[16] Java point; Support machine vector:https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[17] Java point; Random forest:https://www.javatpoint.com/machine-learning-random-forest-algorithm

[18] Sowmiya, C.; Sumitra, P. (2020). *A hybrid approach for mortality prediction for heart patients using ACO-HKNN. Journal of Ambient Intelligence and Humanized Computing, (), –.* doi:10.1007/s12652-020-02027-6

[19] Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). *A decision support system for heart disease prediction based upon machine learning. Journal of Reliable Intelligent Environments, 7(3), 263–275.* doi:10.1007/s40860-021-00133-6

[20] Mohan, Senthilkumar; Thirumalai, Chandrasegar; Srivastava, Gautam (2019). *Effective Heart Disease Prediction using Hybrid Machine Learning Techniques. IEEE Access, (), 1–1.* doi:10.1109/ACCESS.2019.2923707

[21] Fitriyani, Norma Latif; Syafrudin, Muhammad; Alfian, Ganjar; Rhee, Jongtae (2020). *HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. IEEE Access, 8(), 133034–133050.* doi:10.1109/access.2020.3010511

[22] Rushikesh pupale ;(2018);Support Vector Machines(SVM) ;https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

[23] Onesmus Mbaabu*; Introduction to Random Forest in Machine Learning(*2020);https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

[24] Masud Karim; Rashedur M. Rahman;(2013); *Decision Tree and Naïve Bayes Algorithm for Classification and Generation of Actionable Knowledge for Direct Marketing* Journal of Software Engineering and Applications, http://dx.doi.org/10.4236/jsea.2013.64025

VARIOUS OPTIMIZED TECHNIQUE FOR ROUTING INWSN

Prabhjot Kaur[1], Raman Maini[2], Sumandeep Kaur[3]
[123] *Computer Engineering, Punjabi University Patiala*
Kprabhjot07@gmail.com Reasearch_raman@yahoo.com
sumandhanjal@gmail.com

ABSTRACT:— Multimedia applications have become an essential part of our daily lives, and their use is flourishing dayby day. The area of wireless sensor network is not an exception where the multimedia sensors are attracting the at- tention of the researchers increasingly, and it has shifted the focus from traditional scalar sensors to sensors equipped with multimedia devices. The multimedia sensors have the ability to capture video, image, audio, and scalar sensor data and deliver the multimedia content through sensors network. Due to the resource constraints na- ture of WSN introducing multimedia will add more challenges, so the protocols designed for multimedia wireless sensor network should be aware of the resource constraints nature of WSN and multimedia transmission require- ment. This paper discusses the design challenges of routing protocols proposed for WMSN. A survey and compre- hensive discussion are given for proposed protocols of WMSN followed by their limitations and features.

Keywords: WSN, WMSN, Routing

1. BACKGROUND

Wireless networks have experienced explosive growth during the last few years. Nowadays, there are a large variety of networks spanning from the well-known cellular networks to noninfrastructure wireless networks such as mobile ad hoc networks and sensor networks. Communication security is essential to the success of wireless sensor network applications, especially for those mission-critical applications working in unattended and even hostile environments. However, providing satisfactory security protection in wireless sensor networks has ever been a challenging task due to various network and resource constraints and malicious attacks. In Wireless Sensor Networks, these devices are called sensors or motes. Wireless Sensor Networks (WSNs) are composed of a large number of sensor nodes which are densely deployed either inside a physical phenomenon or very close to it [1]. Sensors are tiny devices which monitor various conditions like temperature, humidity, pressure etc. and later convert it into electrical signal. These sensor devices have the ability to communicate either directly to the Base Station (BS) or among each other. Each node hence requires a power source that is smart enough to give a node maxi- mum life in spite of its tiny size. The self-organizing capability of sensor nodes provides several challenges for the researchers in designing the network protocols. WSN applications can be categorized into two: monitoring and tracking [2]. The potential applications include military sensing, air traffic control, traffic surveillance, in- dustrial and manufacturing automation, environment, health, home and other commercial areas. In WSN the network layer aims in maximizing the lifetime by finding ways for energy-efficient route setup and reliable relaying of data from sensor nodes to sink. Many routing protocols have been proposed in order to solve the routing problem in WSNs. The designs of routing protocols are also affected by various factors such as deployment, energy consumption, security etc. Researchers thus focus more on designing energy efficient nodes and protocols that could support various operations.

2. RELATED STUDY

Shashidhar Rao Gandham et al. Energy consumption is a primary anxiety in Wireless Sensor Network (WSN). Sensor node batteries cannot be easily refilled and nodes have a finite lifetime. One of the major issues in wireless sensor networks is developing an energy efficient routing protocol to improve the overall lifetime of the network. Several routing protocols are used in WSN for transforming the messages. In epidemic routing the messages are blindly stored and forwarded to all neighboring nodes. This may contain packet loss and energy consumption. To avoid the packet loss and to reduce energy consumption the Dynamic Source Routing (DSR) algorithm with optimal selective forwarding protocol is used. The DSR protocol is a simple and efficient routing protocol. DSR is specifi- cally designed for Wireless sensor Network. In DSR each node (sensors) maintains a routing table. If the route is available in the table the optimal selective forwarding protocol schemes are applied. And also the forwarding schemes are designed for three different cases to improve the lifetime of sensor node batteries such as when Sensors give maximum importance for their own transmitted messages, Sensors give maximum importance for messages that arrive at sink successfully and Sensors give maximum importance for retransmitted messages to one of its neighbors.

Wei Ye et al. The use of wireless sensor networks (WSNs) has grown enormously in the last decade, pointing outthe crucial need for scalable and energy-efficient routing and data gathering and aggregation protocols in corre- sponding large-scale environments. To maximize network lifetime in Wireless Sensor Networks (WSNs) the pathsfordata transfer are selected in such a way that the total energy consumed along the path is minimized high scalability and better data aggregation, sensor nodes are often grouped into disjoint, non overlapping subsets called clusters.

Clusters create hierarchical WSNs which incorporate efficient utilization of limited resources of sen-sor nodes and thus extends network lifetime.

Stefanos A. Nikolidakis et al. The lifetime of the sensor node is based on battery powered devices. Several authors discussed the different energy consumption techniques for different layers. Consolidated efficient energy consump- tion techniques are missing. This paper provides a detailed study about all the existing energy conservation tech- niques and

also explains about limitations available in the techniques. In this paper, the energy conservation ap- proaches and its algorithms for computing the optimal transmitting ranges in order to generate a network with de- sired properties are discussed.

Xun Li et al. The potential for collaborative, robust networks of microsensors has attracted a great deal of research attention. For the most part, this is due to the compelling applications that will be enabled once wireless micro-sen- sor networks are in place; location-sensing, environmental sensing, medical monitoring and similar applications areall gaining interest. However, wireless micro-sensor networks pose numerous design challenges. For applications requiring long term, robust sensing, such as military reconnaissance, one important challenge is to design sensor networks that have long system lifetimes. This challenge is especially difficult due to the energy constrained natureof the devices. In order to design networks that have extremely long lifetimes, author proposed a physical layer dri- ven approach to designing protocols and algorithms.

Meera Gandhi. G et al. proposed Multi-hop Communication with Localization (MCL), a strategy to localize and route information to nodes present in such areas by determining angles and distances of consecutive nodes hop by hop towards the Base Station. Based on the application area, Subterranean Wireless Sensor Networks are specifical- ly designed to detect underground abnormal conditions and reported to the base station. Many protocols use distancebetween the nodes as one of the criteria for multi-hop communication in the network. It is found to be necessary to know the location of the nodes and the distance between the nodes in many power optimization protocols. But the query of how to attain the distance or the location arises in the same. The main objective here is to design a tech- nique to both localize and transmit data efficiently in subterranean areas. Initially there is a group of nodes deployed in the underground areas all of which bond to a sink that is further connected to the Base Station. It is possible to locate all the nodes through GPS which can be used as a reference in the worst case scenario by the Base Station.

The sink node has a Node Transmission Area (NTA) within which a node can be directly recognized by the sink node otherwise it finds the target node through the intermediate nodes.

Prabha R et al. proposed a QoS Aware Trust Metric based Framework for Wireless Sensor Networks. The proposed framework safeguards a wireless sensor network from intruders by considering the trustworthiness of the forwarder node at every stage of multi-hop routing. Increases network lifetime by considering the energy level of the node, prevents the adversary from tracing the route from source to destination by providing path variation.

3. RESEARCH PROBLEM IN WSN

Subscriber identity confidentiality: A serious weakness of the GPRS security architecture is related to the compro- mise of the confidentiality of subscriber identity. Specifically, whenever the serving network (VLR or SGSN) cannotassociate the TMSI with the IMSI, because of TMSI corruption or database failure, the SGSN should request the MS to identify itself by means of IMSI on the radio path.

Subscriber Authentication: The authentication mechanism used in GPRS also exhibits some weak points regarding security. More specifically, the authentication procedure is one-way, and, thus, it does not assure that a mobile user is connected to an authentic serving network. This fact enables active attacks using a false base station identity.

Data and signalling protection: An important weakness of the GPRS security architecture is related to the fact that the encryption of signalling and user data over the highly exposed radio interface is not mandatory. Some GPRS operators, in certain countries, are never switch on encryption in their networks, since the legal framework in these countries do not permit that. Hence, in these cases signalling and data traffic are conveyed in clear-text over the ra- dio path. This situation is becoming even more risky from the fact that the involved end-users (humans) are not in- formed whether their sessions are encrypted or not.

GPRS backbone: Based on the analysis of the GPRS security architecture (see sect. 3) it can be perceived that the GPRS security does not aim at the GPRS backbone and the wire-line connections, but merely at the radio access network and the wireless path. Thus, user data and signalling information, conveyed over the GPRS backbone, may experience security threats, which degrade the level of security supported by GPRS. In the following, the security weaknesses of the GPRS security architecture that are related to the GPRS backbone network for both signalling and data plane are presented and analyzed

4. ROUTING ESSENTIALS IN WSN

In this research a modified energy efficient GPRS based technique is to be developed with which the energy con- sumption can be reduced and efficiency of the routing protocol can be enhanced. This proposed scheme includes following steps that are to be performed:

First of all the network nodes are to placed in a deployment area in a random manner by setting initial parameters after that the listening and detection period will be started. In this phase following operations are performed. There isa detection time and a listening period in each round. Sensors in the former detect events, whereas sensors in the latter listen to their neighbours for packets to be sent to sinks. Different rounds have the same lengths. In each cycle, the detection times (listening periods) are the same duration. Following the detection phase, communication cooperation among nodes begins, which is characterised as follows. A sensor keeps track of the statuses of its neighbours in its relaying list to determine

when to announce a detected event (i.e., transmit an event packet P) or when to send a receiving/relay packet P, e.g., at time t, which is known as the delivery time point (DTP) of P, regardless of whether P is an announcement packet or a relayed packet.

Following the completion of the cooperation phase, an event packet provides event information such as event location, constrained time point (CTP), and kind of this event, allowing the system administrator to determine where the event is and what has occurred. CTP is an abstract time point at which an event packet should arrive to its sink. During the last phase, packets are relayed in the following way. During a listening period, when sj receives an event packet P, it will relay P to one of its neighbours on the route toward the corresponding sink, e.g., ki, but not immediately because sj's surrounding sensors belonging to the underlying WSN or other WSNs may also be relaying an event announcement packet, denoted by Q, Q!= P, in the same direction as P. It is preferable for sj to aggregate P and Q as one, which is subsequently transmitted, saving some energy when providing P and Q together.

Network Setting

Number of Node are n Number of Sensors are mEnergy for Sensors is E

Assume there are n sinks, K = ki|i = 1,..., n, for event collection, and m sensors, S = sj|j = 1,..., m, for event detection and data transmission in a monitoring environment. Each sensor is aware of its leftover energy as well as the positions of all sinks. There are also p sorts of events, E = eh|h = 1,..., p. Generally, m n, p.

Let kri,h and srj,h be the sink ki and sensorsj parameters on events of type eh, respectively. If sink ki can collect events of type e1, e2, and e3, we say ki is accountable for events of type e1, e2, and e3, and kri,1 = kri,2 = kri,3 = 1. In other words, event packets of type eh can be sent to sink ki,h = 1, 2, 3. Event packets of type eh would not be sent to ki if kri,h = 0. When a sensor sj can detect an event of type eh, we say that sj is responsible for the event, and hence srj,h = 1. When srj,h = 0, sj is unable to identify events of type eh.

Detection and Listening Periods

A round is made up of two parts: detection and listening. Sensors in the former detect events, whereas sensors in the latter listen to their neighbours for packets that will be sent to sinks. The lengths of the several rounds are the same. In each cycle, the detection times (listening periods) are the same duration.

The packet forwarding procedure of the Routing scheme is that when a sensor receives n packets, Pkt(t) ={pi |i = 1,2, . . . n}, in round t, if h packets, P(t) ={p'j |j = 1, 2, h}, $1 \leq h \leq n$, are transmitted to the same direction, the h packets are then aggregated into one and forwarded. Let CD^t_j , CT^t_j , CR^t_j , and CA^t_j be the energy consumed, re-spectively, for event detection, data transmission, packet receiving, and packet aggregation, by sj during time slot t.Let E^t_j be the residual energy of sensor sj in time slot t as

$$E^t_j = E^{t-1}_j - CD^t_j - CT^t_j - CR^t_j - CA^t_j ,$$

$$E^{full} \geq E^t_j \geq 0 \ \& \ s_j \square S$$

where t = 1, 2, and E 0 j = E full denotes a sensor's full energy at t = 1. If a sensor's energy level is zero, it can no longer detect its surroundings. Furthermore, all sensor networks seek to notify sinks of the occurrence of events. When more than ratio p events are uninformed, we consider the sensor network's lifespan to be over. We want to increase the lifespan of environmental monitoring, for example, T, provided that the number of uninformed events is smaller than ratio p.

Maximize T, which is defined as the number of uninformed occurrences that are fewer than the ratio p when u t T, where t and u are the number of rounds and the time period of each round, respectively.

Cooperation Among Networks

To accomplish this, a sensor keeps the statuses of its neighbours in its relaying list to determine when to announce a detected event (i.e., transmitting an event packet P) or when to send a receiving/relay packet P, e.g., at time t, which is known as the delivery time point (DTP) of P, regardless of whether P is an announcement packet or a relayed packet. Before sending P, the waiting time (WT) is defined as

$$WT = DTP - \text{current system time}$$

Events Detected In Detection Period

An event packet contains event information such as event location, constrained time point (CTP), and event kind, allowing system administrators to determine where the event is and what has occurred. CTP is an abstract time point at which an event packet should arrive to its sink. Assume that an event eh is detected by sj and that the related event packet Ph,i is transmitted to sink ki. Let CTPh,i be the event's CTP. As a result, Ph, I should get at ki before CTPh,i.

$$CTP_{h,i} = DCh + DTP_{h,j} + \varepsilon$$

where DTPh,j is the absolute time point at which eh is detected by sj, is the monitoring system's error parameter indicating the synchronous difference between sj and ki, and DCh is the delay constraint imposed to event type eh - Forehand is defined as the longest endurable packet delivery time length between the time sj detects eh and the time Ph,i arrives at ki.

Relaying Packet in Listening Period

During a listening period, when sj receives an event packet P, it will relay P to one of its neighbours on the route to the corresponding sink, e.g., ki, but not immediately because sj's surrounding sensors belonging to the underlying WSN or other WSNs may also be relaying an event announcement packet, denoted by Q, Q!= P, in the same direction as P. It is preferable for sj to aggregate P and Q as one, which is subsequently transmitted, saving some energy when providing P and Q together. As a result, P's WT before delivery, represented by WRj, by sj, is computed as

$$WRj = \{\min(Di,j)/D / (Ej + \gamma\, ERj) \times Cj \times R\} \text{ x } PR/W$$

where D is the working environment's diameter, Di,j is the distance between sj and ki, Ej (ERj) is the residual energy of sj (sj's all neighbours in sj's hot region), as the level of neighbours' importance is determined by the underlying monitoring system, Cj is the current number of packets that will be aggregated by sj, Rj represents the number of time slots/rounds that the earliest packet has waited to be aggregated by

4. CONCLUSION

Appearing of WMSN has opened new large applications in our life and many researches issues emerged that need different solutions. In this paper WMSN technologies are introduced. Challenges and resource constraints are dis- cussed. Current routing protocols are classified according to the existing researches direction. Also the routing pro- tocols proposed for multimedia transmission are surveyed and the performance issues of each routing protocol are highlighted. In our survey, we can realize that the proposed protocols for wireless multimedia sensor network have different methodologies and one goal which satisfies the multimedia transmission requirements.The proposed proto- cols lie under different categories as mentioned, where the first class shows the routing protocols based on ant colony optimization.The ACO displays several features that make it particularly suitable for wireless multimedia sensor networks. Additionally, ant routing has shown excellent performance to solve routing problems in WSNs and ad hoc networks. The second class is geographic routing protocols like TPGF and GPSR. These protocols achieve good performance in hole bypassing and it is suitable for WMSN as it ensures uniform energy consumption and meets the delay and packet loss constraint. The last class of the proposed protocols follows different algorithm types and addresses different QoS metrics that are required for multimedia transmission with resource constraint nature of WMSN. We believe that the researches focus will increase in this area, while developing routing protocols will at- tract more attention since they play the key roles behind the development of WSN.

5. REFERENCES

1. Elhabyan, R.; Shi, W.; St-Hilaire, M.; "Coverage Protocols for Wireless Sensor Networks: Review and Fu- ture Directions", Journal of Communications and Networks, Vol: 21, No: 1, 2019, pp: 45-60
2. Hung, L.L.; Leu, F.Y.; Tsai, K.L.; Yo, C.K.; "Energy-Efficient Cooperative Routing Scheme for Heteroge- neous Wireless Sensor Networks", IEEE Access, Vol: 8, 2020, pp: 56321-56332
3. Li, G.; Peng, S.; Wang, C., Niu, J.; Yuan, Y.; "An Energy-Efficient Data Collection Scheme Using Denois- ing Autoencoder in Wireless Sensor Networks", Tsinghua Science and Technology, Vol: 24, No: 1, 2019, pp: 86-96
4. Akerele, M.; Al-Anbag, I.; Erol-Kantarc, M.; "A Fiber-Wireless Sensor Networks QoS Mechanism for Smart Grid Applications", IEEE Access, Vol: 7, 2019, pp: 37601- 37610
5. Lyu, B.; Qi, T.; Guo, H.; Yang, Z.; "Throughput Maximization in Full-Duplex Dual-Hop Wireless Powered Communication Networks", IEEE Access, Vol: 7, 2019, pp: 158584- 158593
6. Li, Z.; Zhong, A.; "Resource Allocation in Wireless Powered Virtualized Sensor Networks", IEEE Access, Vol: 8, 2020, pp: 40327-40336
7. Khalid, N.; Mirzavand, R.; Saghlatoon, H.; Honari, M.M.; Mousavi, P.; "A Three-Port Zero-Power RFID Sensor Architecture for IoT Applications", IEEE Access, Vol: 8, 2020, pp: 66888- 66897
8. Qiu, T.; Liu, J.; Si, W.; Wu, D.O.; "Robustness Optimization Scheme With Multi-Population Co-Evolution for Scale-Free Wireless Sensor Networks", IEEE, Vol: 27, No: 3, 2019, pp: 1028-1032
9. Tan, X.; Zhao, H.; Han, G.; Zhang, W.; Zhu, T.; "QSDN-WISE: A New QoS-Based Routing Protocol for Software-Defined Wireless Sensor Networks", IEEE Access, Vol: 7, 2019, pp: 61070-61082
10. Gao, D.; Zhang, S.; Zhang, F.; He, T.; Zhang, J.; "RowBee: A Routing Protocol Based on Cross-Technology Communication for Energy-Harvesting Wireless Sensor Networks", IEEE Access, Vol: 7, 2019, pp: 40663-40673
11. Mohit Saini, Rakesh Kumar Saini, "Solution of Energy-Efficiency of sensor nodes in Wireless sensor Net- works", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277-128X, Vol: 3, Issue 5, May 2013, pp: 353-357
12. Satvir Singh, Meenaxi, "A Survey on Energy Efficient Routing in Wireless Sensor Networks", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 7, pp. 184-189, July 2013.
13. K. Arun prabha, K. Hemapriya, "Energy Saving In Wireless Sensor Network Using Optimal Selective For- warding Protocol", International Journal of Advancements in Research & Technology, Volume 2, Issue1, January-2013.
14. Eugene Shih, SeonghwanCho, Nathan Ickes, Rex Min, Amit Sinha, Alice Wang, AnanthaChandrakasan, "Physical Layer Driven Protocol and Algorithm Design for energy efficient Wireless Sensor Networks"
15. Shashidhar Rao Gandham, Milind Dawande, Ravi Prakash, S. Venkatesan, "Energy Efficient Schemes fo

16. Wireless Sensor Networks with Multiple Mobile Base Stations."
17. Wei Ye, John Heidemann, Deborah Estrin, "An Energy-Efficient MAC Protocol for Wireless Sensor Net- works"
18. Stefanos A. Nikolidakis, DionisisKandris, Dimitrios D. Vergados, Christos Douligeris, "Energy Efficient Routin
19. in Wireless Sensor Networks Through Balanced Clustering"
20. Xun Li, Geoff V Merrett, Neil M White, "Energy-efficient data acquisition for accurate signal estimation in wireless sensor networks", Journal on wireless Communications and Networking 2013.
21. Meera Gandhi. G, P. Rama, "GPS based Multi-hop Communication with Localization in Subterranean Wireless Sensor Networks", Vol: 57, 2015, pp: 1189-1198
22. Prabha R, Krishnaveni M, S H Manjula, K R Venugopal, L M Patnaik, "QoS Aware Trust Metric based Framework for Wireless Sensor Networks", International Conference on Intelligent Computing, Communi- cation & Convergence, Vol: 48, 2015, pp: 373-380

DATA MINING TECHNIQUES FOR ELECTRICITY DEMAND FORECASTING

Mandeep Singh[1] and Dr. Raman Maini[2]
[1]*Research Scholar, Department of Computer Engineering*
Punjabi University, Patiala
[2]*Professor and Head of Department, Department of Computer Engineering*
Punjabi University, Patiala.

ABSTRACT:——This paper shows data mining techniques to forecast the electricity load demand of a geographic area using the artificial neural networks methods. History load and temperature data are to be used to perform the learning in network. After the network is trained using back propagation algorithm of learning it can be used to forecast the electricity demand at particular point of instance in future.

KEYWORDS: Data Mining, Load forecasting, Artificial Neural Networks, Back Propagation.

I. INTRODUCTION

A key component of the daily operation and planning activities of an electric utility is short-term load forecasting, i.e., the prediction of hourly loads (demand) for the next hour to several days out. The accuracy of such forecasts has significant economic impact for the utility. This study presents an approach of data mining technique to predict electricity demand of a geographical region based on the meteorological conditions. The value prediction predictive data mining technique can be implementing with the Artificial Neural. The electricity demand for any future period can be predict from the given weather parameters of a geographical region and Previous electric power consumption data. The demand prediction is a sophisticated task in energy deficient countries like India. If the accurate prediction mechanisms exist then demand of electric power required can be efficiently produced and delivered. The purposed model provides the forecast, which has scope of usage in Electricity load scheduling, Electricity load distribution and Meeting the customer requirements in different seasons. The aim of this literature is to present different techniques used for forecasting demand of electricity. There were different techniques available and most common method or we can say technique used for demand prediction was regression analysis technique and at present there are more techniques are available in this era and the most effective from among is artificial neural network (ANN) technique[1]. In this paper we discuss different types of techniques which fall in Data Mining. There are multiple factors which effects to the demand of electricity prediction. Due to up and down in consumption of electricity the prediction of demand is hard process and non-linearity is the factor which introduce the importance of non linear models and techniques for forecasting the demand of electricity.

Artificial Neural Network (ANN) is proposed to be used for electricity demand prediction. Electricity demand is to be dependent of different parameters like historic electricity consumption and meteorological data like temperature, rainfall, day type and day of week.

II. DATA MINING

This is studied under the Data Mining. Data mining is a process of extract and discovers pattern from data sets. Different Data Mining techniques are given in the following table.

Table 1: Data Mining Techniques

Operations	Data Mining Techniques
Deviation detection	Statistics, Visualization
Link analysis	Association's discovery, Similar time sequence discovery, Sequential pattern discovery
Database segmentation	Neural clustering, Demographic clustering
Predictive modeling	Value prediction, Classification

The technique used to perform electric demand forecasting is Predictive Modeling. Under this predictive modeling value prediction is performed using neural networks.

1) Deviation Detection- This deviation detection process is completed by using statistically and visualization based techniques of data mining. This operation can be performed using statistics and visualization techniques or as a by-product of data mining.

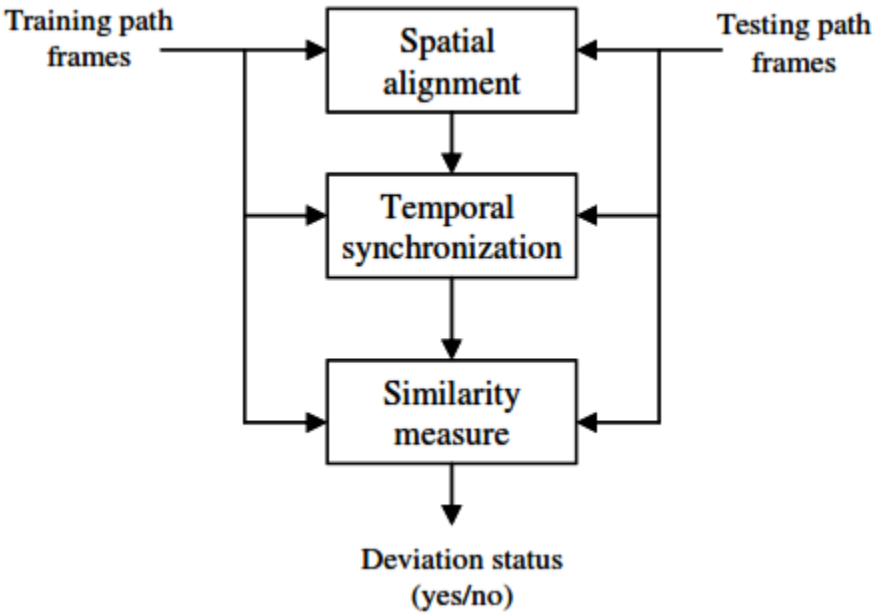


Figure 1: Block diagram for deviation detection during visual path following

Fig. 1 shows the block diagram of deviation detection. The unconstrained nature of the camera motion gives result in overlapping between training and testing the trajectory. Due to that reason the spatial alignment performed in this alignment frames are aligned and temporally synchronization and similarity measures calculate for testing and training purpose.

2) Link Analysis- Link Analysis aim is to establish links between different records in database, or links between sets of records in a database. There are mainly three types of link analysis as follow:

- Similar time sequence discovery
- Sequential Pattern Discovery
- Associations discovery

2.1) Similar time sequence discovery: - The similar time sequence discovery process is based on the associated links between two or more sets in database in which data of different sets are dependent on the time and similar data sets patterns which are associated based on time series.

2.2) Association's discovery: - The Association's discovery is used to find the items from data sets of events with the help of some association discovery rules. These rules can be made according to the situation of the events.

2.3) Sequential pattern discovery: - The Sequential pattern discovery is used to discover the patterns between different events like sequential pattern of tasks or process. Example of **s**equential pattern discovery is used in the applications to understand long term activities of clients.

3) Database Segmentation: - The database segmentation is used to make the segments of similar type and nature of data sets. The segmentations can be performed based on the unknown types and behaviors of the data sets.

These sets can be categorized based on the similar records of clusters which are sharing the properties with other data sets.

4) Predictive Modeling: - Predictive modeling is a normally used technique which is used in arithmetical, geometrical and algebraically data set to predict future behavior. This type of modeling is similar to human knowledge based. This technique used commonly in real world and have the ability to fit new-fangled data into predication of new values. Two types of predictive modeling are described below-

- Classification
- Tree Induction
- Neural Induction

4.1. Classification- Classification is used to establish predetermined class for each record in a database from a finite set of possible.

4.2 Tree Induction-Tree Induction is similar to the decision tree with tuples having training values.

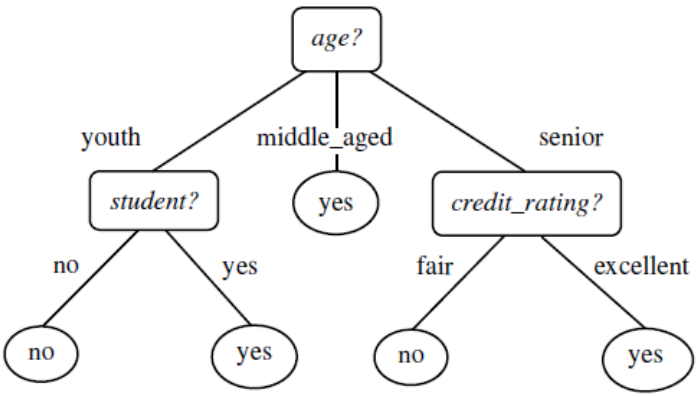


Figure 2: Tree Induction

Figure shows a tree structure having decision tuples. This decision tree shows age of person and has three types of ages one is youth, middle age and senior person. [12] Age is the root node and if the person has age value less than some specific value the person should be considered as youth and if the age lies between some specified values than the person should be considered as middle aged person and the third last is rest of person should be considered as senior persons. Now at the second level if the person is youth than the person will be check that the youth is student or not a student and if the person is belongs to middle aged this will be check. [18] And in third condition if the person is senior person than the bank credit limit is applicable on the person. The credit limit will be based on the usage of the senior person credit card. If the person uses the card repeatedly on daily bases and person transactions are more consumable than the person falls in fair credit limit or excellent credit limit.

4.3 Neural Induction- This Neural Induction method is based on the classification technique and for this work the artificial neural network (ANN) is used for classification and prediction of values. This neural network is made from different nodes; these nodes can be categories in three layers Input Layer, Processing Layer and Output layer. Input layer includes the input nodes which are used for input to the neural network, it can be one node or more than one node and processing layer includes the processing nodes which are used to process the data between input layer and the output layer, and the third one is output layer which include the output nodes these nodes are used for the purpose of output from the neural network [1].

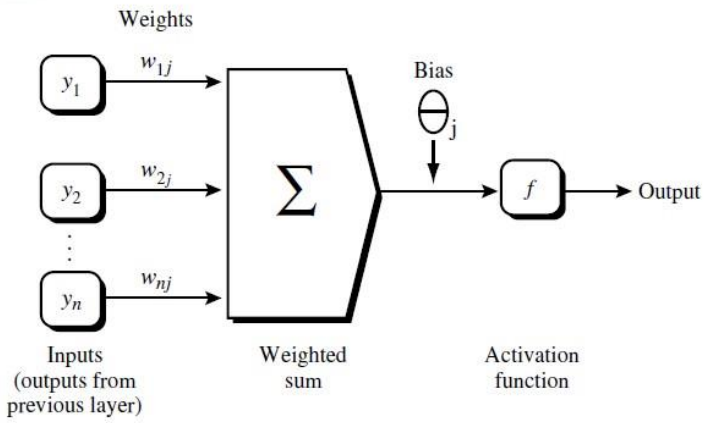


Figure 1.2 Neural Networks

III. CONCLUSIONS

The above discussed techniques are used to predict the values from the previously known data values and patterns and stets or we can say supervised learning techniques. These supervised learning techniques are used when we have the training data sets and we can use these data sets data values and patterns to train our model with different types of techniques. We can use these techniques from previously and we can make new technique according to the data sets which are appropriate to the results we needed from the model. And different parameters like metrological factors are considerable in electricity demand prediction. In the past regression technique was used for predict the future values of demand. Now day's artificial neural network (ANN) gives the best results for electricity demand prediction [10].

REFERENCES

[1] S. Badran and O. Abouelatta, "Forecasting Electrical Load using ANN Combined with Multiple Regression Method", *The Research Bulletin of Jordan ACM*, 2011, vol. 2, Issue no. 2, pp. 152-158.

[2] Slobadan Ilie, Aleksander Selakov, Srdan Vukmirovie, Aleksandar Erdeljan and Filip Kulie, "Short term load forecsting in large scale electrical utility using artificial neural network", *Journal of Scientific & Industrial research,* 2013, Vol. 72, pp. 739-745.

[3] Abdel-Aal , "Univariate modeling and forecasting of monthly energy demand time series using abductive and neural networks", *Computers & Industrial Engineering*, 2007, pp. 903-917, Elsevier.

[4] A. Jain and B. Satish, "Clustering based short term load forecasting using support vector machines", *in Proceedings of the IEEE Bucharest PowerTech,* 2009, pp. 1-8, IEEE.

[5] Heiko Hahn, Silja Meyer-Nieberg and Stefan Pickl, "Electric load forecasting methods: Tools for decision making", *European Journal of Operational Research* , 2009, pp. 902-907, Elsevier.

[6] Tao Hong, David A. Dickey, "Electric load forecasting: fundamentals and best Practices", 2009.

[7] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques, 2nd ed.", *Morgan Kaufmann publications,* 2006.

[8] J. Deng, "Modeling and Prediction of China's Electricity Consumption Using Artificial Neural Network", *Proceedings of the 6th International Conference on Natural Computation*, 2010, IEEE.

[9] W. Mai, C.Y. Chung, T. Wu, and H. Huang, "Electric load forecasting for large office building based on radial basis function neural network," *in Proc. PES General Meeting. Conf. Expo.,* 2014, pp. 1–5, IEEE.

[10] S. Saravanan, S.Kannan, C. Thangaraj, "India's Electricity Demand Forecast Using Regression Analysis and Artificial Neural Networks based on Principal Components", *ICTACT Journal on Soft Computing*, 2012, pp. 365-370.

[11] H. K. Mohamed , S. M. El-Debeiky , H. M. Mahmoud and K. M. El Destawy, "Data mining for electrical load forecasting in egyptian electrical network", Proc. *Int. Conf. Comput. Eng. Syst.,* 2006, pp. 460-465, IEEE.

[12] C. Fan, F. Xiao and S. Wang, "Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques", *Appl. Energy*, 2014, vol. 127, pp. 1-10, Elsevier.

[13] Prakash GL, K. Sambasivarao, Priyanka Kirsali and Vibhuti Singh, "Short Term Load Forecasting for Uttarakhand using Neural Network and Time Series models", 2014, *IEEE.*

[14] Luiz Friedrich and Afshin Afshari, "Short-term forecasting of the Abu Dhabi electricity load using multiple weather variables", *Energy Procedia,* 2015, ELSEVIER.

[15] Fausett L., "Fundamentals of Neural Networks Architectures, Algorithms and Applications", *Dorling Kindersley*, 2008.

[16] Jeyaseeli S.S., Kathirvalavakumar T., "Adaptive modified backpropagation algorithm based on differential errors", *International Journal of Computer Science, Engineering and Applications,* 2011, Vol 1, No. 5, pp. 21-33.

[17] Arunesh Kumar Singh, Ibraheem, S. Khatoon, Md. Muazzam and D. K. Chaturvedi, "Load Forecasting Techniques and Methodologies: A Review", *2nd International Conference on Power, Control and Embedded Systems,* pp. 631-640, 2012, IEEE

[18] Jiangxia Zhong, Xinghuo Yu, Miguel Combariza, and Jinjian Wang, "An Intelligent Relational Pattern Matching System for Electricity Demand Prediction", 2014, pp. 3510-3516, IEEE.

[19] Yu, Z.J., Haghighat, F. & Fung ,B.C., "Advancesand challenges in building engineering and data mining applications for energy-efficient communities", *Sustainable Cities and Society*, 2016. ELSEVIER.

REVIEW ON FACE MASK WEARING DETECTION TECHNIQUES

Urvashi[1], Lakhwinder Kaur[2], Sumandeep Kaur[3], MadanLal[4]
Computer Science & Engineering, Punjabi University
[1]Urvashis064@gmail.com
[2]Mahal2k8@gmail.com
[3]sumandhanjal@gmail.com
[4]mlpbiuni@gmail.com

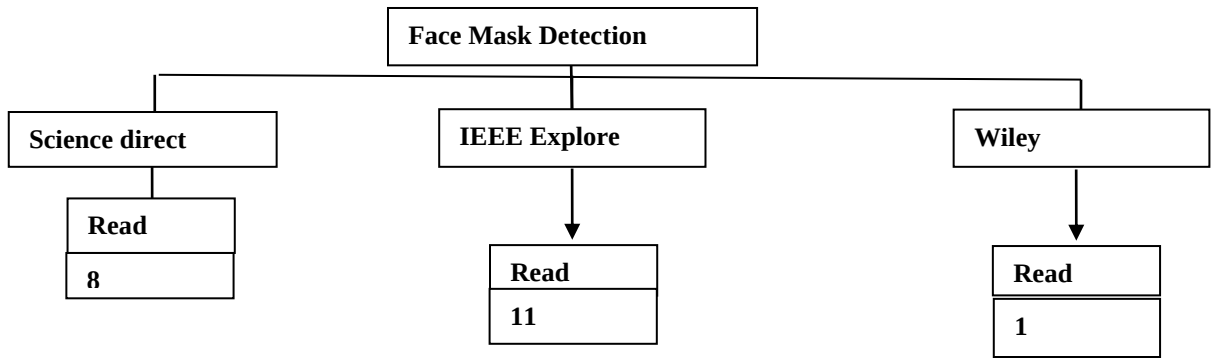
ABSTRACT: Since the outburst of the corona virus, many countries have recommended to their native to adopt physical distancing, face shield wearing and hand hygiene. Wearing face shield has not been adopted by most of citizens. While the reasons were composite. Face shield wearing can block or filter flying virus carrying particles. In this paper a detail study of all accurate recently developed techniques and devices developed for detection of a person wearing a face mask or not is carried out. The main focus of the study is on the effectiveness of face mask wearing detection techniques applied to stop spread of corona virus.

I. INTRODUCTION

The spread of COVID-19 is increasingly worrying for everyone in the world. The virus can be affected from human to human through the droplet and airborne. More than 14 million people have been infected by this virus. According to the instruction from the WHO, to reduce the spread of COVID-19 every people need to wear face mask to prevent the spread of the COVID-19 and the mask can reduce the spray of droplets when worn over the nose and mouth. Wearing a face mask will help prevent the spread of infection and prevent the individual from contracting any airborne infectious germs. Most of people won't wear the face mask properly with so many reasons. To overcome this situation, a different face mask detection technique needs to be developed. The main objective of this research work is to review numbers of recently reviewed research papers to find out which face mask detection technique gives high accuracy.

II. LITERATURE REVIEW

This section summarized of different researchers contribution in terms of technique used, database(s) used for evolution, performance achieved in terms of different parameters and conclusion given by them as shown in table 1.



Various research papers related to face mask detection wearing techniques are listed in the following table.

TABLE-1

Research Contribution	Year	Techniques	Database(s)	Parameters	Conclusion (s)
A Novel Detection Framework About Conditions of Wearing Face Mask for Helping Control the Spread of COVID-19[1]	2021	Context-Attention R-CNN	MAFA	Mean Average Precision (mAP) = 84.1%	Context-Attention R-CNN outperforms many state-of-the-art detectors, including two-stage detectors and single-stage detectors
Face mask detection using YOLOv3 and faster R-CNN models: COVID-19 environment[2]	2021	YOLOv3 and faster R-CNN	MAFA, WIDER FACE	Average Precision of : YOLOv3 is 55% & Faster Rcnn is 62%	YOLO algorithm as it performs single-shot detection and has a much higher frame rate than Faster-RCNN

Face Mask Detection Using MobileNetV2 in The Era of COVID-19 Pandemic[3]	2020	MobileNet	Kaggle dataset and the Real-World Masked Face dataset (RMFD)	Accuracy-96.85%	Study can be an easy move for authorities to use more unstructured data resources for more data-based mitigation, evaluation, prevention, and action planning against COVID-19
Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread[4]	2021	ResNet 50 AlexNet 50 MobileNet 50	FDDB MALF calebA WIDERFACE WIDERFACE SMFRD MFDD	Accuracy -98.2%	The proposed technique efficiently handles occlusions in dense situations by making use of an ensemble of single and two-stage detectors at the pre-processing leve
Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection[5]	2021	YOLO v2 ResNet 50	Medical Masks Dataset Face Mask Dataset	Average precision - 81% using YOLOv2 with ResNet-50	The proposed model improves detection performance by introducing mean IoU to estimate the best number of anchor boxes
Novel Face Mask Detection Technique using Machine Learning to control COVID'19 pandemic[6]	2021	C2D CNN	RWMF CelebA		ML based topology provides better results with higher accuracy and is more effective in controlling the COVID-19 pandemic.
Real time data analysis of face mask detection and social distance measurement using Matlab[7]	2020	Faster R-CNN	RWMF	Accuracy-93.4%	It automatically shown who is Unmasked person and Who is not keep a social distance
A Hybrid Deep Transfer Learning Model with Machine Learning Methods for Face Mask Detection in the Era of the COVID-19 Pandemic[8]	2020	ResNet-50c	RMFD SMFD LFW	Accuracy of RMFD-99.64%, SMFD-99.49% & LFW-100%	The SVM classifier achieved the highest accuracy possible with the least time consumed in the training process
Application of deep learning and machine learning models to detect COVID-19 face masks - A review[9]	2021	InceptionV3	Created Artificially	Acuuracy-99.9%	Deeper and wider deep learning architectures with increased training parameters, such as inception-v4, Mask R-CNN, Faster R-CNN, YOLOv3, Xception, and DenseNet are not yet imple- mented to detect face masks
Face mask detection using deep learning: An approach to reduce risk of Corona virus spread[10]	2021	ResNet-50 AlexNet MobileNet		Accuracy With ResNet-50=98.20%	Efficiently handles occlusions in dense situations by making use of an ensemble of single and two-stage detectors at the pre-processing level
A Deep Learning Based Assistive System to Classify COVID-19 Face Mask for Human Safety with YOLOv3[11]	2019	YOLOV3	UFPR-ALPR	Accuracy-96%	It's perform really well in images and our detection results was also quite good

Face Detection with the Faster R-CNN[12]	2017	RCNN	FDDB WIDER Face		Effectiveness comes from the region proposal network (RPN) module
SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2 [13]	2021	SSDMNV2	WIDER Face IJB-A MALF CELEBA	Accuracy-93%	OpenCV deep neural networks used in this model generated fruitful results
Scaling up face masks detection with YOLO on a novel dataset [14]	2021	YOLO V4	ImageNet MS COCO PASCAL VOC	Precision Yolo v4-78%	The results indicate YOLO v3 and v4, tiny YOLO v3 and v4 and modified tiny YOLO v3 and v4 achieved a higher mAP on MOXA dataset promising these to be suitable for real-time face masks detection
IoT-Enabled Smart Doors for Monitoring Body Temperature and Face Mask Detection[15]	2021	IOT	Artificial Dataset	Accuracy-97%	The test results demonstrate a high level of accuracy in detecting people wearing and not wearing facemasks, as well as it also generates alarms monitored and recorded
COVID-19 Monitoring System using Social Distancing and Face Mask Detection on Surveillance video datasets[16]	2021	YOLO V3 MobileNet V2 DSF	CelebA	Accuracy-91.2%	Efficient solution to monitor social distancing practices in public areas where it is very difficult to monitor manually
Deep Neural Architecture for Face mask Detection on Simulated Masked Face Dataset against Covid-19 Pandemic[17]	2021	CNN VGG16	SMFD	Accuracy of CNN-96.35% & VGG16-97.42%	The proposed work can be used where monitoring is needed
Evaluating the Masked and Unmasked Face with LeNet Algorithm[18]	2021	MTCNN	Face Mask Dataset	Accuracy-98.21%	The algorithm needs to extract each feature for all the unmasked face samples
Mask Wearing Detection Method Based on SSD Mask Algorithm[19]	2020	SSD SeNet VGG16	WIDER Faces MAFA	Accuracy-88.6%	SSD-Algorithm has good accuracy and robustness
Real Time Face Mask Detector Using YOLO V3 Algorithm & Haar Cascade Classifier[20]	2020	YOLO V3 Haar Cascade Classifier	MAFA	Accuracy-90.1%	The propose algorithm images enhancement technique to accuracy of the system

Abbreviations used in table are as follows:-
CNN- Convolution Neural Network
RCNN-Region-based convolution neural network
mAP- Mean Average Precision
YOLO-You Only Look Once
SSD- Single shot Multibox Detector

III. DISCUSSION

Zhang et al. [1] have used Context Attention RCNN technique to detect the fine-grained wearing different state of face mask (without face mask, face with incorrect mask, and face with accurate mask). In future, the authors explore the imbalance problems and a better attention architecture for more accurate detection on conditions of wearing face mask.

Singh et al. [2] use YOLO V3 & RCNN technique to detect the person wear a face mask or not by trained the dataset. In future author extend this application in any public places for future work like stations etc.

Sanjaya et al. [3] also detect face mask with another technique called MobileNet V2. This is image classification technique used in machine learning.

Sethi et al. [4] discover an application to detect face mask in 2021 using deep learning algorithm ResNet 50, AlexNet, MoblieNet.In future the author use this application in video surveillance camera.

Loey et al. [5] discover an application to detect the medical face mask by using technique YOLO V2 with ResNet-50 this is deep learning algorithms. In future author use this application on videos to detect face mask.

Gupta et al. [6] use machine learning technique to control the spread of corona virus by using C2D-CNN technique on dataset. In future reference author also use this application in surveillance cameras to detect people wearing mask or not.

Meivel et al. [7] use Faster R-CNN algorithm to detect face shield and individual space on the dataset (RWMF) this contain more complex images.

Loey et al. [8] discover an application to detect face protection mask to control the spreading of corona virus by machine learning feature extraction technique ResNet-50. They also use this application for feature extraction problems.

Mbunge et al. [9] used the machine learning technique InceptionV3 CNN method for facial mask wearing detection to put break on spread of corona virus. In future author use this application on real world images.

Sethi et al. [10] discover the method to detecting face protective shield by using deep learning method ResNet-50, AlexNet, MobileNet. Author use this technique for future references in biometric to detect face landmarks.

Bhuiyan et al. [11] used a deep learning technique (YOLO V3) to detect human face mask. In future references author trained more data to obtained more accuracy.

Jiang et al. [12] discoved a technique to detect the facial mask by using a Faster R-CNN method. In future author trained the dataset for achieving more accuracy.

Nagrath et al. [13] used Single Shot Multibox detector & MobileNet technique to detect face shield .This is DNA based technique used by author.

Kumar et al. [14] discoved a technique to detect facial shield to stop the spread of corona virus. The technique used by author is YOLO V4.

Varshini Bet al. [15] used Internet of things (IoT) implant smart door to detect facial mask. In future author develop the mobile application to monitor the face mask wearing detection.

Srinivasan et al. [16] detect the face shield by using the technique called YOLO V3, Dual Shot Face Detector and MobileNet.

Negi et al. [17] developed an architecture of deep neural is used to detect the face shield on the trained dataset (SMFD). Author use CNN and VGG16 technique to detect face shield.

Rusli et al. [18] used an algorithm called LeNet which comes under MultiTask Cascaded Neural Network. For future references author use this technique for face recognition systems.

Xu et al. [19] used Single shot Multibox Detector to detect human face mask on the trained dataset (WIDER Faces,MAFA).Author used SeNet & VGG 16 method to detect face shield.

VINH et al. [20] used YOLO V3 & Haar Cascade Classifier to detect the facial shield to stop the spread of corona virus. For future references author work on improved the accuracy by adding more images on dataset.

IV. CONCLUSION

This paper is review of work done by different researchers till now on face mask wearing detection techniques. This study includes the different technique which is used to detect the face mask and their major features. Recently developed techniques are compared based on different parameters. Most of the technique discussed in this work uses different dataset

which contain thousand of images. A single technique is there which uses face images from real time video cameras. Among the reviewed papers a technique InceptionV3 gives highest accuracy of 99.9%.

V. FUTURE WORK

After reading these paper and examine the different technique used for face mask wearing detection, the following sub point are bring out, which give the idea to researchers to work in this field:-

1. There are many technique used to detect the face mask wearing but there is one technique called RCNN which work best for face mask wearing detection.
2. In this paper many techniques and algorithms are discussed but all these are varies upon their parameters and dataset. InceptionV3 give the highest accuracy of 99.9%.
3. In this paper main focus is on face mask wearing detection but work can also be done on type of face mask.

VI. REFERENCES

1. Zhang, J., Han, F., Chun, Y., & Chen, W. (2021). A novel detection framework about conditions of wearing face mask for helping control the spread of covid-19. *IEEE Access*, *9*, 42975–42984. https://doi.org/10.1109/access.2021.3066538
2. Singh, S., Ahuja, U., Kumar, M., Kumar, K., & Sachdeva, M. (2021). Face mask detection using yolov3 and faster R-CNN models: COVID-19 environment. *Multimedia Tools and Applications*, *80*(13), 19753–19768. https://doi.org/10.1007/s11042-021-10711-8
3. Sanjaya, S. A., & Adi Rakhmawan, S. (2020). Face mask detection using mobilenetv2 in the era of COVID-19 pandemic. *2020 International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*. https://doi.org/10.1109/icdabi51230.2020.9325631
4. Sethi, S., Kathuria, M., & Kaushik, T. (2021). Face mask detection using Deep learning: An approach to reduce risk of coronavirus spread. *Journal of Biomedical Informatics*, *120*, 103848. https://doi.org/10.1016/j.jbi.2021.103848
5. Loey, M., Manogaran, G., Taha, M. H., & Khalifa, N. E. (2021). Fighting against covid-19: A novel deep learning model based on Yolo-V2 with resnet-50 for medical face mask detection. *Sustainable Cities and Society*, *65*, 102600. https://doi.org/10.1016/j.scs.2020.102600
6. Gupta, S., Sreenivasu, S. V. N., Chouhan, K., Shrivastava, A., Sahu, B., & Manohar Potdar, R. (2021). Novel face mask detection technique using machine learning to control COVID'19 pandemic. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.07.368
7. Meivel, S., Indira Devi, K., Uma Maheswari, S., & Vijaya Menaka, J. (2021). Real time data analysis of face mask detection and social distance measurement using MATLAB. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2020.12.1042
8. Loey, M., Manogaran, G., Taha, M. H., & Khalifa, N. E. (2021). A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the COVID-19 pandemic. *Measurement*, *167*, 108288. https://doi.org/10.1016/j.measurement.2020.108288
9. Mbunge, E., Simelane, S., Fashoto, S. G., Akinnuwesi, B., & Metfula, A. S. (2021). Application of deep learning and machine learning models to detect COVID-19 face masks - A Review. *Sustainable Operations and Computers*, *2*, 235–245. https://doi.org/10.1016/j.susoc.2021.08.001
10. Sethi, S., Kathuria, M., & Kaushik, T. (2021). Face mask detection using Deep learning: An approach to reduce risk of coronavirus spread. *Journal of Biomedical Informatics*, *120*, 103848. https://doi.org/10.1016/j.jbi.2021.103848
11. Bhuiyan, M. R., Khushbu, S. A., & Islam, M. S. (2020). A deep learning based assistive system to classify covid-19 face mask for human safety with yolov3. *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. https://doi.org/10.1109/icccnt49239.2020.9225384
12. Jiang, H., & Learned-Miller, E. (2017). Face detection with the faster R-CNN. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. https://doi.org/10.1109/fg.2017.82
13. Nagrath, P., Jain, R., Madan, A., Arora, R., Kataria, P., & Hemanth, J. (2021). SSDMNV2: A real time DNN-based face mask detection system using single shot multibox detector and mobilenetv2. *Sustainable Cities and Society*, *66*, 102692. https://doi.org/10.1016/j.scs.2020.102692
14. Kumar, A., Kalia, A., Verma, K., Sharma, A., & Kaushal, M. (2021). Scaling up face masks detection with Yolo on a novel dataset. *Optik*, *239*, 166744. https://doi.org/10.1016/j.ijleo.2021.166744
15. Varshini, B., Yogesh, H. R., Pasha, S. D., Suhail, M., Madhumitha, V., & Sasi, A. (2021). IOT-enabled smart doors for monitoring body temperature and face mask detection. *Global Transitions Proceedings*, *2*(2), 246–254. https://doi.org/10.1016/j.gltp.2021.08.071
16. Srinivasan, S., Rujula Singh, R., Biradar, R. R., & Revathi, S. A. (2021). Covid-19 monitoring system using social distancing and face mask detection on surveillance video datasets. *2021 International Conference on Emerging Smart Computing and Informatics (ESCI)*. https://doi.org/10.1109/esci50559.2021.9396783
17. Negi, A., Kumar, K., Chauhan, P., & Rajput, R. S. (2021). Deep Neural Architecture for face mask detection on simulated masked face dataset against covid-19 pandemic. *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*. https://doi.org/10.1109/icccis51004.2021.9397196

18. Rusli, M. H., Sjarif, N. N., Yuhaniz, S. S., Kok, S., & Kadir, M. S. (2021). Evaluating the masked and unmasked face with lenet algorithm. *2021 IEEE 17th International Colloquium on Signal Processing & Its Applications (CSPA)*. https://doi.org/10.1109/cspa52141.2021.9377283
19. Xu, M., Wang, H., Yang, S., & Li, R. (2020). Mask wearing detection method based on SSD-mask algorithm. *2020 International Conference on Computer Science and Management Technology (ICCSMT)*. https://doi.org/10.1109/iccsmt51754.2020.00034
20. Vinh, T. Q., & Anh, N. T. (2020). Real-time face mask detector using yolov3 algorithm and Haar Cascade classifier. *2020 International Conference on Advanced Computing and Applications (ACOMP)*. https://doi.org/10.1109/acomp50827.2020.00029

CYBERCRIME IN INDIA AMID COVID-19: ANALYSIS OF CYBER-ATTACKS AND CORRELATIONS BETWEEN EVENTS & CYBER-CRIMINAL CAMPAIGNS

Jashanpreet Singh Toor [1], Dr Abhinav Bhandari[2]
[1,2,] *Department of Computer Science and Engineering, Punjabi University*
[1]jashanpreet@pbi.ac.in
[2]bhandarinitj@gmail.com

ABSTRACT— Although cybersecurity is the contrary of cybercrime, they cannot exist without each other. The Coronavirus pandemic was one of its kind event that has affected lives of billions of citizens in the world resulting in an extraordinary change in the lifestyle and business. As the whole world went in to lockdown it shifted the world into online mode and increased the likelihood of cyber-attacks succeeding, corresponding with an increase in the number and range of cyber-attacks. This paper illustrates the variety of cyber-attacks experienced in India related to the Covid-19 pandemic and correlations between key events related to pandemic & cyber-criminal campaigns is explained. The analysis proceeds to utilise how cyber-criminals leveraged key events and governmental announcements to carefully craft and design cyber-crime campaigns.

KEYWORDS— Cybercrime, covid-19, cyber-attack, work from home, pandemic.

I. INTRODUCTION

Cybercrime is a wide array of illegal actions that are made possible by access to an information technology structure. They include electronic fraud, unauthorized access, ID theft, systems interference, and many more actions and new types of crime that came into existence with the creation of the original internet. This led to increasing unease concerning the state of one's security in cyberspace, and thus the concept of cybersecurity was conceived. [1]. Despite the fact that the numbers of internet users are growing quickly, access of internet is not evenly distributed between the users of different countries. A major gap is there among developed and developing nations in introducing cyber security practices and arrangements [2]. The coronavirus pandemic (COVID-19) that started in 2020 and suddenly became a pandemic that affected lives of 100s of millions of citizens across the globe with nations imposing lockdowns, curfews, and many other emergency steps to curb the pandemic [3]. Due to lockdowns work from home was encouraged as most of the businesses shifted to online mode for sustainability. This shift of physical to online system of industry and citizenry led to realisation of a level of cyber security concerns and challenges never faced before. This whole scenario revealed a general level of unpreparedness of cyber security in India. [4]

In this paper we aim to study cyber-attacks related to the COVID-19 pandemic. This paper highlights cyber-attacks and campaigns which typically follow events such as announcements of policy. This allows us to track how quickly cyber-attacks and crimes were witnessed. Thus, the main aim of the present paper is to provide further understanding of the relationship between the changes in daily activities brought about by the COVID-19 pandemic and cybercrime in India.

II. CHANGE IN TECHNOLOGY DUE TO COVID-19

Ever since the Covid1-19 pandemic, people are using the Internet for work, study, shop, visit the doctor, entertain and for many other daily routine activities, as a result, the demand for online communication systems have increased, with some fixed and mobile operators reporting a 60% increase in Internet traffic. India's internet consumption rose by 13% since the nationwide lockdown was put in place to check the spread of Covid-19, according to telecom ministry data that showed Indians consumed 308 petabytes (PB) or 308,000 terabytes (TB) of data daily on an average for the week beginning March 22. According to the department of telecom, which collated reports from service providers, the daily average consumption in this period was 9% higher than 282PB data used on March 21 (the day the Janta curfew was announced) and 13% more than March 19, when consumption was 270 PB. The change reflected how people consumed more streaming content and logged on to work from home, which was also captured in how data demand from residences rose as compared to commercial areas. [6]

Like most organisations, crime also went digital during the pandemic. As per the National Crime Records Bureau (NCRB) report for the year 2020, cyber-crime surged 12% across the country, even as other crimes such as murder, theft and cheating witnessed a drop due to the national and regional lockdowns. Uttar Pradesh crossed Karnataka to emerge as the state with the highest number of cybercrimes reported in 2020. It was followed by Karnataka and Maharashtra. States such as Tamil Nadu, Telangana, Assam, Bihar and Odisha have witnessed sharp jumps in cybercrimes over the previous year. [7].

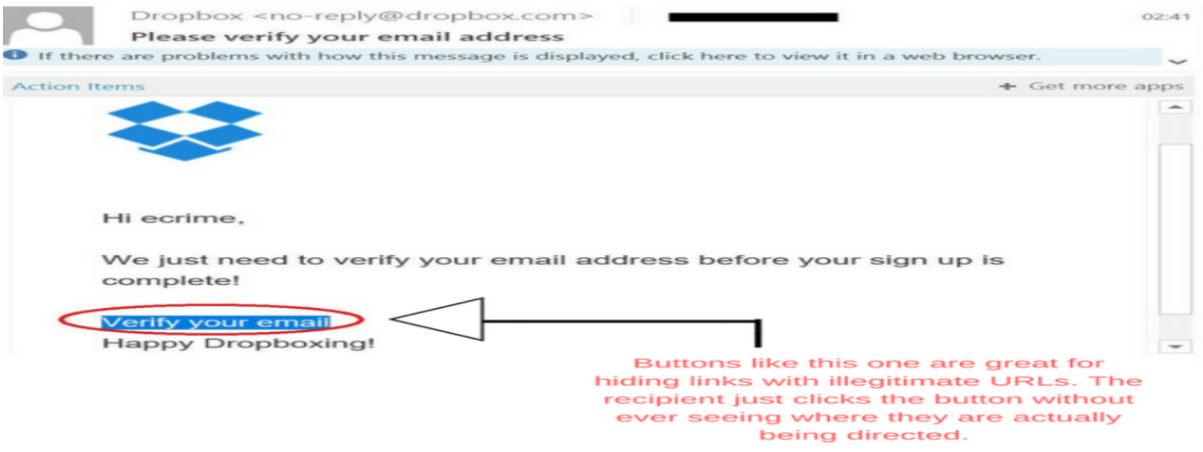
The spike in cybercrimes and attacks has effected the wallets and personal data, given the sharp increase in the percentage of India workforce remotely working as a result of the nationwide lockdown instituted by the government. According to data by National Cyber Security Coordinator (NCSC), cyber criminals had launched thousands of "fraud portals" related to the coronavirus. These sites have lured thousands of Indians eager to contribute to the fight against coronavirus into making donations. Many of these phony sites are quite sophisticated, virtually indistinguishable from their genuine counterparts. [5]

III. Cyber Attacks

Cybercrime and cybersecurity are like two sides of the same coin. They are opposites but cannot exist without each other. A cyberattack is where an attacker tries to gain unauthorized access to a system for the purpose of theft, extortion, disruption or other nefarious reasons. The person who carries out a cyberattack is termed as a hacker/attacker. . When an attack is carried out, it can lead to data breaches, resulting in data loss or data manipulation. Organizations incur financial losses, customer trust gets hampered, and there is reputational damage. To put a curb on cyberattacks, we implement cybersecurity. Cybersecurity is the method of safeguarding networks, computer systems, and their components from unauthorized digital access. The COVID-19 situation has had an adverse impact on cybersecurity with more people working remotely or online post-2020. [7] Many of the methods cybercriminals use to breach organizations rely on human error. Even the sharpest employees can become your greatest weakness if they click on a malicious link without realizing it.

NN. Phishing Attacks

Phishing attacks attempt to steal information from users or trick them into downloading malware by sending malicious emails or text messages (SMS) that look like real requests but are, in fact, a Scam.



Dropbox email asking users to verify their email address that's actually a phishing attack [8]

Phishing attacks are one of the most prominent widespread types of cyberattacks. It is a type of social engineering attack wherein an attacker impersonates to be a trusted contact and sends the victim fake mails. Unaware of this, the victim opens the mail and clicks on the malicious link or opens the mail's attachment. By doing so, attackers gain access to confidential information and account credentials. They can also install malware through a phishing attack. [9]

OO. Malware Attacks

Malware refers to many different types of malicious software designed to infiltrate, spy on, or create a backdoor and control an organization's systems or data. This includes ransomware, worms, Trojans, adware, and spyware. Experts report that malware usage is up almost 800% since early 2020. Malware has the potential to cause major data breaches and severely disrupt business operations. Microsoft was the victim of a major ransomware attack, where WannaCry took advantage of a weak spot in their operating system and displayed the following message to banks, health care providers, manufacturers, and other businesses across the globe.[9]

C. Denial-of-Service Attack

A Denial-of-Service Attack is a significant threat to companies. Here, attackers target systems, servers, or networks and flood them with traffic to exhaust their resources and bandwidth. When this happens, catering to the incoming requests becomes overwhelming for the servers, resulting in the website it hosts either shut down or slow down. This leaves the legitimate service requests unattended. It is also known as a DDoS (Distributed Denial-of-Service) attack when attackers use multiple compromised systems to launch this attack.[10]

D. Man in the Middle

Man-in-the-middle (MitM) attacks, also known as eavesdropping attacks, occur when attackers insert themselves into a two-party transaction. Once the attackers interrupt the traffic, they can filter and steal data. [11]

E. Zero Day Exploit

A zero-day exploit hits after a network vulnerability is announced but before a patch or solution is implemented. Attackers target the disclosed vulnerability during this window of time. Zero-day vulnerability threat detection requires constant awareness.[12]

Review of cyber-attacks in India related to Covid-19 Pandemic

With pandemic disrupting businesses and with remote working becoming reality, cyber criminals have been busy exploiting vulnerabilities. Year 2020 saw one of the largest numbers of data breaches and the numbers seem to be only rising. According to Kaspersky's telemetry, when the world went into lockdown in March 2020, the total number of brute force attacks against remote desktop protocol (RDP) jumped from 93.1 million worldwide in February 2020 to 277.4 million 2020 in March—a 197 per cent increase. The numbers in India went from 1.3 million in February 2020 to 3.3 million in March 2020. From April 2020 onward, monthly attacks never dipped below 300 million, and they reached a new high of 409 million attacks worldwide in November 2020. In July 2020, India recorded its highest number of attacks at 4.5 million. In February 2021—nearly one year from the start of the pandemic—there were 377.5 million brute-force attacks—a far cry from the 93.1 million witnessed at the beginning of 2020. India alone witnessed 9.04 million attacks in February 2021. The total number of attacks recorded in India during Jan & Feb 2021 was around 15 million. [13].

The Prime Minister's Citizen Assistance and Relief in Emergency Situations Fund (PM CARES Fund) was created on 27 March 2020, following the COVID-19 pandemic in India. The stated purpose of the fund is for combating, and containment and relief efforts against the coronavirus outbreak and similar pandemic like situations in the future. [14].

The number of cybercrime cases reported in the national capital spiked during last year's lockdown period, from nearly 2,000 in March to over 4,000 in May, as fraudsters adopted new methods to cheat people, the Delhi Police said on Friday. According to data shared by the city police, 62 per cent of the cybercrimes reported in Delhi were related to online financial fraud, 24 per cent were related to social media and 14 per cent to other cybercrimes. The data showed that from March to May 2020, when restrictions were in place due to the outbreak of COVID-19, there was a rise in cases related to cybercrimes. It went up from around 2,000 such cases in March to more than 4,000 in May. [15]

People created fake government websites providing jobs to doctors and nurses for COVID patients. There were people selling sanitisers, PPE kits, food and groceries with help of fake websites and cheating people [16].

Barracuda Networks researchers also add that cybercriminals prefer to use well-known email providers like Gmail because they are free, easy to register, and "have a higher reputation in the market." Many times, attackers customised malicious email addresses using terms like 'principal,' 'head of department,' 'school,' and 'president' to make them look authentic. Similarly, emails carrying the phishing links are sent with eye-grabbing subject lines such as COVID-19 New Updates, COVID-19 School Meeting, COVID-19 Update, and Follow Up Right Now, to grab the receiver's attention which tracks cloud-based security threats across various sectors, educational institutions are more than twice as vulnerable to phishing attacks than an average organisation. The company says that its researchers over the last few months, evaluated over 3.5 million spear-phishing attacks executed on various sectors, and it was found that over 1,000 schools, colleges, and universities in India were affected. Researchers at Barracuda also noticed that phishing attacks on educational institutes were mainly carried using two methods between July and September - email scams and service impersonation. In many cases, Gmail accounts were the primary medium for hackers to launch phishing attacks, accounting 86 percent of total spear-phishing attacks, the report notes. These emails via Gmail were part of carefully-crafted business email compromise (BEC) attack, that is a form spear-phishing attack, the company notes [17].

A prominent school in Kolkata ditched online classes after hackers sneaked into several lectures and displayed obscene videos on the screen and threatened the students and teachers. The hackers used abusive language and threatened students with rape and murder. As a result, teachers had to suspend the online classes. The Coronavirus-induced lockdown has left schools and colleges with no other option but to conduct online classes. But with educational institutes going online, the risks and vulnerabilities have increased accordingly [18].

There has been a sharp rise in people working from home as a precautionary measure towards lowering the spread of Covid-19. As a result, work from home related tools have seen an uptick in demand and usage. The video-conferencing app Zoom has enjoyed a surge in demand the last few months and this has caught the attention of malware writers[19].

SonicWall Capture Labs threats research team has observed malicious Android apps that use the name, user interface (UI) elements and parts of code of the legitimate Zoom app to infect unsuspecting users [22].

Cybercriminals are encouraging people to register through fake APK files. In such a situation, the Indian Computer Emergency Response Team (CERT-In) has issued a new advisory by alerting people about the fake CoWin vaccine registration app. These fake apps are gaining popularity from viral SMS. In its officially advisory, CERT-In said that the SMS message carries a link that installs the malicious app on Android-based devices. The app also gains unnecessary permissions that attackers could leverage to acquire user data such as contact list," CERT-In said. Scammers are using multiple variants of SMSs to trap citizens into their fishy net. The apps are downloaded as APK files so that scammers can easily inject malware into their smartphones. [24]

Table I
Description of Covid 19 related cyber attacks

Id	Attack type	Description	Month of Attack	Ref
1	Phishing , Financial Fraud	Fake Accounts created of PMCARES fund	March 2020	[14]
2	Phishing , Financial Fraud	Fake social media accounts targeting people in distress over the pandemic	May 2020	[15]
3	Phishing , Financial Fraud	Fake government websites proving jobs to doctors and nurses for Covid-19 patients.	April-May 2020	[16]
4	Phishing , Financial Fraud	Fake websites selling sanitizers ,PPE kits, food and groceries.	April-May 2020	[26]
5	Hacking , Financial fraud	Hackers gaining access to bank accounts by using fake websites and links of KYC platforms.	April-May 2020	[17]
6	Phishing	Spear phishing attacks on schools and colleges	July-Sept 2021	[18]
7	Phishing	Fake COWIN websites for duping people and stealing information.	Jan –March 2021	[19]
8	Phishing , Financial Fraud	Fake websites claiming to be govt. portal for Covid-19 vaccination registration	Jan – March 2021	[27]
9	Hacking	Online classes in school and colleges disrupted by hackers and posting of obscene images.	March-June 2021	[20]
10	Phishing	Fake video conferencing android apps compromising personal details of users is circulated on play store.	May 2021	[22]

Table II
Correlations between events and cyber-criminal campaigns

EVENT DATE	EVENT	RELATED CYBER-CRIME CAMPAIGNS	TYPE OF CYBER CRIME
MARCH 2020	Govt. of India announced nationwide lockdown across India with only essential services kept out of its purview.	Fake Social Media Accounts targeting people in distress over pandemic.	Phishing, Financial Fraud
		Fake websites and portal claiming to be govt. websites providing jobs to doctors and nurses for Covid-19 patients.	Phishing, Financial Fraud
MARCH –APRIL 2020	Govt. of India announced PMCARES fund following the Covid-19 pandemic in India.	Fake bank accounts and UPI id's created and being circulated on social Media.	Phishing, Financial Fraud
APRIL 2020		Hackers gaining access to bank accounts by fake websites and links for KYC verification of various online payment platforms.	Hacking Phishing, Financial Fraud

APRIL –MAY 2020	Demand of PPE kits, sanitizers,antiviral medications etc. increases as nationwide lockdown is extended.	Fake websites selling Sanitizers, PPE kits, fake medicines , Food and groceries.	Phishing, Financial Fraud
JULY – SEPT 2020	Online Teaching and work from home is encouraged and use of online video conferencing software is increased.	Online classes in school and colleges disrupted by hackers and posting of obscene images.	Hacking Phishing,
		Fake video conferencing apps compromising personal details of users is circulated	Hacking Phishing,
JAN – MAY 2021	Indian government web portal for COVID-19 vaccination registration is launched.	Fake COWIN websites for duping people and stealing information	Phishing, Financial Fraud
JAN – MAY 2021		Fake websites claiming to be govt. portal for Covid-19 vaccination registration	Phishing, Financial Fraud

Vulnerability to Cybercrime at the Onset of the COVID-19 Pandemic

With the rapid and massive shift online, there is concern that individuals are insufficiently trained, are using unfamiliar tools, are inexperienced with the technology, and, as a result, becoming easy targets for cybercriminals. The increase in cybercrime during the pandemic mainly impacted individuals rather than organizations. With cybercrime, individuals often actively participate in the fraudulent process to which they become the victim, such as by responding to a phishing email and providing private information. Individuals may not be sufficiently suspicious, may not be able to detect fraudulent messages, or may not pay sufficient attention to stop a fraudulent process. Spammers make their offers look like they come from official institutions or legitimate businesses that people routinely trust, and use persuasion principles found to be effective in legitimate emails. Vulnerability to online fraud involves a combination of psychological and demographic factors, and online activities, where victim profiles vary with the type of cybercrime. Overall, victims of online fraud are older, impulsive, sensation seeking, have an addictive disposition, and follow routine activities placing them at risk for fraud like online banking and shopping. Individuals with healthcare concerns may have increased susceptibility to health related phishing. The importance of human factors in cybercrime cannot be overstated. The problems of cybersecurity cannot be solved just by adding more technology. Humans are involved in every aspect of cybersecurity in our complex, interconnected, digitalized world as software and hardware developers, systems administrators, managers, end users, consumers, attackers, and victims. The ways in which humans interact with each other, process information and make decisions, handle workload and stress, and interface with technology are fundamental to cybersecurity. Humans often place inappropriate levels of trust in automated systems.

CONCLUSION

Our lives have been radically altered by a pandemic that is considered to be among the most widespread in history. The shift to the digital world undoubtedly creates new opportunities and platforms for motivated offenders to engage in various illegal activities. This shift increases the number of suitable targets, as millions of people are confined to their homes and forced to work, study, and socialize online. The solution lies in adopting cybersecurity as a way of life. The realisation that attackers are trying to get into our lives is the starting point of journey of self-discovery. Research in cybersecurity is shifting its focus from technology point of view towards human factors as well. Researches need to realise that the focus of most attacks is on human vulnerabilities, it is critical to understand how humans routinely interface with technology, including cyber security products. This paper gives rise to the recommendation that governments, the media and other institutions should be aware that announcements and the publication of stories are likely to give rise to the perpetration of associated cyber-attack campaigns which leverage these events. The events should be accompanied by a note / disclaimer outlining how information relating to the announcement will be relayed. This research has shown correlation between events and cyber-attacks. Further study should analyse and determine whether a predictive technique can be used to confirm this relation.

REFERENCES

1. -Aleksandra Pawlicka a, *, Michał Choras a,b , Marek Pawlicki a,b , Rafał Kozik a,b , A $10 million question and other cybersecurity-related ethical dilemmas amid the COVID-19 pandemic R. Sarath, K. Boddu, and V. R. Bendi, “Cyber Crime and Security , a Global Vulnerable Coercion : Obstacles and Remedies,” vol. 7, no. 5, pp. 5–8, 2017.
2. S. Alotaibi, A. Alruban, S. Furnell, and N. Clarke, “A Novel Behaviour Profiling Approach to Continuous Authentication for Mobile Applications,” no. February, 2019.
3. J. R. C. Nurse, “Cybercrime and You: How Criminals Attack and the Human Factors That They Seek to Exploit,” in The Oxford Handbook of Cyberpsychology. OUP, 2019

4. Over 900 cases of fraud involving cards, net banking registered in Apr-Sep 2018 https://economictimes.indiatimes.com/industry/banking/finance/banking/over-900-cases-of-fraud-involving-cards-net-banking-registered-in-apr-sep-2018/
5. [Online]. Available https://www.hindustantimes.com/india-news/india-s-internet-consumption-up-during-covid-19-lockdown-shows-data/story-ALcov1bP8uWYO9N2TbpPlK.html
6. [Online]. Available https://auth0.com/blog/the-7-most-common-types-of-cybersecurity-attacks-in-2021/
7. [Online]. Available https://www.bankinfosecurity.com/locky-returns-via-spam-dropbox-themed-phishing-attacks-a-10250
8. [Online]. Available: https://www.simplilearn.com/tutorials/cyber-security-tutorial/types-of-cyber-attacks
9. [Online]. Available:https://blog.malwarebytes.com/cybercrime/2017/05/wanacrypt0r-ransomware-hits-it-big-just-before-the-weekend/
10. [Online]. Available: https://auth0.com/blog/the-7-most-common-types-of-cybersecurity-attacks-in-2021/
11. [Online]. Available https://www.simplilearn.com/tutorials/cyber-security-tutorial/types-of-cyber-attacks
12. [Online]. Available https://www.business-standard.com/article/technology/india-becomes-favourite-destination-for-cyber-criminals-amid-covid-19-121040501218_1.html
13. [Online]. Available: https://economictimes.indiatimes.com/tech/internet/cyber-chiefs-warning-as-hackers-target-pms-covid-fund/articleshow/74877953.cms?from=mdr
14. [Online]. Available: https://timesofindia.indiatimes.com/city/bengaluru/bengaluru-cybercrooks-cash-in-on-covid-angst-cheat-ailing-patients-and-relatives/articleshow/82788482.cms
15. [Online]. Available: https://www.business-standard.com/article/current-affairs/delhi-reported-steep-spike-in-cybercrimes-during-lockdown-period-police-121022000052_1.html
16. [Online]. Available https://www.business-standard.com/article/current-affairs/delhi-reported-steep-spike-in-cybercrimes-during-lockdown-period-police-121022000052_1.html
17. [Online]. Available https://www.news18.com/news/tech/online-education-due-to-covid-19-is-causing-massive-spike-in-cyber-attacks-on-schools-colleges-3024551.html
18. [Online]. Available https://economictimes.indiatimes.com/news/india/beware-of-covid-online-scams-here-are-five-dangerous-websites/fake-covid-vaccine-apps/slideshow/82903248.cms
19. [Online]. Available https://www.outlookindia.com/website/story/india-news-online-classes-disrupted-by-hackers-schools-fight-flood-of-obscene-and-abusive-messages/355420
20. [Online]. Available https://www.outlookindia.com/website/story/india-news-online-classes-disrupted-by-hackers-schools-fight-flood-of-obscene-and-abusive-messages/355420
21. [Online]. Available https://securitynews.sonicwall.com/xmlpost/fake-android-zoom-video-meeting-apps-harbor-malware-adware-components/
22. [Online]. Available https://www.dnaindia.com/technology/report-these-fake-cowin-vaccine-registration-apps-could-steal-your-personal-data-2890486
23. [Online]. Available https://www.tribuneindia.com/news/ludhiana/fraudsters-target-people-on-pretext-of-covid-19-vaccine-registration-367731
24. [Online]. Available https://www.business-standard.com/article/current-affairs/delhi-reported-steep-spike-in-cybercrimes-during-lockdown-period-police-121022000052_1.html
25. [Online]. Available https://economictimes.indiatimes.com/news/india/beware-of-covid-online-scams-here-are-five-dangerous-websites/fake-covid-vaccine-apps/slideshow/82903248.cms
26. Scott Monteith1 & Michael Bauer2 & Martin Alda3 & John Geddes4 & Peter C Whybrow5 & Tasha Glenn6, Current Psychiatry Reports (2021 Increasing Cybercrime Since the Pandemic: Concerns for Psychiatry:)

ENHANCED I-SEP PROTOCOL USING FITNESS FUNCTION FOR CLUSTER HEAD SELECTION

Navneet kaur[1], Er. Gurpreet Singh[2]
Department of computer science & engineering, Punjabi University, Patiala
[1]niwassekhon92@gmail.com
[2]gurpreet.1887@gmail.com

ABSTRACT - Sensor Networks build a self-organized network by utilizing a range of low-cost sensor nodes scattered across the region. These sensor nodes detect files and send the information to the base station, which uses a lot of energy. The creation of an energy-efficient WSN routing protocol is a significant problem. Clustering is a novel technique for increasing the energy efficiency of a sensor network. In heterogeneous protocols, two or three node energy levels are often defined, however there is now a large range of energy levels in heterogeneous WSNs. For many years, energy efficiency has been a highly hot and demanding research topic for WSNs. It is not possible to change the sensor batteries for a significant number of sensor nodes installed in a hostile environment. The suggested technique is a clustering algorithm based on an improved mobile agent-based FFI-SEP protocol. The major objective of this work is to create an energy-efficient WSN protocol based on the SEP protocol.

KEYWORDS - Fitness Function, Cluster Head, Sensor Networks, Stable Election Protocol, Throughput, Energy Efficiency, Clustering, Stable Election Protocol.

I. INTRODUCTION

1.1 INTERNET OF THINGS(IOT)

IoT is an evolving analysis field that integrates a variety of fields of study. The key IoT concept is to link all devices to the Internet, such as home appliances, mobile phones, vehicles, houses, robots, machines, and so on. The concept is to provide a virtual equivalent for all applications in the real world that detects essential data from the environment in order supply advanced end- user services[1]. Energy efficiency is one of the major concerns when using these devices, as communication and computation on the restricted device could rapidly discharge its battery capacity resources. In the scenario Nodes depend on self-organizing multi-hoping networking methodologies of WSN, which could operate in the absence of a base station and comparable strategies could be implemented for IoT. A main research problem is the creation of DR algorithms that could effectively observe paths among mobile nodes. Due to the obvious rise in computational burden on mobile nodes, dynamic networks do not use traditional algorithms. Connect performance and network topology could differ while a message packet has been routed[2]. The maintenance of a high-quality connection therefore involves regular measurement and upgrading of routing paths. Clustering is known to be the most efficient way to solve the performance issues of ad hoc networks and inevitably opposes their usage in the sense of IoT due to similar difficulties.

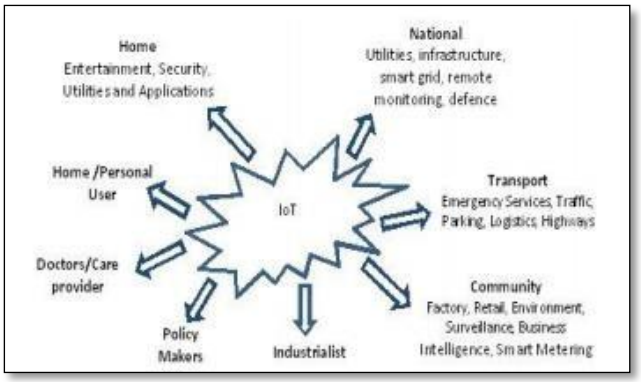


Figure 1: Internet of Things (IOT)[1]

TABLE 1
Evolution of IOT Technologies

Technology	Time Span	Description
RFID	1999	Passive identification, wireless networks
WSN	2005	WSN, Cloud computing, Web2.0, low energy communication
Smart Things	2012	Mobile computing, cooperating operation of objects, connecting devices
IOT	2017	Advanced sensor fusion, faster wireless connectivity, predictive analysis

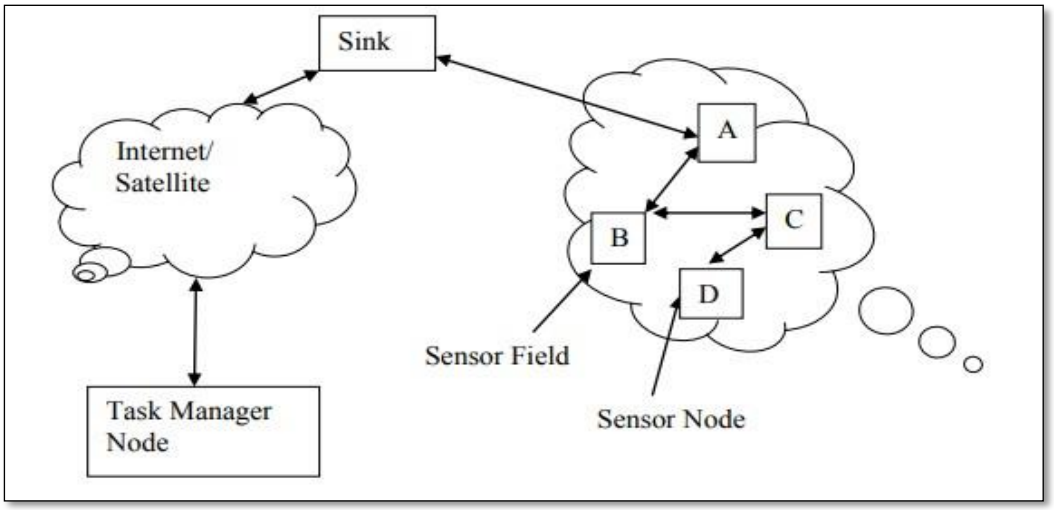
1. WIRELESS SENSOR NETWORK

Mass production of small size and low sensors has become commercially viable as a result of recent technological advances. The deployment of large-scale Wireless Sensor Networks (WSNs) has been made possible by continuous advancements in wireless communication systems [3]. WSN is a big network comprised of a collective of spatially distributed autonomous sensors interlinked by communication links to monitor physical or environmental conditions like temperature, sound, vibration, pressure, movement, or particulates at numerous places and collaboratively transmit their data via the network to a centralized location.

The sensor nodes as well as access points are the most important components of WSN (BSs). Sensor nodes form a system and interact with one another either directly or via nodes. The BS allows one or more nodes to communicate with the user. The base Station allows communication either directly or via current wireless connections.

Figure 2 depicts a typical sensor network topology. Sensor nodes are depicted as smaller pieces. Every other sensor node is made up of five parts: a sensor unit, an analog digital converter (ADC), a central processing unit (CPU), a power unit, as well as a communication unit. The ADC converts the sensory information as well as notifies the CPU about what the sensor unit is responsible of. The communication server gets commands or queries as well as sends data from the CPU to the outside world. The CPU recognizes the ADC command or query, managing or monitoring power as needed, procedures receive messages, evaluates the next hop to the sink, etc [4].

Figure 2: Architecture of Sensor Network[2]



CHARACTERISTICS

The following are the factors of wireless sensor nodes[5]:

- Sensor nodes are densely distributed.
- Sensor nodes are prone to failures.
- The topology of a sensor network alterations on a regular basis.
- The set of sensor nodes could be various orders of magnitude greater as compared to that of ad hoc network nodes.
- Sensor nodes primarily use a telecast communication paradigm, whereas the majority of ad hoc networks rely on point-to-point communication systems.
- Sensor nodes primarily use a telecast communication paradigm, whereas the majority of ad hoc networks rely on point-to-point communication systems. Because of the large quantity of overhead as well as the huge number of sensors, sensor nodes would not have Global identification [6].

REQUIREMENTS

Precise to the implementation WSNs are made up of hundreds to thousands of low-power multi- functional sensor nodes. It works without human intervention in an area with minimal computation as well as sensing capability. They request the preceding conditions[4]:

- Sensor nodes are cheap.
- Data collection procedures extend the channel's life.
- Sensor nodes can form their own networks without any external setup;
- Sensor nodes should be willing to coordinate and collate their relevant data.

APPLICATIONS

WSN is utilized in a multitude of scenarios, including medical purposes where sensor nodes involve battlefield monitoring, smart missiles, and medical products such as patient diagnosis or monitoring, environmental monitoring, industrial uses, incident response (including power grid monitoring, and so on.), as well as other miscellaneous applications like commercial applications at home and industries to be remote-controlled. WSNs are particularly useful for controlling or managing numerous manufacturing processes [7].

ISSUES

The faster advancement of mobile sensing devices is primarily because of the many design requirements. The greatest difficulty in WSNs is developing a system while keeping in mind restricted battery life, resource limitations, random as well as huge integration, as well as a dynamic and uncontrolled atmosphere. Enabling secure transmission while maintaining network energy consumption is a complicated issue in WSN studies because sensors have limited battery system for processing and transmitting data to the base station. Aside from effectiveness in regards to energy preservation, bandwidth is an essential criterion for WSNs[8]. Even so, the number of packets obtained by the BS enables evaluation of data transmission dependability as well as throughput[9] . Another difficult issue in a static WSN is network Coverage, which is a

significant aspect in sensor nodes, particularly in WSNs that require large accessibility of collected data. Clustering enables sensors to effectively organize their local interactions in order to accomplish policy objectives such as power efficiency, throughput, coverage, as well as scalability in a routing algorithm.

1. CLUSTERING

Clustering is a process for grouping similar objects into classes[10]. A cluster is a set of information objects that are similar to one another within the same group but distinct from objects in many groups. A cluster could be defined at a high - level of abstraction: a grouping of data items could be defined as a group. Clustering is a well-known energy consumption method. There are various groups in which the Sensor nodes (SN) are divided. Every group contains a single Cluster Head and a large number of Members' nodes. All devices transmit messages to the Cluster Head (CH) and it gathers all analyzed information, compiles it, as well as delivers it to the Base Station (BS) for the final processing of the information. When compared to the standard Sensor Node , every cluster head depreciates at a quick speed. Clustering reduces the distance among nodes as well as base stations for communication, reduces energy consumption, or extends the network's lifetime.

The blue circle in Figure 3 symbolizes a sensor node (SN) made up of smaller batteries, resulting in a shorter lifespan, less computation power, as well as less memory. A cluster Head (CH) is defined through a red circle, and it collects and sending information from the cluster's various nodes to the Base Station (BS). Cluster Head is intended to be extremely effective, secure, as well as believed by the entire sensor network. A base station is a node element with computation complexity, energy, and action capacities. The arrow denotes the linkages that link the sensor node and the base station.

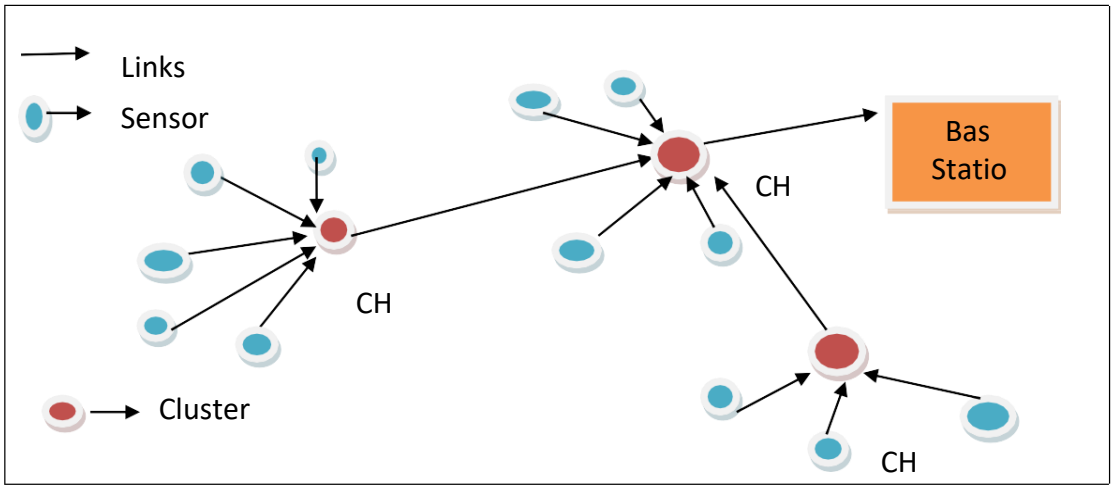


Figure 3: Cluster in WSN

Clustering in sensor nodes has been highly recommended by the academic researchers as a solution to sensor network scalability, energy, as well as lifetime issue of the network.

2.1 IMPORTANCE OF CLUSTERING IN WSN

1) Data from sensors can only be transported so far.
2) As a result of network subsistence, there is an enhancement in energy competence [11].
3) Clustering contributes to network performance.
4) By incorporating CH-level data, we were effective to lessen energy consumption.

5) Decreased network security as well as channel contention.
6) Interaction bandwidth preservation.

2. *Overview Stable Election Protocol*

In the SEP routing protocol, the election of new CH and the formation of new clusters occurs on a regular basis for each round. This, in turn, results in unnecessary energy consumption due to routing overhead, influence the production of IoT devices attached to the sensor network. As per the traditional SEP method, a CH elected in the current round will be unable to engage in the CH election process in the following round . However, there may be cases where a CH did not expend enough energy in the preliminary round and is therefore qualified for the CH election process in the subsequent round[20]. It is also possible that a sensor with a relatively lower amount of energy is appointed as CH in the subsequent selection process, resulting in the network's premature death. Furthermore, every round involves the formation of a new cluster, which absorbs node power by sending messages such as ADV (advertisement) as well as ACK (acknowledgement) to CHs back and forth. The aforementioned constraint in SEP encourages researchers to explore as well as develop an efficient CH replacement strategy.

2 RELATED WORK

(Bensaid et al.,(2020) suggest a new clustering method based on FCM for WSN-based IoT applications . To pick the optimal CH in each transmission round, the method utilizes an FCM strategy to create the clusters as well as a minimization of the overall consumed energy in each cluster. To evaluate the behavior as well as implement methodology, a comparison with the LEACH protocol is addressed. The experiments demonstrated that the suggested FCM system involves network lifetime by increasing residual energy by 50 percent[12].

Behera et al.,(2019) concentrates on an efficient cluster head election approach that turns the cluster head position between many nodes that have a faster speed than others. To select the next group of cluster heads for the network that is suitable for IoT applications like environmental observing, smart cities, and systems, the technique calculates remaining energy, power consumption, and an average value of cluster heads. Based on simulation results, the different version outperforms the LEACH approach by increasing throughput by 60%, lifetime by 66%, and residual energy by 64%[13] .

Ahmed et al.,(2019) Energy-Efficient Scalable Routing Algorithm brings an energy-efficient clustering as well as hierarchical routing algorithm (EESRA). The suggested application's idea is to maximize network lifespan despite rising network size. The method computes a three-layer structure to reduce cluster head load as well as randomly select cluster head selection. Furthermore, in order to achie ve a hybrid WSN Authentication scheme, EESRA employs multi-hop transmissions for intra-cluster communications. The article describes EESRA to other WSN routing protocols in terms of system performance as network scale changes. According to simulation results, EESRA outperforms benchmarked schemes in terms of load balancing and power generation on large scale WSNs[14].

Wang et al.,(2018) Recognize a WSN-assisted IoT network that is hierarchical. The base station serves as the network's central controller, as well as nodes are circulated at random throughout the network. The information is collected by CH nodes and routed to the base station through a multi-hop network. The proposed energy-efficient clustering routing method's focus is to enhance energy efficiency as well as network lifetime. First, designers suggest an uneven cluster formation system to improve energy efficiency while also balancing the various traffic loads in each layer. Following that, the CH node rotates adaptively focused on residual energy as well as relative position to maintain energy consumption within each cluster. Eventually, the routing path would be created new opportunities to prevent the energy hole issue when the CH node spins or the energy level changes. According to simulation outcomes, suggested energy- efficient clustering routing method outperforms existing routing protocols in means of network lifetime, throughput, as well as energy efficiency[15].

Ali et al.,(2018) To extend the lifespan of WSN-based IoT, design an efficient energy-efficient clustering protocol (IEECP). The suggested IEECP is divided into 3 areas that must be completed in order. First, an ideal set of clusters for the overlapping balanced clusters is motivated. The balanced-static clusters are then formed using an altered fuzzy C-means method and a framework to decrease and stability the energy consumption of the sensor nodes. Finally, cluster heads (CHs) are chosen in optimal locations by rotating the CH function between many cluster members using a new CH selection-rotation method that integrates a back-off timer methodology for CH selection as well as a rotation process for CH rotation. The suggested protocol reduces as well as balances node energy consumption by enhancing the clustering structure, in which IEECP is useful for channels that take a significant lifespan. The results demonstrate that the IEECP outperforms existing protocols[16].

Durairaj et al.,(2020) The suggested hybrid multi-hop routing algorithm seeks to extend the life of a WSN utilized in a widely spread network. Despite the fact that its performance is superior, chain-based CH selection as well as data routing via MST consumes more energy, decreasing network lifetime. To avoid this, three novel multihop routing methods are suggested: Shortest Grid Routing, Partitioned Super Grid Direct Routing, and Shortest Super Grid Routing. The initiatives are evaluated using modeling as well as evaluated using real-time innovation for various network situations. The effectiveness of the proposals is evaluated by comparing to the recently suggested approach protocol utilizing metrics such as first node death, energy consumption, stability, as well as network capacity. The outcomes show increased network

lifetime as well as dependability, demonstrating the utility of the suggested routing algorithms for CPS-based infrastructure networks[17].

M.Ganesan et al.,(2019) To create the best use of the available energy, a clustering model is developed. The goal of this method is to organize IoT devices into clusters, with an advanced node, i.e. a specific node with strong computing abilities, serving as the cluster head (CH) as well as nearby IoT devices joining the cluster as cluster members. This proposed technique could be used to collect data on patients and infrastructure details like biological data such as heart rate and blood pressure, temperature, parking facility, pharmacy, and so on. The effectiveness of the proposed model is studied and compared to a traditional solution that does not require a clustering process. Two performance measures, such as the amount of data transmissions as well as the total set of data transmissions to the cloud platform, are for experimentation. The computation findings confirm that the inclusion of the clustering algorithm significantly reduces energy consumption[18].

A. Selva et al.,(2021) This research proposed a clustering-based novel Fuzzy C-means clustering approach as well as an enhanced ECC-ElGamal encryption approach to suggest secure data transmission while also reducing transmission time. Using the cluster coefficient, the ideal cluster heads CHs in the network are chosen, with the most neighboring nodes, the shortest distance from the base station, as well as the highest residual energy. The base station then predicts the encrypted key utilizing adapted co-efficient based key generation and transmits the nodes via CH. The information is then encrypted utilizing improved ECC-ElGamal encryption, and the decryption method is enabled. As a result, secure transmission with authentication is already accomplished, as well as an overall output analysis has been performed as well as presented with existing techniques. When the suggested method's performance is contrasted to existing ECC and RSA approaches, the overall processing time is reduced, as well as the network is extremely valid in the case of differing the nodes from 10 to 100[19].

Sushanta et al.,(2020) an enhancement to the existing stable election procedure (SEP) that executes a threshold-based CH selection for a heterogeneous network The threshold ensures that energy is distributed uniformly among member and CH nodes. To transfer the network load evenly, sensor nodes are classified into three categories: normal, intermediate, as well as advanced, based on their initial energy supply. According to the simulation outcomes, the suggested approach outperforms SEP and DEEC procedures by 300 percent in network lifetime as well as 56 percent in throughput[20].

3. PROPOSED METHODOLOGY

The proposed protocol is termed as FFI-SEP, i.e. fitness function based I-SEP. In this work, the cluster head, selection process will be modified by adding fitness function value for each node. The fitness function will be computed using two parameters namely cluster compactness and distance from base station. The cluster compactness parameter will reflect the proximity of cluster head and its members.

$$\text{Cluster compactness} = \frac{d_0}{\sum_{i=1}^{n} (D_s - D_i)/n}$$

Where $d0$ is the communication range of the node 's'
$Ds - Di$ is the distance between node 's' and neighbour node 'i' 'n' is the number of neighbours
Now the fitness function of the node will be:

$$f = \text{Cluster compactness} + \frac{d_0}{b_s} D$$

This fitness function will be included while computing the probability of the node to become cluster head.
Once the cluster heads have been selected, they will form cluster with their neighbouring nodes. After cluster formation, the next step is the data forwarding step in which instead of transmitting data directly to the base station the cluster head will make use of crow search optimization to select the best neighbouring cluster head. The optimal neighbour chosen will act as relay node to forward the data to the base station. This marks the end of one round. Now, at the beginning of the second round, the decision of the retaining the cluster head will be taken according to the concept of threshold value given in the existing scheme.

4. CONCLUSION

Wireless Sensor Networks has a range of low-cost sensor nodes spread across the region to create a self-organized network. Such sensor nodes sense the files and transmit the information to the base station requiring a lot of energy. The development of an energy-efficient WSN routing protocol is a major challenge. Clustering is an innovative method used to improve the energy efficiency of the sensor network. Generally, two or three node energy levels are established in heterogeneous protocols, but there is currently a wide variety of energy levels in heterogeneous WSNs. Energy management and energy efficiency have been a very hot and challenging subject of research for WSNs for many years. As a large number of sensor nodes are deployed in a very harsh environment, it is not feasible to change the sensor batteries for them. The proposed algorithm is an enhanced FFI-SEP protocol, which is a clustering algorithm. The main purpose of this dissertation is to establish an energy-efficient protocol for WSN based on the SEP protocol. The research work

presents energy efficient methodologies for energy efficiency selection of the cluster heads based on the fitness function. Final results are concluded by comparing the performance of the network based on average residual energy, number of alive nodes, number of dead nodes and throughput of the network. Simulation results depicts that the proposed algorithm showed better results than the previous technique. The findings show that the FFI-SEP was stronger than the existing algorithm, because it reduced the rate of packet losses.

5. REFERENCES

[1] Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M.and Ayyash, M. , "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," IEEE Communica- tions Surveys & Tutorials, Vol. 17 No. 4, pp. 2347-2376,2015.

[2] Yao, X., Wang, J., Shen, M., Kong, H., & Ning, H., "An Improved Clustering Algorithm and Its Application in IoT Data Analysis" Computer Networks,2019.

[3] Xu, M., Ma, L., Xia, F., Yuan, T., Qian, J., & Shao, M., "Design and Implementation of a Wireless Sensor Network for Smart Homes", 7th International Conference on Ubiquitous Intelligence & Computing and 7th International Conference on Autonomic & Trusted Computing,2010.

[4] Warrier, M. M., & Kumar, A., "An Energy Efficient Approach for Routing in Wireless Sensor Networks", Procedia Technology, 25, 520–527,2016.

[5] Yick, Jennifer, Biswanath M., " Wireless sensor network survey", Computer Networks- Elsevier, Vol. 52, No. 12 ,pp. 2292-2330,2008.

[6] Nagamalar T & Rangaswamy TR, "Sleeping Cluster based Medium Access Control Layer Routing Protocol for Wireless Sensor Networks", Journal of Computer Science, vol. 8, no. 8, pp. 1294-1303,2012.

[7] Lewis FL., "Wireless Sensor Networks', Smart Environments: Technologies, Protocols, and Applications", 2004.

[8] Singh, Shio, Kumar, Singh, M,P & Singh, D., "Energy-efficient Homogeneous Clustering Algorithm for Wireless Sensor Network", International Journal of Wireless & Mobile Networks (IJWMN) ,Vol. 2., No. 3 pp. 49- 61,2010.

[9] Nabila,Labraoui, Mourad Gueroui, ,Aliouat & Jonathan, "Reactive and adaptive monitoring to secure aggregation in wireless sensor networks", Telecommunication systems, vol. 54,no. 1, pp. 3-17,.2013.

[10] H. Kasban, S. Nassar & Mohsen A. M. El-Bendary, "Medical images transmission over Wireless Multimedia Sensor Networks with high data rate", Analog Integrated Circuits and Signal Processing ,vol. 108, pp. 125–140 , 2021.

[11] Geetha, V., Kallapur, P. V., & Tellajeera, S., "Clustering in Wireless Sensor Networks: Performance Comparison of LEACH & LEACH-C Protocols Using NS2", Procedia Technology, Vol. 4,pp. 163–170,2012.

[12] Bensaid, R., Ben Said, M., & Boujemaa, H., "Fuzzy C-Means based Clustering Algorithm in WSNs for IoT Applications", International Wireless Communications and Mobile Computing (IWCMC),2020.

[13] Behera, T. M., Mohapatra, S. K., Samal, U. C., "Residual Energy Based Cluster-head Selection in WSNs for IoT Application", IEEE Internet of Things Journal, 2019.

[14] Ahmed, E. F., Omar, M. A., Wan, T.-C., & Altahir, A., "EESRA: Energy Efficient Scalable Routing Algorithm for Wireless Sensor Networks", IEEE Access, 2019.

[15] Wang, Z., Qin, X., & Liu, B., "An energy-efficient clustering routing algorithm for WSN- assisted IoT", IEEE Wireless Communications and Networking Conference (WCNC),2018.

[16] Ali Abdul-Hussian H., Wahidah Md ," An Improved Energy-Efficient Clustering Protocol to Prolong the Lifetime of the WSN-Based IoT", Vol. 14, No. 5,2018.

[17] Durairaj, U. M., & Selvaraj, S., "Two - Level Clustering and Routing Algorithms to Prolong the Lifetime of Wind Farm based WSN", IEEE Sensors Journal, 2020.

[18] M.Ganesan, Dr.N.Sivakumar, "An energy efficient IoT based Healthcare System based on clustering technique", Third International Conference on Electronics Communication and Aerospace Technology (ICECA) , 2019.

[19] A. Selva Reegan & V. Kabila, "Highly Secured Cluster Based WSN Using Novel FCM and Enhanced ECC-ElGamal Encryption in IoT", Wireless Personal Communications volume 118, pp. 1313–1329, Feb 2021.

[20] Trupti Mayee B., Sushanta M., Umesh Chandra Samal, "I-SEP: An Improved Routing Protocol for Heterogeneous WSN for IoT-Based Environmental Monitoring", IEEE Internet of Things Journal, Volume: 7, Issue: 1, pp. 710-717, Jan. 2020.

A COMPREHENSIVE REVIEW ON PLANT DISEASE DETECTION USING MACHINE LEARNING

Deepak Sidana[#1], Neelofar Sohi [#2]
[#]*Computer Engineering Department, Punjabi University, Patiala"*
[1]sidana.deepak15@gmail.com
[2]neelofarsohi7@gmail.com

ABSTRACT— The identification of crop diseases is the key to minimizing productivity and quantity losses in plant/crop and agricultural goods. The manual control of crop diseases is very arduous. It needs a great deal of effort, experience in Plant diseases, and often-excessive processing time. The Plants disease detection technique based on the image processing and machine learning. In this research work, technique is designed for the Plant disease detection using machine Learning. The Plant disease detection system consists of five major steps: image acquisition, image pre-processing, segmentation, feature extraction, and classification. There are various classifiers are available to abstract the features of an image such as random forest, artificial neural network, support vector machine (SVM), fuzzy logic, K-means method, KNN etc. The proposed model will implemented in MATLAB and results will compared with other available classifiers in terms of accuracy, precision, recall for Plant quality prediction.

KEYWORDS— Plant Disease detection; Machine Learning; SVN; KNN; Random forest; Artificial neural network; Image Processing

INTRODUCTION

In the Growing world, population has brought a lot of pressure on our agriculture. It is crucial to obtain maximum yield from crop in order to sustain the population and the economy. Crop/Plant diseases are the main source of plant damage, which cause economic and production losses in agricultural areas. Owing to distressed climatic and environmental conditions, occurrence of plant diseases is on the rise.

There are various types of diseases in plants, variety of symptoms such as spots or smudge arising on the plant leaves, seeds and stanches of the plant. In order to manage these diseases effectively, there is a need to introduce automatic method of plant surveillance that can scrutinize plant conditions and apply knowledge-based solutions to detect and classify various diseases. Machine Learning is an intrinsically appropriate framework to support this problem.

A variety of techniques has been proposed recently for identification and classification of plant diseases from images using Machine Learning. While these automated techniques have paved way for remote monitoring and expert surveillance of plant diseases, there are challenges of accuracy and robustness that need to be addressed for reaping practical benefits from these techniques. This report presents experimental results of using various Machine Learning techniques for the task of plant disease identification and classification.

Modern approaches such as machine learning and deep learning algorithm has been employed to increase the recognition rate and the accuracy of the results. Various researches have taken place under the field of machine learning for plant disease detection and diagnosis, such traditional machine learning approach being random forest, artificial neural network, support vector machine(SVM), fuzzy logic, K-means method, Convolutional neural networks etc.

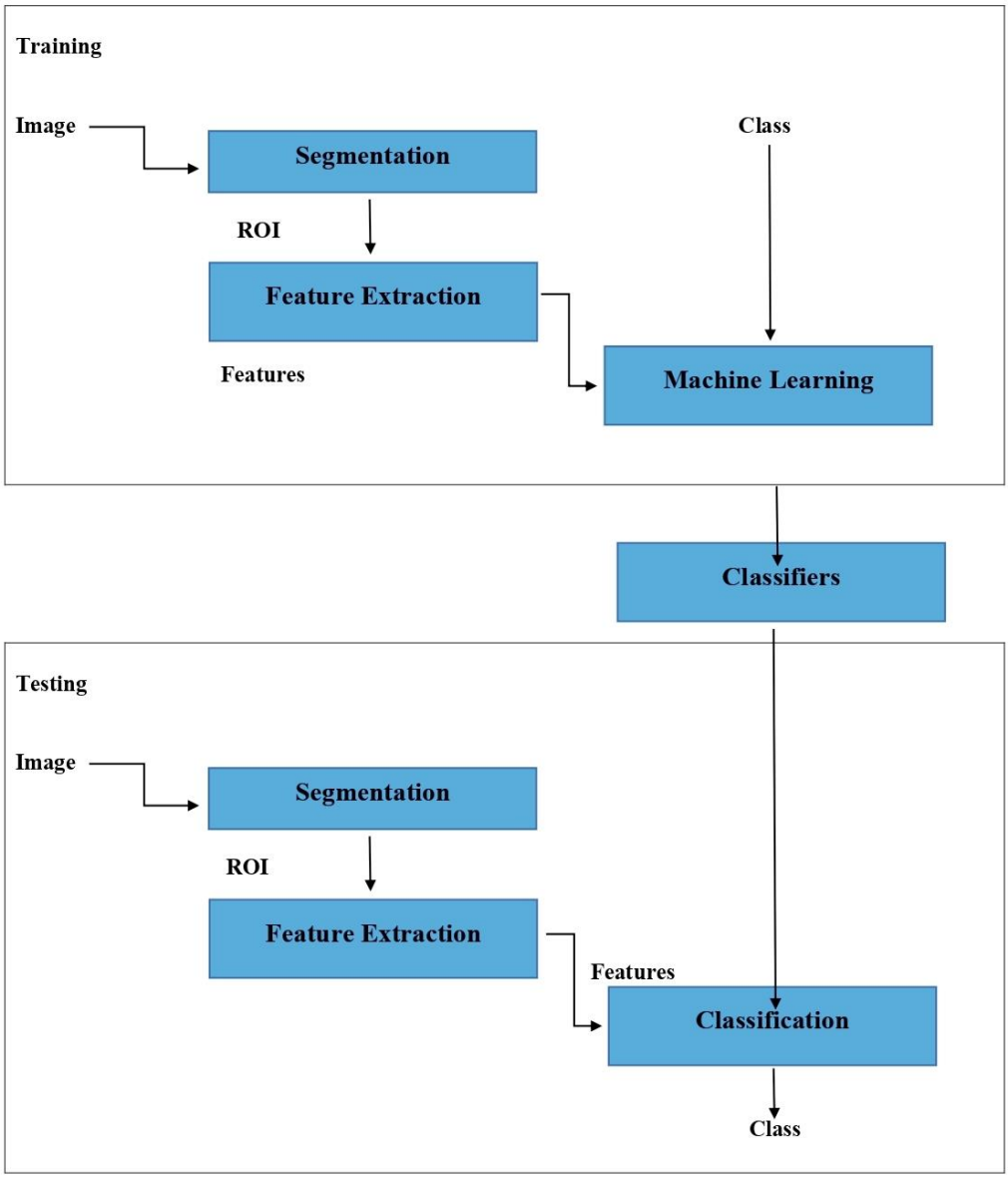


Fig. 1 Machine Learning Framework for Automated Plant Disease Analysis [1]

Fig.1 shows the basic setup of automated plant disease analysis using Machine Learning Techniques. In the training section, segmentation is performed from images and features extracted, these features are then used for classification.

EXISTING WORK ON PLANT DISEASE DETECTION

Different researchers have used different Image processing, segmentation and classification techniques in there research work. Some of these are listed below:

Nitesh Agrawal et al (2017) [2] have studied and proposed the automatic detection of grape leaf diseases is presented, this method is based on K-means as a clustering procedure and Multiclass SVM as a classifier tool using some texture feature set. They have implemented the code and tested on three diseased leaf they are: Black rot, Esca, LBlight and healthy leaves.

Shima Ramesh et al (2018) [3] have worked and finds the disease Detection algorithm for papaya leaves plant Using Machine Learning. The goal is to create datasets for diseased and healthy leaves from Random Forest to classify the diseased and healthy images.

K. Jagan Mohan et al (2016) [4] worked for detection and recognition of diseases from Paddy Plant Leaf Images. They used an image processing system that can identify and classify the various paddy plant diseases affecting the cultivation of paddy namely brown spot disease, leaf blast disease and bacterial blight disease.

M. P. Vaishnnave et al (2019) [5] have researched for the detection and classification of Diseases on Groundnut Leaf using KNN classifier. The team worked on a software determination to robotically classify and categorize groundnut leaf diseases.

Md. Ashiqul Islam et al (2020) [6] proposed a Machine Learning based Image Classification for Papaya Disease Recognition. This research is mainly required to support agriculture to make it highly effective and helpful particularly for papaya cultivation. User captures an image via mobile app, sends it to the system for disease Detection, and compare some algorithms accuracy. They Used K-means clustering for Segmentation of image.

Minu Eliz Pothen et al (2020) [7] have proposed a solution for detection of Rice Leaf Diseases Using Image Processing. They come with a solution to detect the most common disease Leaf smut, Bacterial leaf blight and Brown spot on Rice leafs. SVM+HOG classifiers used for the research work.

L. Sherly Puspha Annabel et al (2019) [8] worked on Artificial Intelligence - Powered Image-Based machine learning system for Tomato Leaf Disease Detection. They proposed tomato leaf disease detection, which comprises of four different phases that includes image preprocessing, segmentation, feature extraction and image classification. GLCM and random forest classifier used in the research work.

N. Nandhini et al (2020) [9] have done the research on Feature Extraction for Diseased Leaf Image Classification using Machine Learning. Diseased leaf image classification performed in two stages, (i) extraction of color and shape feature (ii) classification using machine learning. SVM, KNN and Decision trees, the total three number of different classifiers was used in the course work.

COMPARISON OF DIFFERENT TECHNIQUES FOR PLANT/CROP DISEASE DETECTION USING MACHINE LEARNING

Study	Description	Classifier	Accuracy	Disease Recognized	Year Of Publication	Conclusion / Future Work
Detection and Recognition of Diseases from Paddy Plant Leaf Images	To develop an image processing system that can identify and classify the various paddy plant diseases affecting the cultivation of paddy namely brown spot disease, leaf blast disease and bacterial blight disease	SVM, KNN	91.10%, 93.33%	This work mainly concentrates on three main diseases of paddy plant namely Brown spot, Leaf blast and Bacterial blight	2016	For future work, some alternative methods can be used to extract features and some other classifiers can be used to improve the result accuracy.
Grape Leaf Disease Detection and classification Using Multi-class Support Vector Machine	This project tries to attempt for improvement in classifying the leaf diseases	Support Vector Machines (SVM)	90%	Major disease commonly observed in Grapes plant such as Brot, Esca and LBlight	2017	Along with LAB and HSI, features can be extracted from YCbCr Color space and also wavelet based features can be added to the database which might increase the accuracy.
Plant Disease Detection Using Machine Learning (papaya leaves)	To create datasets for diseased and healthy Leaves from Random Forest to classify the diseased and healthy images	Random forest Classifier	70%	--	2018	The accuracy can be increased when trained with vast number of images and by using other local features together with the global features such as SIFT (Scale Invariant Feature Transform), SURF (Speed Up Robust Features) and DENSE along with BOVW (Bag Of Visual Word)
Detection and Classification of Groundnut Leaf Diseases using KNN classifier	Software determination to robotically classify and categorize groundnut leaf diseases	K Nearest Neighbour (KNN).	NA	A) Early Leaf spot: Cercospora arachidicola B)Late leaf spot: Phaeoisariopsis personatum C) Rust: Puccinia arachidis D)Bud Necrosis	2019	The examination work can added extended to decrease the false classification by using extra classifiers for feature extraction among the various groundnut crop diseases

Machine Learning based Image Classification of Papaya Disease Recognition	This research is mainly required to support agriculture to make it highly effective and helpful particularly for papaya cultivation	Forest, K-means clustering, SVM and CNN).	98.4%	Anthracnose, Black Spot, Phytophthora, Powdery Mildew, Ring spot	2020	The present research work will be extended to work on a large dataset to predict the factor that is mainly responsible for the papaya diseases
Detection of Rice Leaf Diseases Using Image Processing	To detect the most common disease Leaf smut, Bacterial leaf blight and Brown spot on Rice leafs	SVM+HOG with polynomial Kernel function	94.6%	Leaf smut, Bacterial leaf blight and Brown spot on Rice leafs	2020	SVM+HOG with polynomial kernel function can be used to detect other diseases in rice leaves as well as the leaves of other plants.
AI-Powered Image-Based Tomato Leaf Disease Detection	Novel tomato leaf disease detection is proposed which comprises of four different phases that includes image pre-processing, segmentation, feature extraction and image classification	GLCM and random forest	94.1%	Acterialspot, Lateblight, Tomatomosaic.	2019	NA
Feature Extraction for Diseased Leaf Image Classification using Machine Learning	The approach to leaf image-based disease recognition	SVM, KNN and Decision trees	91%, 85%, 82%	NA	2020	The approach suggested uses the colour characteristics of diseased leaf images to classify the images of the lesion

CONCLUSIONS

A studied above a comparative study carried out on several of machine learning techniques for recognition of plant disease in this review. SVM and KNN classifier used by many authors for classification of diseases when compared with other classifiers. The result shows that SVN classifier detects more number of diseases with high accuracy. In future, other classification techniques in machine learning may use for disease detection in plants so that it will help farmers an automatic detection of all types of diseases in crop to be detected.

REFERENCES

[1] Asma Akhtar, Aasia Khanum, Shoab A. Khan, Arslan Shaukat, "Automated Plant Disease Analysis (APDA): Performance Comparison of Machine Learning Techniques 2013", 11th International Conference on Frontiers of Information Technology

[2] Nitesh Agrawal, Jyoti Singhai and Dheeraj K. Agarwal "Grape Leaf Disease Detection and classification Using Multi-class Support Vector Machine" Proceeding International conference on Recent Innovations is Signal Processing and Embedded Systems (RISE-2017) 27-29 October,2017

[3] Shima Ramesh, Mr. Ramachandra Hebbar, Niveditha M, Pooja R, Prasad Bhat N, Shashank N, Mr. P V Vinod, "Plant Disease Detection Using Machine Learning", 2018 International Conference on Design Innovations for 3Cs Compute Communicate Control

[4] K. Jagan Mohan, M. Balasubramanian, S. Palanivel,"Detection and Recognition of Diseases from Paddy Plant Leaf Images", International Journal of Computer Applications (0975 – 8887) Volume 144 – No.12, June 2016

[5] M.P.Vaishnnave, P.Srinivasan, G.ArutPerumJothi, K.Suganya Devi, "Detection and Classification of Groundnut Leaf Diseases using KNN classifier", Proceeding of International conference on System Computation Automation and Networking 2019

[6] Md. Ashiqul Islam, Md. Shahriar Islam, Md. Sagar Hossen, Minhaz Uddin Emon, Maria Sultana Keya, Ahsan Habib, "Machine Learning based Image Classification of Papaya Disease Recognition", Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020) IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1

[7] Minu Eliz Pothen, Dr.Maya L Pai, "Detection of Rice Leaf Diseases Using Image Processing", Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020), IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2

[8] L. Sherly Puspha Annabel, V. Muthulakshmi "AI-Powered Image-Based Tomato Leaf Disease Detection", Proceedings of the Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019), IEEE Xplore Part Number:CFP19OSV-ART; ISBN:978-1-7281-4365-1

[9] N. Nandhini, R. Bhavani "Feature Extraction for Diseased Leaf Image Classification using Machine Learning", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), Jan. 22-24, 2020, Coimbatore, INDIA

A SURVEY ON BRAIN TUMOR CELL IMAGE SEGMENTATION AND DETECTION TECHNIQUES

Harjeet Singh[1], Harpreet Kaur[2]
1. *Harjeet Singh Lecturer Thapar Polytechnic College, Patial,*
2. Harpreet Kaur Assistant Professor Punjabi University *Patiala*
1. harjeetsinghbiviaan@gmail.com
2. harpreet.ce@pbi.ac.in

ABSTRACT:—— Tumor in our brain is one of the main causes of death in human beings. It becomes more dangerous or malignant if not detected at an early stage and right time. So brain tumor detection is the first stage in the requirement for finding better results or cures the cancer problem. In the past few years, there are so many cases of a brain tumor in children, adults and old age persons. Many scientists and researchers are working together to develop the best technique in order to find the brain tumor category and the affected area. Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) pictorial representation of imaging is the best way for finding an affected area, shape, size, and stage of the tumor. MRI is better than Computed Tomography that mostly doctor prefers for tumor diagnosis. There are many automated algorithms and methods in machine learning that are helpful for the doctor to detect tumors with high accuracy to provide treatment to the patients. Some popular approaches are used for detecting tumor-like Convolutional Neural Network (CNN), Deep Neural Network (DNN), Growing Convolutional Neural Network (GCNN), Fuzzy C mean and weighted fuzzy kernel clustering (WKFCOM), Fuzzy C-mean and Support Vector Machine(SVM), Particle Swarm Optimization (PSO and Artificial Neural Network (ANN) etc. This paper sheds the light on the study of various methods and also summarized various approaches from 2014 to 2021 with theoretical measurement of performance, False Positive Rate (FPR), True Positive Rate (TPR)/ sensitivity/Recall for detecting brain tumor cell.

KEYWORDS: Brain Tumor, Image Segmentation, Cancer, Detection, Cell Inage Segmentation

1. INTRODUCTION

The human brain is the most important and complex part of the body which is the combination of cells that perform a specific task. The working of the brain depends on these billion cells. So it can be said that there may be some abnormal cells inside the brain that create major problems in our brain and these abnormal group of cells in our brain affect the functionality of healthy cells. In this situation, the brain will be out of control and do an undefined task we call this stage as a brain tumor. We can divide the brain tumor into two types namely: *Harmless or lower-grade* (type I and II) and *Cancerous tumors or high-grade* (type III and IV). Harmless tumor cells are conservative (Benign) and are known as non-aggressive cells. They grow very slow anywhere in our body but they are not dangerous. On the other hand, malignant tumors are very aggressive and dangerous as they grow with very high speed in the body. These cells can be originated from the human body itself called *primary malignant* or can be created from some external issues in human body parts which are called a *secondary malignant* or cancerous brain tumor. Functional magnetic resonance imaging (fMRI) is the latest methodology in the field of medicine. It is used to detect the tumor by researchers and doctors during treatment phase. MRI provides detailed useful data related to the internal partition and problems in brain cells because of a clear depiction of the picture. Presently researchers represent various automatic techniques using fMRI scan when it used as a tool for fetching and decode medical pictures in the labs. Some techniques are used for detecting brain tumors using MRI images. Support Vector Machine (SVM) and Neural Networks (NN) is the frequently used technique for their accurate prediction of brain Tumor over the last few years **[1].** But in the present time, the Deep learning method is at the top in machine learning for detecting brain tumor.

Cell image segmentation (CIS): The cell is the fundamental, functional and biological unit in all living body parts having different size, shape and function. Cell image segmentation is the process of identifying individual cells or structures with in an image. It is the key step for analysis workflows in biomedical images. It includes segmenting or dividing the pixels in the image into region of interests (ROI) shown in Fig.1. It has been observed that image extraction, study of cells and their sub-cellular alcove is required to various research areas for example cellular dynamics characterization in normal and pathologic conditions [11] as well as drug making. Recently enhancement in high-resolution fluorescent microscopy brick the way for exact representation of the cells and their sub-cellular structures.

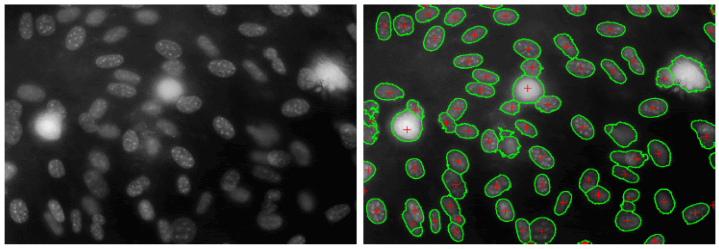


Fig.1: A fluorescence microscopy image (left) and the output of segmented cells (right). The boundaries of the detected cells and of the ground truth cell centroids are plotted with green curves and red pluses, respectively [12].

To achieve to better result and develop a well-controlled system we have to face many challenges as required good knowledge of the technology, cell structure, algorithm. There are many challenges that has to face, discussed major ones in following. As the cells in our body are tightly connected with each other and have a challenge to detect and analyses them as shown in Fig.2.

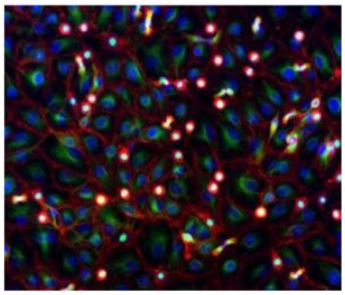


Fig. 2**:** Discontinuous object boundaries which are still required to be connected is still a challenge

It is complicated to show the relationship between brain cells as it requires many numbers of nodes in superficial architectures. For example, SVM, K-nearest neighbor is used for this purpose. So, because of the above-said reason, MRI and CT scan methods expand at a very fast speed to become the state of the art in various health information fields such as Biocomputing, medical information and analyzing medical images. Some of the techniques are discussed below.

1.1 Magnetic Resonance Imaging (MRI): In the medical field, MRI plays a very important role in neurology for detecting and envisioning the internal structure of our mind and other cranium building block. It is very helpful for envisioning the makeup in the main position namely prime, coronet and sagged. Fig.3 [2] and Fig.4 [3] depict three makeup of the body of brain. Fig.2 represents the prime, Sagged, Coronet position of the Brain taken from MRI. There are three main series that come on brain MRI during scanning depend upon the time known as T1 low grade, T2 high grade, flair type. Low grade (T1) is produced by short time to echo (TE) and repition time (RT) on the other hand T2 grade is produced by long time to echo (TE) and repition time. T1 is most commonly used during these days as we can see in Fig.3. The third one is flair type which is very important and mostly used for finding problems in our brain. The MRI scanning process has very high resolution and contrast features and less radiation that's why it is more successful than a CT scan [4].

It can find the flow of blood and hidden secular defects. It is also able to catch the problems related to the nerve system. MRI imaging methods are used to detect critical issues regarding the human brain. This is also applicable to other problems that are concerned with the brain such as Mental Decay problem [5], Parkinsonism problem [6] and mental deterioration [7], etc. The process of MRI is traceable in the human body as the electric and magnetic field is applied on hydrogenic particles present in our body by consuming Radio Frequency (RF) pulsation after that hydrogenic nuclei uses the power and pass that into form of an electric wave after pausing the Radio Frequency (RF). The particle discharges the power and comes in its initial phase; this situation is called a relax phase. The time consumed by this process is known as a relaxing time. We can distinguish the disorder cells and normal cells from brain images. In the laboratories, the radiologist mainly focuses on 3 particular parts of the brain that is: Gray Matter (GM), White Matter (WM) and Cerebrospinal Fluid Spaces (CSF). Fig.6 represents the normal image of our brain partitioning White Matter, Grey Matter and CSF [8], Fig.5 shows the MR image of Grey Matter, White Matter, and Cerebrospinal Fluid MRI brain image respectively [9]. Grey matter (GM) consists numerous cells bodies and relatively few myelinated axons. White matter (WM) consists relatively few cell bodies and is possessed mainly of large-range myelinated axons. Cerebrospinal Fluid Spaces (CFS) is very neat and clean, black and white body fluid in the brain and spinal cord.

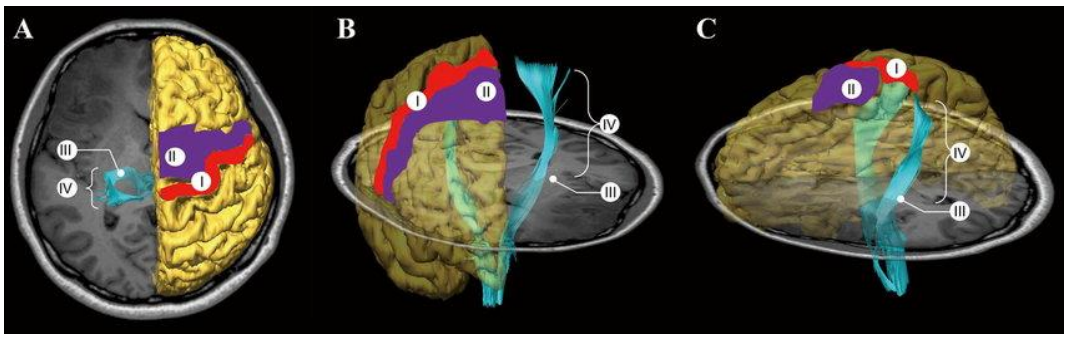


Fig.3. (a) The axial plane, (b) Coronal plane and (c) Sagittal plane of the brain image**[2]**

Fig.3 constitute the category of 3 parts of brain cell. a axial aircraft, b coronal aircraft and c sagittal aircraft. Type I, tumors that invading the precentral gyrus, kind II, tumors that invadeding premotor region and/or supplementary motor regions however did now no longer invade the pre-valuable gyrus, kind III, tumors that invaded or near the posterior limb of the inner pill and sort IV, tumors that invaded different supra-tentorial regions

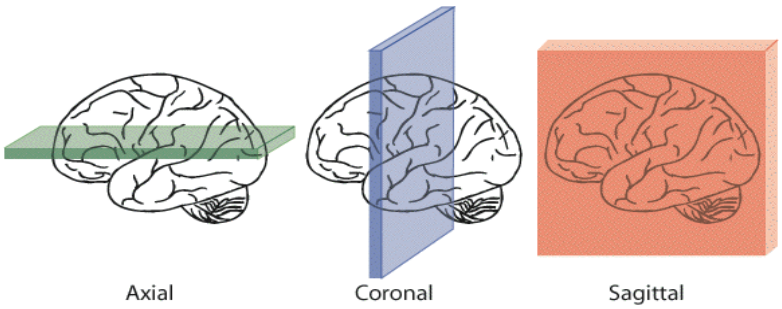


Fig.4: (a) The axial plane, (b) Sagittal plane and (c) Coronal plane of the brain observed by MRI**[3]**

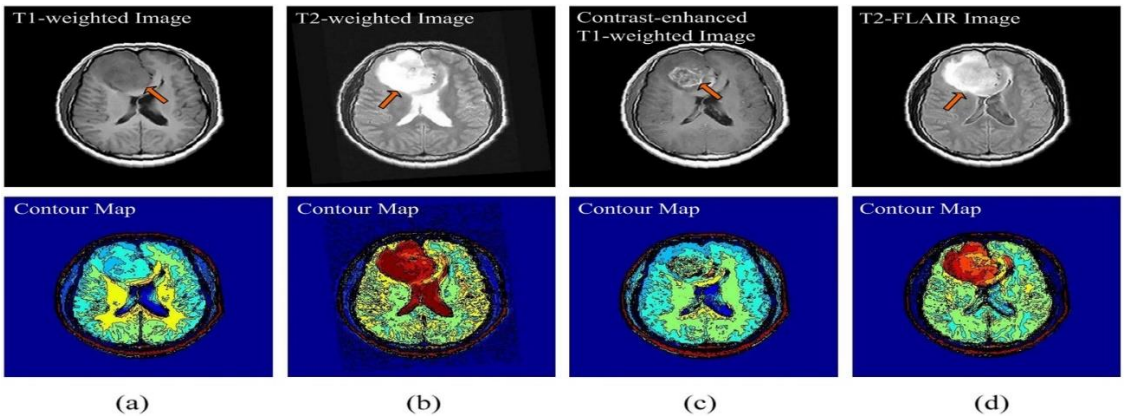


Fig.5: (a) T1- weighted, (b) T2-weighted and (C) Flair **[10]**

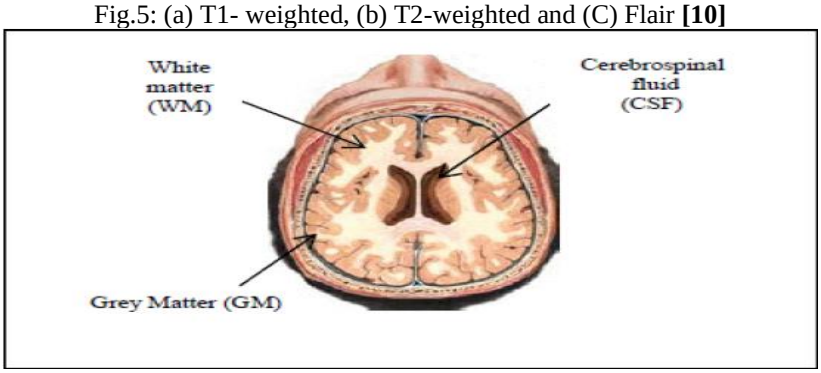


Fig.6: Normal Brain

1.2 Computed Tomography

A Computed Tomography (CT) scan is used in a medical imaging procedure that uses the computer to analyze images from different angles with a combination of many X-Ray Measurements and virtual slices. It is a mostly used technique these days. It generates the data facts and computes them in order to describe the various bodily structures based on their power to absorb the X-Ray beam. After that cross-sectional image is created. This process is repeated on all slides of the image. It provides details and clear information about different parts of the bodies during a CT scan in a very effective manner as compared to X-ray. CT scan can be done for diagnosing any problem in our brain like other body parts such as bones, head, lungs, etc. But MRI is very useful in the case of a brain tumor. Some CT scan images of the brain tumor are given below in Fig.7.

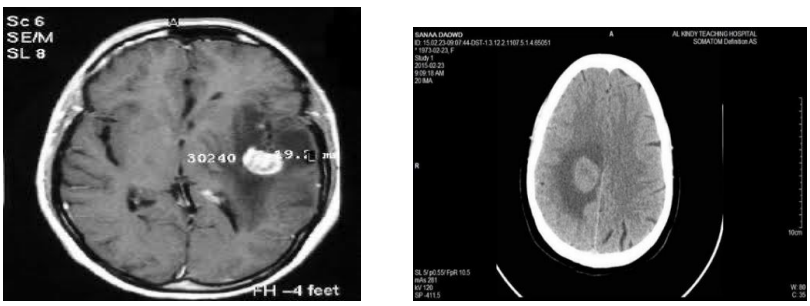


Fig.7: CT Scan of Brain Tumor Cell

2. LITERATURE SURVEY

Several researchers have presented their ideas, methodologies, and implementations for the detection and prediction of Brain Tumor. This section describes the approaches for the detection and classification of brain tumors by using MRI scanning. There are different approaches for the detection of brain tumor but most of them are time-consuming and manual based like a traditional analytical technique to segment the brain problems is done manually based on past results and knowledge in a particular field. The manual segmentation process is very slow or time taking and very prone to misclassify. There are various automated tools and techniques have been proposed in order to find tumor in the brain.

J. Amin et al. suggested an automated approach for segmenting and classifying brain tumours in [1]. Following the segmentation of the area of interest (ROI), which includes intensity, shape, and texture, multiple kernels of the Support Vector Machine (SVM) classifier are used to classify the various phases of malignant or non-cancerous pictures. The suggested technique has been cross-validated on three different datasets- Local, Harvard, and Rider- using AUC (Area under Curve) and ACC (accuracy) performance metrics. The results demonstrate the efficacy of the proposed method. S. Deepak and P.M Ameer focused on identifying various brain cancers in [2], with glioma, meningioma, and pituitary tumours being the most common. Because there are various criteria that contribute to categorization, such as the shape and size of brain tumours exhibiting higher diversity, classification is hampered. Furthermore, different types of tumours tend to have similar appearances, further complicating classification. The utilisation of typical machine learning algorithms is made difficult by this issue. To address this issue, the suggested system provided a contribution, namely, using transfer learning, a greater degree of accuracy was reached compared to earlier models; even with a smaller dataset, a significant increase was accomplished. At the softmax level, the proposed model employed an existing GoogleNet with slight adjustments for the different types of tumour categorization. The accuracy of the CNN-based deep learning GoogleNet model was 92.3 percent, which was enhanced to 97.8 percent with the application of multiclass SVM. Ryo Ito and his colleagues. [3] illustrates a semi-supervised learning approach for brain tumour segmentation from MRI data. This technique outperformed the current registration-based and Deep Neural Network (DNN)-based methods in terms of accuracy. The author has introduced a probabilistic model to overcome the label error problem of registration-based (Label Propagation) methods. Using the Expectation Maximization (EM) algorithm, we can find the true label of a latent (un-labelled) image with the condition that the probability distribution governing those latent images is known. By adding the specific noise to the genuine label, the expected label is predicted. To train this probabilistic model, the DNN model is combined with the EM method. The inaccurate label is recovered to the latent image during the maximising state. The approach has been tested on two different datasets: open benchmark human MR images from the Internet Brain Segmentation Repository (IBSR) and a benchmark brain image dataset.

In [4,7,9,11,12] there is a method for detecting a brain tumor by using the process of segmentation, then the process of checking the stage, size, shape of tumor is performed with help of SVM classifier. Since classification is based on the shape and not based on other factors like size, type, nature because the same type of tumor has same shape and size, so to remove this problem author focus on classification various brain tumor mainly based on glioma, beningioma and pituitary [13]. There is a modified approach i.e CNN based deep learning GoogleNet is used as compared to the traditional method with a small database. It achieves the highest accuracy of 92.3% and can be improved up to 97.8% using multiple class SVM. In [14] the study shows the partially supervised approach of segmentation is used to detect brain tumors. It produces good results than Deep Neural Network (DNN), as there was a problem with an error label. In this, Expectation Maximum (EM) algorithm has been applied and gets a very true result as per guesses and expectation by implementing on Brain Segmentation Repository (BSR) and marmoset brain image dataset. In order to enhance the accuracy of CNN, there is a combo of Stationary Wavelet Transform (SWT) and Growing Convolution Neural Network (GCNN) which is utilized. SWT has been adopted for feature extraction that produces the high result for alternate data as compared to the Fourier transform technique. It boosts up the accuracy of 2% in Peak Signal to Noise Ratio (PSNR). In the study of [15] the process of finding problems in the brain comprises three stages: Firstly, a Deep Learning Model is made. Secondly the k-mean algorithm for dividing MRI images into subparts is applied. In the last normal and abnormal images are separated by using the CNN model. In [16] to examine the brain tumor problem in detail from roots an automated deep learning method is proposed. This method is applied to Harvard medical school MR dataset. It can find critical problems such as stroke, Alzheimer's, autism disease. In [17] the dual approach Kernel-based Fuzzy C means (KFCOM) and Weighted fuzzy kernel clustering (WKFCOM) algorithms are adopted and the result of WKFCOM is better than KFCOM with 2.36% minimum misclassification. In [18] it is observed that the MRI images are further divided into three groups' namely white matter, grey matter, and cerebrospinal fluid spaces by using new method Adaptive fuzzy k means (AFKM). This method is better than the Fuzzy-C-Means (FCM). In [1] the multi procedure approach has been introduced for the segmentation of multi-level features of brain tissues from MRI images. The process involving skull stripping means removing unwanted features from an image by a new technique combo of Fuzzy C mean and SVM is recommended in [19]. After skull stripping, abnormal and normal images are extracted.

2.1 Classification of Brain Tumor

Based on nature or malignancy and benignity, brain tumors can be categorized into different types like the American Association of Neurological Surgeon (AANS) has presented the details of the classification of brain tumors for the researcher and education perspective. As depicted in Fig.8 brain tumor is classified into basic brain tumor (growing slowly) and secondary (growing rapidly) brain tumor. Further classification represents that the basic brain tumors are divided into eleven other branches and sub-branches Fig.9 represents the branches of Gliomas; one of the sub-categories of basic brain tumor which is further expanded into four other types namely *ground level (lowest), Minor, highest* and *superlative grade malignancy.*

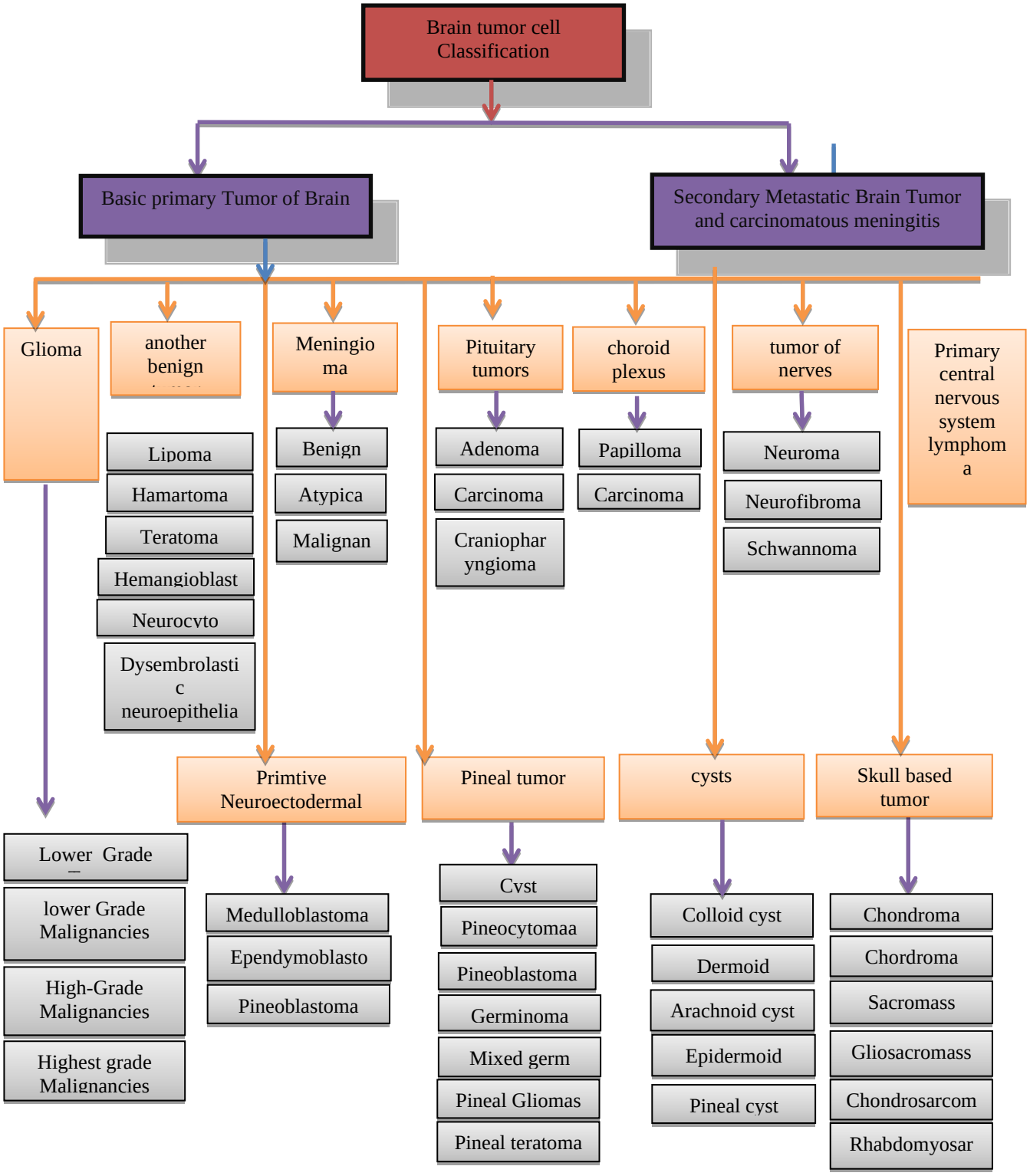


Fig.8: Brain Tumor Classification

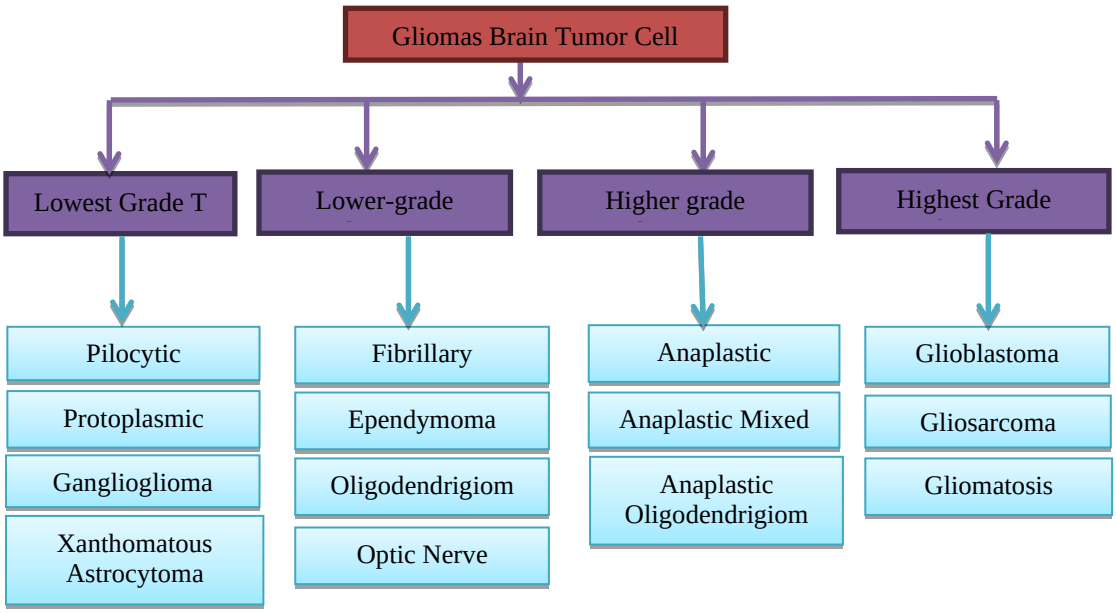


Fig.9: Classification of Gliomas tumor cell

3. METHODS FOR SEGMENTATION AND CLASSIFICATION FOR DETECTING BRAIN TUMOR CELL

In Digital Image Processing (DIP) two main pillars are *Image segmentation* and *image classification* play a very important role in tumor detection. There are several methods used for segmentation and classification of the image. Medically image segmentation is the process to find the region of interest or divide the image into the various regions on basis of foreground and background by using pixel similarities from two dimensional or three-dimensional images captured by a different method like MRI, CT, X-Ray. These techniques divide the picture by differentiating forepart and backdrop on the basis of identical pixels from the 2 Dimensional or 3 Dimensional pictures captured in various methods; fMRI, X-ray, CT, Microscopy, Endoscopy, etc. Fig.10 shows commonly used approaches for classification and segmentation of images. There are mainly two types namely *supervised* and *unsupervised* methods of image classification as shown in Fig.11. Supervised methods are methods in which an experienced person in a particular field manually detects the exact class or problem. On the other hand, unsupervised approach segmentation is performed on the basis of a numeric relationship then grouping the images in the various clusters. In [4,8,9,12], the author has recommended automatic tools and techniques in order to perform segmentation and classification of brain tumor. The author has applied various SVM's kernel classifiers to find out the cancer cell and noncancerous cell and the level of tumor in the image.

This process is done only when segmentation is complete which includes a Region Of Interest (ROI) consist of depth, architecture and text information in the image. The recommended method has been tested on 3 various databases, Local, Harvard, and Rider based on Area under Curve (AUC) and accuracy performance measures. In [13,14,] author has targeted analyzing different brain tumors mainly glioma, meningioma, and pituitary. It seems to be very difficult to analyze the brain tumor because there are so many aspects, characteristics which need to be considered for classification.i.e. the structure and length display upper layer of abnormality so cripple the classification issues, various kinds of tumors bear to present the same look, it also creates interference again in the process of classification. So it can be said that it becomes more critical to handle this problem with old methods. So in **[**15] in order to solve this problem, the recommended technique provides little bit sharing i.e. by using transferring learning a top class of accuracy was accomplished correlated with past methods; a considerable enhancement was accomplished still with the smallest database.

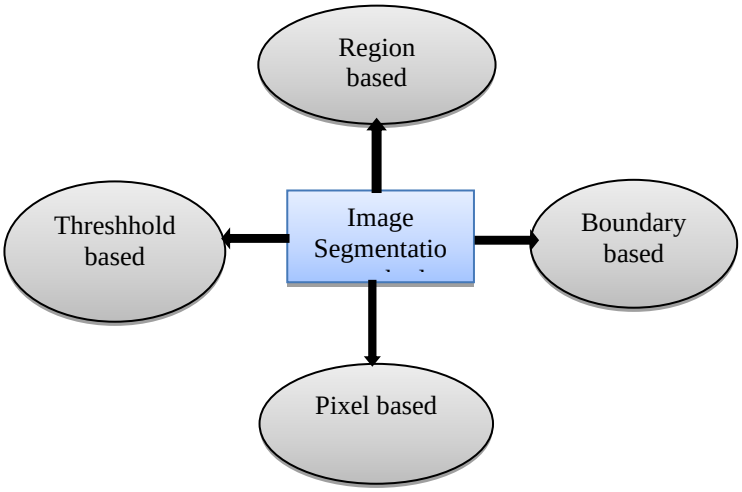


Fig.10: Methods of Image Segmentation

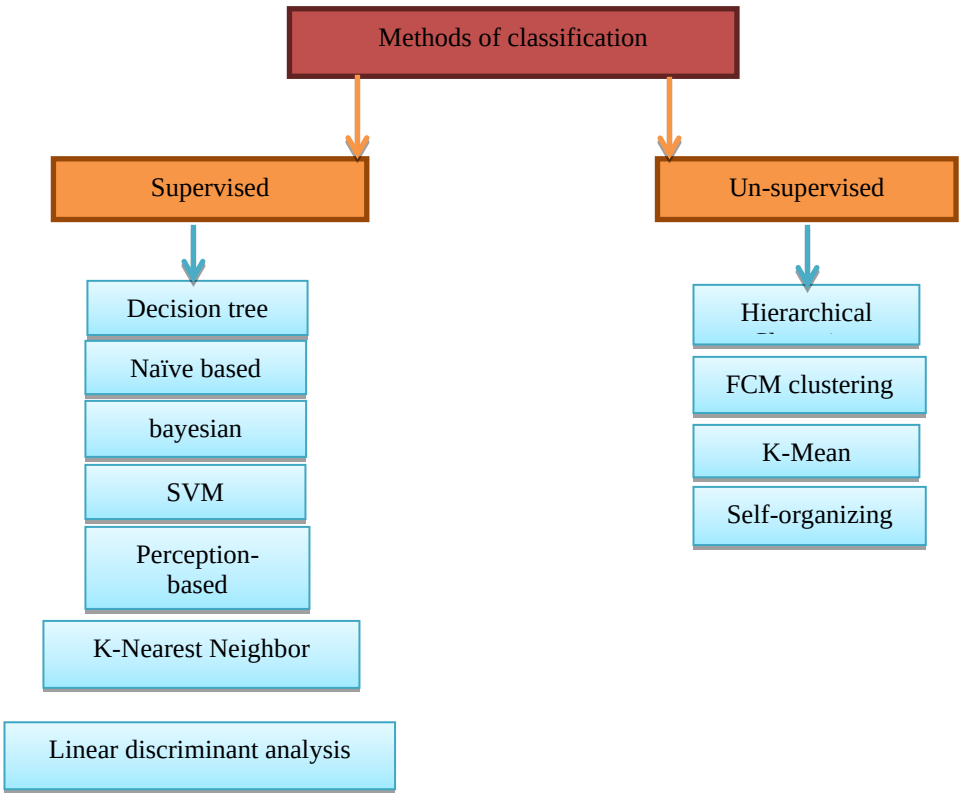


Fig.11: Methods of Image Classification

3.1 *Histogram/ Threshold-based Techniques*

This method deals with image histogram where the existence of pixels is counted. After that the histogram is divided into sub (m) parts, p1, p2, p3….and pm. An image consists numbers of pixels that are connected with each other by same characteristics. So the detection of same region and other of pixels then grouping these pixels is also done in histogram-based technique. This task is done with threshold value that is why is also known as Thresholding technique. The division may be done by colors, gray-level deviation, 2D, 3D or structural equity. Most of the image processing techniques works on color/grey-level pixels in binary by using threshold value.

Thresholding is a technique to compress and make image in a simple form as is required for segmentation. Threshold process is complicated when applied on 3D image but easy in 1D. There are any number of thresholds may be used in a histogram-based method. In some case threshold is more required. Basically there are two main reasons to adopt threshold method. Firstly, inequity amongst object and background only main objective of segmentation like in cell image segmentation. Secondly, some methods are used to handle alteration and iterate of threshold value in histogram method for image analysis that are not developed for detecting multiple thresholds

3.2 *Boundary/ edge-based techniques*

There are a number of applications are used in biomedical science to detect various objects in an image It can be done by detection of edge as each object is connected to boundary by its edge. An edge is a set of pixels that lies on boundary between two regions. Edge detection can be done by using first derivative and second derivative step. First derivative tells us where is the edge in an image and second derivative tells the direction of edge as edge going from black to white or vice versa. So both paly important role in boundary based method.

It has been noted that derivative based edge detection is extremely sensitive to noise. So here is need to apply special filter to remove extra noise. The edge detection shows the edge information and the relationship between pixels in an image. If pixel of image has gray value it means not an edge at that point. However, if a pixel has a neighbor with widely varying gray level it may be an edge point. In general edges are caused by change in color and texture in image or specific lighting condition present during image acquisition process.

3.3 *Region-based techniques*

In this method the objective of the segmentation is to partition an image into regions. We have approached this problem by finding boundaries between region based on discontinuities in gray levels. Also it is accomplished via threshold based on the distribution of pixel properties like gray level values or color. The region based method is applied into two ways (a) growing-and-merging and splitting-and-merging, as well as, more recently, region-based techniques rely on (b) morphological operations. The region growing is a procedure that groups pixels or sub regions into large region based on predefined criteria. To start with the set of "seed" points and from these grow regions by appending to each seed those neighboring pixels that have same properties as seed selected. Selecting a set of one or more starting points often can be based on the nature of the problem. The seed is selected on basis of the features of an images like gray level is high or low. This process is repeat again and again with another seed until more than one seeds cannot be detected.

3.4 *Pixel-based Techniques*

In this the image segmentation has been done in terms of making clusters of pixels in two groups that is foreground (cell) and background pixels. The pixels that belongs to same group grouped together for making one region. Final output after performing the segmentation is come in the form of fore-/back ground significant. Selecting of algorithm that works on pixels of image and analyze, whether one can use statistical or neural computing-based learning techniques here. Intensity is the fundamental point of pixel-based method. For color images the intensity is typically measured for each of three primary channels, red/blue/green [16] while for gray images only the gray-level is used [17]. Recall that cell images taken under different illuminations table 1 shows that such images cannot easily be segmented using pixel-based techniques. The large background in these images is segmented into many regions even they have a very constant intensity that is close to black.

TABLE 1. COMPARISON OF SEGMENTATION TECHNIQUES [18]

Parameters/Orignal cell image	Thresholding	Edge based segmentation	Region based Segmentation	Pixel-based segmentation
Performance	Best	Excellent	Good	Good
Color space	HIS images, combination of YIQ, values and RGB and grayscale	Gray levels and RGB images	Gray levels and RGB images, intensity and saturation	Gray levels and RGB images, intensity
Segment level	Homogeneity	Discontinuity in homogeneity	Homogeneity	Homogeneity
Segmentation effect	Good	Average	Normal	Normal
Complexity	Very low	Average	High	Average
Quality measurement	Depends upon threshold value	Depends upon intensity variation	Depend upon same pixels in region	Depends upon no of pixels in cluster

4. PERFORMANCE MEASURES

There are several ways to detect the tumor with respect to various performance measures. There are many techniques used for this purpose by evaluators and researchers that are mostly adopted in previous studies. The following are some important performance measures that will be used in this paper:

1. Confusion Matrix: It is a method for compiling the performance of the classification algorithm. The only accuracy of classification may be ambiguous if there is any an unequal
2. Mean Square Error (MSE): It estimates the unobserved quantity and calculates the average square of the error between actual and estimated results. It shows the square difference between the final result and the expected result.

$$Mean\ Square\ Error\ (MSE) = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[A(i,j) - B(i,j)]^2 \quad (I)$$

The variable m, n in the Equation (1) represents the row and column of the image. A shows the actual output and B shows the expected output calculated with the used techniques.

3. Peak Signal to Noise Ratio (PSNR): It is the ratio between the highest power of the signal and corrupted signal power that affects the accuracy of the final result.

$$Peak\ Signal\ to\ Noise\ Ration\ (PSNR) = 10\log_{10}\left(\frac{MAXi^2}{MSE}\right) \quad \text{(II)}$$

$MAXi$ is used to represent the total pixel in the MRI brain images.

4. Jaccard Index (Tanimoto Co-efficient): It is used to find the common features (similarities) amongst the finite sample dataset. Jaccard Index shows the pixels that are similar to each other pixels between ground truths (A) and the segmented result (B). A higher Jaccard index means a very accurate result is calculated.

$$\text{jaccard Index J (A, B)} = \frac{|A \cap B|}{|A \cup B|} \quad \text{(III)}$$

5. Dice Similarity Coefficient (DSC) is also known as the proportion of particular compromise by Fleiss. It shows the overlap pixels between the ground truths (A) and the segmented result (B).

$$Dice\ Similarity\ Coefficien\ (DSC\ (A, B)) = \frac{2|A \cap B|}{|A| + |B|} \quad \text{(IV)}$$

Calculation of DSC can be done by the use of Jaccard Index value as,

$$\text{DSC}\ (A, B) = 2* \frac{j(A,B)}{1+j(A,B)} \quad \text{(V)}$$

Calculation of DSC can be done by the use of Confusion Matrix value as,

$$DSC = \frac{2TP}{2TP+FP+FN} \quad \text{(VI)}$$

6. Dice-overlap- index (DOI) or Dice Similarity Coefficient (DSC): It counts only positive once in both the numerator and denominator. DSC is the quotient of the common range between 0 and 1.
7. Accuracy metric: It is used to evaluate the result of the algorithm.

$$Accuracy = \frac{TN+TP}{TP+FP+TN+FN} \quad \text{(VII)}$$

8. Specificity: It is known as a true negative rate as given by Equation (VIII):

$$Specificity = \frac{TN}{TN+FP} \quad \text{(VIII)}$$

Sensitivity: It is also known as true positive rate or Sensitivity/Recall/True Positive Rate.

$$Sensitivity = \frac{TP}{TP+FN} \quad \text{(IX)}$$

Recall and Precision. A recall is a fraction of relevant objects among the retrieved objects.

9. Precision. It represents a close estimate between two different samples. The confusion matrix helps in providing the necessary data related to the true result and the approximated results calculated by the segmentation or classification techniques as shown in Table 2.

TP: TRUE POSITIVE Accurate result (Tumor cell exist)
FP: FALSE POSITIVE Partial result (Partially tumor cell)
FN: FALSE NEGATIVE false result (No tumor)
TN: TRUE NEGATIVE false (No tumor)
Precision

$$Precision = \frac{TP}{TP+FP} \quad \text{(X)}$$

False Positive Rate

$$FPR=\frac{FP}{FP+TN} \quad \text{(XI)}$$

TABLE 2: CONFUSION MATRIX

	PREDICTED CLASS 1	**PREDICTED CLASS 2**
ACTUAL CLASS 1	TP	FN
ACTUAL CLASS 2	FP	TN

5. DATASETS

There are various datasets used for the detection and validation of brain tumors cell as mentioned in the below Table 3.

TABLE 3. DATASETS USED FOR BRAIN TUMOR CELL

S.No	**Name of dataset**	**Numberof images/ slices**	**Type of images**	**Types of tumor cell**	**Reference**
1	Local Harvard	21,100,613	MRI	Low- and High-Grade Gliomas, Metastatic Bronchogenic Carcinoma,	[28],[4],[116]
2	Rider	126	MRI	Early stage tumor detection	[4]

3	IBSR	18	MRI	extraction of the brain MRtissue (WM, GM and CSF)	[12],[30],[31]
4	Marmoset	-	MRI	CC,SGM, WM, LCC and RCC	[12]
5	BRATS 2012-2013	60,285,30	MRI	glioma HGG ,LGG (T1, T1C, T2 and FLAIR)	[32],[33],34],[35],[36],[37],[38]
6	BRATS 2015	354	MRI	classified tumor benign and malignant tumor cell	[39],[32],[33],[40],[34],[41],[42]
7	BRATS 2016	274	MRI	Flair, T1c, T2,LGG,HGG	[34]
8	BRATS 2017	210	MRI	Glioblastomas (with both high and low grade)	[43]
9	BRATS 2018	285	MRI	glioma HGG and LGG	[32]
10	Rembrandt	130	MRI	low or high grade Gliomas	[44]
11	ISLES 2015	61	MRI	Gliomas	[40]
12	Figureshare	3064	MRI	(meningioma, glioma and pituitary tumors)	[13]
13	DICOM	22	MRI	(T1-, T2-, and proton density-(PD-) weighted) and a variety of slice thicknesses, noise levels, and levels of intensity non-uniformity	[45],[46]
14	PGIMER, PSL	428	MRI	Astrocystoma (AS), Glioblastoma Multiforme (GBM), childhood tumor-Medullobla stoma (MED) and Meningioma(MEN), along with Secondary tumor-Metastatic (MET).	[47]
15	MRbrains	48	MRI	brain Cell like gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF)	[1]
16	ABIDE	-	MRI	T2	[48]
17	BraTS2019	336	MRI	glioma patients (259 HGG and 76 LGG).	[49] [50]

TABLE 4. OVERVIEW OF VARIOUS METHODS USED FOR CELL IMAGE SEGMENTATION ON THE BASIS OF THEIR PERFORMANCE.

Ref No	Method Used	Performance measure (Accuracy)	False-positive Rate	True Positive Rate Sensitivity/ Recall	Database Used
[1]	Hybrid Fuzzy C-Mean + SVM	91.88% (Linear)		0.9000 (Linear)	
[4]	SVM, Gaussian radial based function	97.0%	Minimum 0.0 Maximum 0.40 in 5 folds) Minimum 0.0 Maximum 0.34 in 30 folds)	Minimum 0.904 Maximum 0.969 In 5 folds Minimum 0.809 Maximum 0.952 In 30 folds	Local, Harvard Rider
[5]	SWT+GCNN	98.9%	0.019	0.9834	BRAINIX medical images
[8]	CNN	86.70% Gray matter 89.88%, white matter, and 85% and cerebrospinal fluid, (Dice Coefficient)			MRBrainS Challenge database
[13]	SVM+K- Nearest Neighbor	92.5% deep transfer Learning 98.5% (SVM+DCNN) 98.3 (KNN+DCNN)		1 meningioma 0.990 glioma, 0.995 Pituitary	Figureshare brain tumor dataset
[14]	DNN	93.8% (Mean Dice Coefficient)			(IBSR) Marmoset brain image dataset
[15]	K-mean+CNN	96%	0.02	0.99	-
[16]	Pre-trained Convolutional Neural Network ResNet34	99.7%	0.02 (5 folds)	1.06 (5 folds)	Rider
[17]	KFCOM + WKFCOM	93.5%			
[19]	CNN	92.9%		0.94	BRATS 2013, BRATS 2015
[22]	BWT +SVM	97%	0.022	0.6441	BrainWeb, Digital Imaging and Communications in Medicine (DICOM)
[23]	DCNN +CRF	90.3% (Dice Coefficient)		0.989	BRATS 2015, ISLES 2015
[29]	FCM + SVM	More than 99%			
[30]	FFT + minimum redundancy MRMR)+ SVM	99%			

[31]	KIFCM (hybrid K-mean + Fuzzy C-Mean) +Active Contour (Level Set Contouring)	100%	0.03 (DS1), 0.00 (DS2), 0.01 (DS3)	90.5(DS1), 100 (DS2),100 (DS3)	BrainWeb, Digital Imaging and Communications in Medicine (DICOM)
[32]	Wiener filter + (PCA) + Radial Basis Function - kernel based SVM	94.7%	0.05	0.80	
[33]	Convolutional Neural Network (CNN)	89.2% (Dice Coefficient)		0.8832	BRATS 2013, BRATS 2015
[34]	Cuckoo Search Algorithm Tsallis entropy + regulized data set	More than 99%	Min 0.0905 (Slice 60) Max 0.4715 (T2)	0.9810(Slice 60) 0.9311 (T2)	
[35]	Particle Swarm Optimization (PSO) + Bacteria Foraging Optimization (BFO) + Modified Fuzzy C- Mean	97.22% Index)		0.9545	Harvard Brain Web, BRATS 2013
[36]	Kernelized fuzzy entropy clustering +PSO	91%	(Jaccard Index)		Internet Brain Segmentation Repository (IBSR)
[39]	Tsallis entropy + Bat Algorithm	94.78%	0.0495	0.9649	BRAINIX medical images
[40]	NN	82.8	0.1665	1.02	
[41]	PSO +SVM	96.23%		1.09	BRATS 2015
[42]	PSO + FCM	93.71%			BRATS 2013
[43]	TLBO, Shannon' s entropy, the level set method	95.07%	0.0399	0.9815	CEREBRIX, BRAINIX, BRATS 2012
[44]	PSO + SVM + Cuckoo Search	99.69%	0.00	0.967	
[45]	BFO MFKM	96%		0.9714	Brain Web
[46]	Kapur’s Entropy-based Cuckoo Search Optimization	97%		0.9765	Harvard Brain Web, BRATS 2013
[47]	PSO + FCM	96..39% White Matter, 96.61% Jaccard Index)			Internet Brain Segmentation Repository (IBSR
[48]	MFCM + Optimized Ant Colony	97.93%			REMBRANDT dataset (TCIA)
[49]	CNN GA	94.2%	Min 0.004 Max 0.044	Max 0.998 Min 0.868	REMBRANDT dataset (TCIA)
[50]	ANN + GA	> 90%			PGIMER dataset, SPL database
[51]	Firefly Algorithm (FA) + Tolerance Rough Set (TRS)	> 90%		Max 0.924 Min 0.76	REMBRANDT dataset (TCIA)
[52]	DCNN +GA	69.25%		0.6905	

6. CONCLUSION

In this article, the most popular methods for detecting brain tumor segmentation and classification are reviewed by using in fMRI pictures. The actual objective of this paper is to describe an overall view of the most used techniques for brain image segmentation and classification. The final result of analyzing shows the various techniques used in the field of machine learning like (SVM and SOM), some methods are based on deep learning like (CNN, DCNN, G-CNN), meta-heuristic algorithms (GA, DE, PSO, Bat algorithm, ABC), data mining tools (FCM) and habituation methods have been used for the segmentation and classification of brain abnormality. It is observed that KIFCM has 100% accuracy, PSO+SVM+ cuckoo search in combo has the highest accuracy 99.69%, SVM+KNN 98.5%, and also observe that BRATS data set most of the time used in the time period of 2013 to 2015. The overall performance of all the methods is given in table 4. These methods can be used for segmenting and classification of some critical issues of the brain like Parkinsonism's issues, Presenile dementia's issues, stroke, and autism. This article also discusses the availability of database sets that are used by the researcher for the verification of final output. In order to enhance the accuracy of the system, some other combinations of different classifiers can also be used. Brats 2013 and brats 2015 are the basic datasets that are mostly being used by the researchers.

7. REFERENCES

[1]. Amin J, Sharif M, Yasmin M, Fernandes SL. A distinctive approach in brain tumor detection and classification using MRI. Pattern Recognition Letters. (2017) pp. 1-10.

[2]. S. Deepak, P.M. Ameer, Brain tumor classification using deep CNN features via transfer learning, Computers in Biology and Medicine, 111, (2019).

[3]. Ryo Ito, Ken Nakae, Junichi Hata, Hideyuki Okano, Shin Ishii, Semi-supervised deep learning of brain tissue segmentation, Neural Networks, 116, (2019),pp. 25-34.

[4]. Mamta Mittal, Lalit Mohan Goyal, Sumit Kaur, Iqbaldeep Kaur, Amit Verma, D. Jude Hemanth, Deep learning-based enhanced tumor segmentation approach for MR brain images, Applied Soft Computing, 78, (2019), pp. 346-354.

[5]. S. T. Kebir and S. Mekaoui, An Efficient Methodology of Brain Abnormalities Detection using CNN Deep Learning Network, International Conference on Applied Smart Systems (ICASS*)*, Medea, Algeria, (2018), pp. 1-5.

[6]. Muhammed Talo, Ulas Baran Baloglu, Özal Yıldırım, U Rajendra Acharya, Application of deep transfer learning for automated brain abnormality classification using MR images, Cognitive Systems Research, 54, (2019), pp. 176-188

[7]. Tianbao Ren, Huanhuan Wang, Huilin Feng, Chensheng Xu, Guoshun Liu, Pan Ding, Study on the improved fuzzy clustering algorithm and its application in brain image segmentation, Applied Soft Computing, 81, (2019).

[8]. S. N. Sulaiman, N. A. Non, I. S. Isa and N. Hamzah, Segmentation of brain MRI image based on clustering algorithm, IEEE Symposium on Industrial Electronics & Applications (ISIEA), Kota Kinabalu, (2014), pp. 60-65.

[9]. Jinghong Li, Zhu Liang Yu, Zhenghui Gu, Hui Liu, Yuanqing Li, MMAN: Multi-modality aggregation network for brain segmentation from MR images, Neurocomputing, 358, (2019), pp. 10-19.

[10]. Parveen and A. Singh, Detection of a brain tumor in MRI images, using a combination of fuzzy c-means and SVM, International Conference on Signal Processing and Integrated Networks (SPIN), (2015), pp. 98-102.

[11]. Omid Tarkhaneh, Haifeng Shen, An adaptive differential evolution algorithm to optimal multi-level thresholding for MRI brain image segmentation, Expert Systems with Applications, 138, (2019).

[12]. Thaha, M. Mohammed, Kumar, K. Pradeep Mohan, Murugan, B. S., Dhanasekeran, S., Vijayakarthick, P., Selvi, A. Senthil, Brain Tumor Segmentation Using Convolutional Neural Networks in MRI Images, Journal of Medical Systems, 43 (9), (2019), pp. 294

[13]. Kai Hu, Qinghai Gan, Yuan Zhang, Shuhua Deng, Fen Xiao, Wei Huang, Chunhong Cao, and Xieping Gao, Brain Tumor Segmentation Using Multi- Cascaded Convolutional Neural Networks and Conditional Random Field, IEEE Access,7, (2019), pp. 92615-92629.

[14]. R. A. Rajam, R. Reshmi, A. Suresh, A. Suresh, and S. Sindhuja, Segmentation and Analysis of Brain Tumor using Meta-Heuristic Algorithm, International Conference on Recent Trends in Electrical, Control, and Communication (RTECC), Malaysia, Malaysia, (2018), pp. 256-260.

[15]. M. I. Razzak, M. Imran, and G. Xu, Efficient Brain Tumor Segmentation with Multiscale Two-Pathway-Group Conventional Neural Networks, IEEE Journal of Biomedical and Health Informatics, 23 (5), (2019), pp.1911-1919.

[16]. Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, Daniel Rueckert, Ben Glocker, Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation, Medical Image Analysis, 36, (2017), pp. 61-78.

[17]. Rajinikanth V., Fernandes Steven Lawrence, Bhushan Bharath, Harisha and Sunder Nayak Ramesh, Segmentation and analysis of Brain Tumor using Tsallis Entropy and Regularized Level Set, International Conference on Micro-Electronics, Electromagnetics and Telecommunications, Springer Singapore, (2018), pp. 313-321.

[18]. Damodharan, S., Raghavan, D., Combining tissue segmentation and neural network for brain tumor detection, International Arab Journal of Information Technology, 12 (1), (2015), pp. 42-52.

[19]. Quratul Ain, M. Arfan Jaffar, Tae-Sun Choi, Fuzzy anisotropic diffusion-based segmentation and texture-based ensemble classification of brain tumor, Applied Soft Computing, 21, (2014), pp. 330-340.
[20]. Alfonse, Marco, and Abdel-Badeeh M. Salem, an automatic classification of brain tumors through MRI using a support vector machine. Egyptian Computer Science Journal 40 (03), (2016).
[21]. Eman Abdel-Maksoud, Mohammed Elmogy, Rashid Al-Awadi, Brain tumor segmentation based on a hybrid clustering technique, Egyptian Informatics Journal,16 (1), (2015), pp. 71-81.
[22]. Kumar, P., and B. Vijayakumar, Brain tumor MR image segmentation and classification using PCA and RBF kernel-based support vector machine.Middle-East Journal of Scientific Research, 23 (9) (2015), pp. 2106-2116.
[23]. Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, Yong Fan, A deep learning model integrating FCNNs and CRFs for brain tumor segmentation, Medical Image Analysis, 43, (2018), pp. 98-111.
[24]. Wenlu Zhang, Rongjian Li, Houtao Deng, Li Wang, Weili Lin, Shuiwang Ji, Dinggang Shen, Deep convolutional neural networks for multi-modality isointense infant brain image segmentation, NeuroImage,108, (2015), pp. 214-224.
[25]. Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, Hugo Larochelle, Brain tumor segmentation with Deep Neural Networks, Medical Image Analysis, 35, (2017), pp. 18-31.
[26]. Nilesh Bhaskarrao Bahadure, Arun Kumar Ray, and Har Pal Thethi, Image Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM, International Journal of Biomedical Imaging, (2017), pp. 1-12.
[27]. Agus Pratondo, Chee-Kong Chui, Sim-Heng Ong, integrating machine learning with region-based active contour models in medical image segmentation, Journal of Visual Communication and Image Representation, 43, (2017), pp. 1-9.
[28]. Mostefa Ben naceur, Rachida Saouli, Mohamed Akil, Rostom Kachouri, Fully Automatic Brain Tumor Segmentation using End-To-End Incremental Deep Neural Networks in MRI images, Computer Methods and Programs in Biomedicine,166, (2018), pp. 39-49.
[29]. Vasupradha Vijay, A.R. Kavitha, S. Roselene Rebecca, Automated Brain Tumor Segmentation and Detection in MRI Using Enhanced Darwinian Particle Swarm Optimization (EDPSO), Procedia Computer Science, 92, (2016), pp. 475-480.
[30]. Kamanasish Bhattacharjee, Millie Pant, Hybrid particle swarm optimization-genetic algorithm trained multi-layer perceptron for classification of human glioma from molecular brain neoplasia data, Cognitive Systems Research, 58, (2019), pp. 173-194.
[31]. Wang, S, Zhang, Y, Dong, Z. Du., S., Ji. G., Yan, J., Yang, J., Wang, Q., Feng, C. and Phillips, Feed- forward neural network optimized by hybridization of PSO and ABC for abnormal brain detection . Int. J. Imaging Syst. Technol., 25, (2015), pp. 153-164.
[32]. Thuy Xuan Pham, Patrick Siarry, Hamouche Oulhadj, integrating fuzzy entropy clustering with an improved PSO for MRI brain image segmentation, Applied Soft Computing, 65, (2018), pp. 230-242.
[33]. A. R. Deepa and W. R. Sam emmanuel, MRI Brain Tumor Classification Using Cuckoo Search Support Vector Machines and Particle Swarm Optimization Based Feature Selection, 2nd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, (2018), pp. 12131216.
[34]. N. Kumari and S. Saxena, Review of Brain Tumor Segmentation and Classification,International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, (2018), pp. 1-6.
[35]. A. Kumar, A. Ashok and M. A. Ansari, Brain Tumor Classification Using Hybrid Model of PSO and SVM Classifier, International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida (UP), India, (2018), pp. 1022-1026.
[36]. H. A. M. Ali, M. A. A. Ahmed and E. M. Hussein, MRI Brain Tumour Segmentation Based on Multimodal Clustering and Level-set Method, International Conference on Computer, Control, Electrical, and Electronics Engineering (ICCCEEE), Khartoum, (2018), pp. 1-5.
[37]. M. Ü. Özıç, Y. Özbay and Ö. K. Baykan, Detection of tumor with Otsu-PSO method on brain MR image, 22nd Signal Processing and Communications Applications Conference (SIU), Trabzon, (2014), pp. 1999-2002.
[38]. R. Kaur and G. Singh, Hybrid Technique Using PSO and Region Growing Algorithm for Brain Tumor Detection, Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, (2018), pp. 1286-1289.
[39]. S. A. Taie and W. Ghonaim, Title CSO-based algorithm with support vector machine for brain tumor's disease diagnosis,,IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), Kona, HI, (2017), pp. 183-187.
[40]. Mohammad Majid al-Rifaie Ahmed Aber ; Duraiswamy Jude Hemanth, Deploying swarm intelligence in medical imaging identifying metastasis, microcalcifications and brain image segmentation, IET Systems Biology, 9 (6) ,(2015), pp. 234 – 244.
[41]. Rajinikanth, V., Satapathy, S. C., Fernandes, S. L., & Nachiappan, S. Entropy based segmentation of tumor from brain MR images–a study with teaching learning-based optimization. Pattern Recognition Letters, 94, (2017), pp. 87-95.
[42]. Asmita Dixit ,Aparajita Nanda, Brain MR image Classification via PSO based Segmentation, Twelfth International conference on Contemporary Computing(IC3), 8-10 Aug. 2019, Noida, India.

[43]. Fernandes, S. L., Tanik, U. J., Rajinikanth, V., & Karthik, K. A. A reliable framework for accurate brain image examination and treatment planning based on early diagnosis support for clinicians. Neural Computing and Applications, (2019), pp. 1-12

[44]. Ahmed Elazab, Ahmed M. Anter, Hongmin Bai, Qingmao Hu, Zakir Hussain, Dong Ni, Tianfu Wang, Baiying Lei, An optimized generic cerebral tumor growth modeling framework by coupling biomechanical and diffusive models with treatment effects, Applied Soft Computing, 80, (2019), pp. 617-627.

[45]. Anitha Narayanan, M. Pallikonda Rajasekaran, Yudong Zhang, Vishnuvarthanan Govindaraj, Arunprasath Thiyagarajan, Multi-channeled MR brain image segmentation: A novel double optimization approach combined with clustering technique for tumor identification and tissue segmentation, Biocybernetics and Biomedical Engineering, 39 (2), (2019), pp.350-381.

[46]. AnithaVishnuvarthanan, M. Pallikonda Rajasekaran, Vishnuvarthanan Govindaraj, Yudong Zhang, Arunprasath Thiyagarajan, An automated hybrid approach using clustering and nature inspired optimization technique for improved tumor and tissue segmentation in magnetic resonance brain images, Applied Soft Computing, 57, (2017), pp. 399-426.

[47]. R.Sumathi, M.Venkatesulu, Sridhar P.Arjunan, Extracting tumor in MR brain and breast image with Kapur's entropy based Cuckoo Search Optimization and morphological reconstruction filters Biocybernetics and Biomedical Engineering, 38 (4), (2018), pp. 918-930.

[48]. Abdenour Mekhmoukh, Karim Mokrani, Improved Fuzzy C-Means based Particle Swarm Optimization (PSO) initialization and outlier rejection with level set methods for MR brain image segmentation, Computer Methods and Programs in Biomedicine, 122, (2), (2015), pp. 266-281.

[49]. G. S. Raghtate and S. S. Salankar, Modified Fuzzy C Means with Optimized Ant Colony Algorithm for Image Segmentation, International Conference on Computational Intelligence and Communication Networks (CICN), Jabalpur, (2015), pp. 1283-1288.

[50]. Amin Kabir Anaraki, Moosa Ayati, Foad Kazemi, Magnetic resonance imaging-based brain tumor grades classification and grading via convolutional neural networks and genetic algorithms, Biocybernetics and Biomedical Engineering, 39, Issue 1, 2019, Pages 63-74.

[51]. G. Rajesh Chandra, Kolasani Ramchand H. Rao, Tumor Detection In Brain Using Genetic Algorithm, Procedia Computer Science, 79, 2016, pp. 449-457.

[52]. Amiya Halder, Anuva Pradhan, Sourjya Kumar Dutta ; Pritam Bhattacharya, Tumor extraction from MRI images using dynamic genetic algorithm based image segmentation and morphological operation, International Conference on Communication and Signal Processing (ICCSP), (2016).

[53]. V. Kiruthika Lakshami, C. A. Feroz and J. Asha Jenia Merlin, Automated Detection and Segmentation of Brain Tumor Using Genetic Algorithm, International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, (2018), pp. 583-589.

[54]. T. Chithambaram and K. Perumal, Brain tumor segmentation using genetic algorithm and ANN techniques, IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, (2017), pp. 970-982.

[55]. N. Menon and R. Ramakrishnan, Brain Tumor Segmentation in MRI images using unsupervised Artificial Bee Colony algorithm and FCM clustering, International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, (2015), pp. 0006-0009.

[56]. Kamalam Balasubramani P Madhura ; Ramya V. Kulkarni ; P Pavithra, Hybridized approach of artificial bee colony algorithm for detection of suspicious brain pattern using magnetic resonance images, 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), (2017),pp.21-22.

[57]. Saravanan Alagarsamy, Kartheeban Kamatchi, Vishnuvarthanan Govindaraj, Yu-Dong Zhang, Arunprasath Thiyagarajan, Multi-channeled MR brain image segmentation: A new automated approach combining BAT and clustering technique for better identification of heterogeneous tumors, Biocybernetics and Biomedical Engineering, 2019.

[58]. Taranjit Kaur, Barjinder Singh Saini, Savita Gupta, An optimal spectroscopic feature fusion strategy for MR brain tumor classification using Fisher Criteria and Parameter-Free BAT optimization algorithm, Biocybernetics and Biomedical Engineering, 38 (2) (2018), pp. 409-424.

[59]. Jothi G., HannahInbarani H., Hybrid Tolerance Rough Set–Firefly based supervised feature selection for MRI brain tumor image classification, Applied Soft Computing,Volume 46, (2016), pp. 639-651.

[60]. J. Kennedy and R. Eberhart, Particle swarm optimization, Proceedings of ICNN'95 International Conference on Neural Networks, Perth, WA, Australia, 4, (1995), pp. 1942-1948.

[61]. J. Kennedy and R. Eberhart, Particle swarm optimization, Proceedings of ICNN'95 International Conference on Neural Networks, Perth, WA, Australia, 4, (1995), pp. 1942-1948.

[62]. Acharya, U. Rajendra, Steven Lawrence Fernandes, Joel En WeiKoh, Edward J. Ciaccio, Mohd Kamil Mohd Fabell, U. John Tanik, V. Rajinikanth, and Chai Hong Yeong. Automated Detection of Alzheimer's Disease Using Brain MRI Images–A Study with Various Feature Extraction Techniques. Journal of Medical Systems 43 (9), (2019), 302.

[63]. S. Clare, 1997, Functional MRI: Methods and Applications, Ph.D. Thesis, University of Nottingham, Nottingham, United Kingdom.

[64]. Kuntegowdenahalli LC, Jacob LA, Komaranchath AS, Amirtham U. A rare case of primary anaplastic large cell lymphoma of the central nervous system. J Can Res Ther (2015), 11 (4), pp. 943-945.

[65]. Chang C-W, Ho C-C and Chen J-H, ADHD classification by a texture analysis of anatomical brain MRI data. Front. Syst. Neurosci, (2012), 6, pp. 66.

[66] Tiwari, Arti, Shilpa Srivastava, and Millie Pant. "Brain tumor segmentation and classification from magnetic resonance images: Review of selected methods from 2014 to 2019." *Pattern Recognition Letters* 131 (2020): 244-260.

[67]. Nicola Amoroso, Marianna La Rocca, Alfonso Monaco, Roberto Bellotti, Sabina Tangaro, Complex networks reveal early MRI markers of Parkinson's disease, Medical Image Analysis, Volume 48, (2018), pp. 12-24.

[68]. Marie Bruun, Juha Koikkalainen, Hanneke F.M. Rhodius-Meester, Marta Baroni, Le Gjerum, Mark van Gils, Hilkka Soininen, Anne M. Remes, Päivi Hartikainen, Gunhild Waldemar, Patrizia Mecocci, Frederik Barkhof, Yolande Pijnenburg, Wiesje M. van der Flier, Steen G. Hasselbalch, Jyrki Lötjönen, Kristian S. Frederiksen, Detecting frontotemporal dementia syndromes using MRI biomarkers, NeuroImage: Clinical, Volume 22, (2019), pp. 101711,

[69]. Fang, Shengyu, et al. "Awake craniotomy for gliomas involving motor-related areas: classification and function recovery." Journal of neuro-oncology 148.2 (2020): 317-325.

[70]. Deng, H., Deng, W., Sun, X. et al. Adaptive Intuitionistic Fuzzy Enhancement of Brain Tumor MR Images. Sci Rep 6, 35760 (2016)

[71] Zeineldin, Ramy A., et al. "DeepSeg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance FLAIR images." International journal of computer assisted radiology and surgery 15.6 (2020): 909-920.

[72] Rehman, Mobeen Ur, et al. "BrainSeg-Net: Brain Tumor MR Image Segmentation via Enhanced Encoder–Decoder Network." Diagnostics 11.2 (2021): 169.

ATTRIBUTES FOR DATA QUALITY IN DATA PLATFORMS: MONITOR DATA HEALTH

Neha Sharma[1] and Er Gurpreet Singh[2]
[1.] Research Scholar, ernehavatsyan@gmail.com,
Department of computer science and Engineering, Punjabi University, Patiala
[2.] Assistant Professor , Gurpreet.1887@gmail.com ,
Department of computer science and Engineering, Punjabi University, Patiala

ABSTRACT:—— With the ever increase of data in these past years and the increasing awareness around data as an asset, various organisations/developers have been pro-active in building large scale, scalable data pipelines with huge infrastructures to process, store and analyse big data. Data is being used to make Machine learning models smart, it is being used to drive analytics giving never like before statistics around data. Data is being used to automate operations using Artificial Intelligence enabling our machines to be smart. With advances in technology, we have seen how data is enabling Intelligent Machines. For example, AI is being used to diagnose patients, Neural Networks have had a great back behind Self-Driving Cars, Robots etc. These advancements have only been possible due to data, though these inventions are helping to make our lives very easy but these are very critical. If in any chance this data gets corrupted or is not a quality data, we cannot even imagine how fatal these inventions can be to someone. Consider you have set your car on self-drive mode but it receives wrong data on a car moving ahead of you and you would suddenly bump into it. As much as large volumes of data is required for smooth functioning of software's, governing that data is equally important so that standards are up to the mark and are met for high quality results. This article is a contribution to creation of survey set for various attributes for checking data quality that play key role with real large scale data platform pipelines.

KEYWORDS: Introduction, Survey Methodology, Existing Work Survey, Conclusions, References

1. INTRODUCTION

All the digital machines we use in day-to-day life are designed to take some input and produce output, which directly depicts the importance of data in our day-to-day life. In order to build any intelligent system, one has to create a data management plan, but what are the parameters that will decide if the data at rest and motion is of quality and will create value. Usually organisations building software products, rely on expertise of their Data Engineers for Data Quality, but as we already understand the cruciality of data it is important that the validation becomes process/framework driven rather than person driven. Let's first understand what Data is and what are the various dimensions of Data.

Data:

As defined by Merriam Webster Dictionary [1] Data is

a. Facts or information used to usually calculate, analyse or plan something.
b. Information that is produced or stored by computer

According to Herencia, there are 5 Dimensions of Data when it is Big, which are popularly known as 5V's of Big Data and are namely Volume, Velocity, Veracity, Variety, Value.

Volume is a huge amount of data. Velocity stands for refers to the high speed of accumulation of data. Variety refers to nature of data that is structured, semi-structured and unstructured data. Veracity refers to inconsistencies and uncertainty in data. After having the 4 V's into account there comes one more V which stands for Value. The bulk of Data having no Value is of no good to the company, unless you turn it into something useful.

Data Platform :

A Data Platform is used for enriching, transforming and loading data. It is a central platform that stores and is used to manage data. Data Platform may consist of 100's of ETL Data Pipelines, Databases, Data Warehouses etc.

2. Survey Methodology:

We followed the systematic literature review for finding out various Data Quality Frameworks and Data Quality Metrics being used till date for measuring Data Quality. In this section we describe step by step the way to select and filter papers, analyse the research proposals and contributions in the papers. As well as synthesized the results. We defined that our work should cover all the aspects of data.

Firstly, older papers were only selected whenever they were important for understanding definitions as were looking for the most recent and most up to date techniques the authors were using.

Finally, preferred papers having deep description of techniques, experiments and concepts. Having selected our study set, we analysed those papers by first reading the abstract, introduction, conclusion to separate different and most interesting ones. Having those most interesting ones, we proceed to second deeper reading those, in order to review their techniques, definition, related work done and results. During our preliminaries reading theme related papers, we found that many authors used social aspects of the entries in order to better classify them, those aspects varied from comments, entry sharing, relationships between consumers of those entries to the writer's profile information and pictures. We

classified those aspects as relevant also, hoping to find new and better practices on how to use that external contextual information in favour of better predictions.

3. Existing Work Survey

In the 1950s, researchers began to check quality issues, especially for the standard of products, and a series of definitions, as an example, quality is "the degree to which a collection of inherent characteristics fulfill the requirements" (General Administration of Quality Supervision, 2008); "fitness for use" (Wang & Strong, 1996); "conformance to requirements" (Crosby, 1988) were published.

Later, with the rapid development of data technology, research turned to the study of the info quality. Research on data quality started abroad within the 1990s, and lots of scholars proposed different definitions of knowledge quality and division methods of quality dimensions.

The Entire Data Quality Management group of MIT University led by Professor Richard Y. Wang has done in-depth research within the data quality area. They defined "data quality" as "fitness for use" (Wang & Strong, 1996) and proposed that data quality judgment depends on data consumers.

At the identical time, they defined a "data quality dimension" as a collection of knowledge quality attributes that represent one aspect or construct of knowledge quality. They used a two-stage survey to spot four categories containing fifteen data quality dimensions.

Some literature regarded web data as research objects and proposed individual data quality standards and quality measures.

Alexander and Tate (1999) described six evaluation criteria - authority, accuracy, objectivity, currency, coverage/intended audience, and interaction/transaction features for web data. Katerattanakul and Siau (1999) developed four categories for the knowledge quality of a personal website and a questionnaire to check the importance of every of those newly developed information quality categories and the way web users determine the data quality of individual sites.

For information retrieval, Gauch (2000) proposed six quality metrics, including currency, availability, information-to-noise ratio, authority, popularity, and cohesiveness, to research.

From the attitude of society and culture, Shanks and Corbitt (1999) studied data quality and founded an emiotic-based framework for data quality with 4 levels and a complete of 11 quality dimensions.

Knight and Burn (2005) summarized the foremost common dimensions and therefore the frequency with which they're included within the different data quality/information quality frameworks. Then they presented the IQIP (Identify, Quantify, Implement, and Perfect) model as an approach to managing the selection and implementation of quality related algorithms of a web crawling computer programme.

According to the U.S. National Institute of Statistical Sciences (NISS) (2001), the principles of knowledge quality are: 1. data are a product, with customers, to whom they need both cost and value; 2. as a product, data have quality, resulting from the method by which data are generated; 3. data quality depends on multiple factors, including (at least) the aim that the info are used, the user, the time, etc.

Research in China on data quality began later than research abroad. The 63rd Research Institute of the PLA staff Headquarters created an information quality research group in 2008. They discussed basic problems with data quality like definition, error sources, improving approaches, etc. (Cao, Diao, Wang, et al., 2010).

In 2011, Xi'an Jiaotong University founded a pursuit group of knowledge quality that analyzed the challenges and importance of assuring the standard of massive data and response measures within the aspects of process, technology, and management (Zong & Wu, 2013).

The pc Network Information Center of the Chinese Academy of Sciences proposed a knowledge quality assessment method and index system (Data Application Environment Construction and repair of the Chinese Academy of Sciences, 2009) within which data quality is split into three categories including external form quality, content quality, and also the utility of quality. Each category is subdivided into quality characteristics and an evaluation index.

In summary, the prevailing studies specialize in two aspects: a series of studies of web data quality and studies in specific areas, like biology, medicine, geophysics, telecommunications, scientific data, etc. Big data as an emerging technology, acquires more and more attention but also lacks research leads to establishing big data quality and assessment methods under multi-source, multi-modal environments (Song & Qin, 2007)

4. IDENTIFYING DIFFERENT ATTRIBUTES FOR DATA QUALITY

Following list includes with many authors along the year. What Attributes were defined for measuring Data Quality and whether they were user based, architecture based, organization based etc.

Authors [Year]	Attributes/Categories/Determinants
Wang & Strong, 1996	Dimensions of Data
U.S. National Institute of Statistical Sciences (NISS) (2001)	Customers, Data Source , time

Gauch (2000)	Currency, Availability, Information-to-noise ratio, Authority, Popularity, and Cohesiveness, To Investigate
Data Application Environment Construction and Service of the Chinese Academy of Sciences, 2009	Form quality, Content quality, Utility of quality
Alexander and Tate (1999)	Authority, Accuracy, Objectivity, Currency, coverage/intended audience, and interaction/transaction
Knight and Burn (2005)	Identify, Quantify, Implement, and Perfect
Klara Nelson (2002)	Industry Environment, Organisational Environment, It Environment
(IJACSA) International Journal of Advanced Computer Science and Applications (2021)	Organizational, Managerial, Stakeholder, Technological and External

CONCLUSION

Many researches have been done to define what is Data Quality and define various attributes to check Data Quality and to make a data quality and assessment and still the researches are going on. Since the world is moving towards Data Driven and Microservice Architecture, having a generic Quality assessment framework that can be used with and architecture is the need. Many researches have been done many are still in progress. If the world has to move towards Intelligent Machines Architecture and wants to rely on them, the need is for Standard Data Quality Framework.

REFERENCES

1) https://www.merriam-webster.com/dictionary/data
2) Alan, F. K., Sanil, A. P., Sacks, J., et al. (2001) Workshop Report: Affiliates Workshop on Data Quality, North Carolina: NISS.
3) Alexander, J. E., & Tate, M. A. *Web wisdom: How to evaluate and create information on the web*, Mahwah, NJ: Erlbaum.
4) Cao, J. J., Diao, X. C., Wang, T., et al. (2010) Research on Some Basic Problems in Data Quality Control. *Microcomputer Information 09*, pp 12–14.
5) Cappiello, C., Francalanci, C., & Pernici, B. (2004) Data quality assessment from user‘s perspective. *Procedures of the 2004 International Workshop on Information Quality in Information Systems*, New York: ACM, pp 78–73.
6) Crosby, P. B. (1988) *Quality is Free: The Art of Making Quality Certain*, New York: McGraw-Hill.
7) Data Application Environment Construction and Service of Chinese Academy of Sciences (2009) Data Quality Evaluation Method and Index System. Retrieved October 30, 2013 from the World Wide Web: http://www.csdb.cn/upload/101205/1012052021536150.pdf
8) Demchenko, Y., Grosso, P., de Laat, C., et al. (2013) Addressing Big Data Issues in Scientific Data Infrastructure. *Procedures of the 2013 International Conference on Collaboration Technologies and Systems*, California: ACM, pp 48–55.
9) Feng, Z. Y., Guo, X. H., Zeng, D. J., et al. (2013) On the research frontiers of business management in the context of Big Data. *Journal of Management Sciences in China 16*(01), pp 1–9.
10) Gantz, J., & Reinsel, D. (2012) THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. Retrieved February, 2013 from the World Wide Web: http://www.emc.com/collateral/analyst-reports/idc-digital-universe-western-europe.pdf
11) General Administration of Quality Supervision (2008) Inspection and Quarantine of the People’s Republic of China. *Quality management systems-Fundamentals and vocabulary* (GB/T19000—2008/ISO9000:2005), Beijing.
12) Katal, A., Wazid, M., & Goudar, R. (2013) Big Data: Issues, Challenges, Tools and Good Practices. *Procedures of the 2013 Sixth International Conference on Contemporary Computing*, Noida: IEEE, pp 404–409.
13) https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1079&context=acis2007
14) https://core.ac.uk/download/pdf/301346399.pdf
15) https://thesai.org/Downloads/Volume12No2/Paper_24-Factors_Influencing_Master_Data_Quality.pdf

USE OF DATA ANALYSIS AND ARTIFICIAL INTELLIGENCE TO IMPROVE MANUFACTURING PERFORMANCE: A REVIEW PAPER

Abrar Ali Khan, Amisha Tiwari, Jashanpreet Singh Toor, Santbir Singh

ABSTRACT: Data-driven error detection and prediction using artificial neural networks during the production process has become an essential tool to improve the manufacturing process. Various fault detection and Isolation (FDI) techniques have been implemented to improve an engineering system's reliability, accuracy, and safety. Time domain, frequency domain and time-frequency domain analysis have been performed to identify characteristics or patterns linked to fault conditions. Early problem detection provides crucial forewarning time, allowing for the implementation of necessary countermeasures to avert unwanted failures or poor product quality in industrial operations. This paper reviews various data analysis and Artificial Intelligence methods used to improve production performance in different manufacturing sectors.

1. INTRODUCTION

The possibility of changing business models, implementing new operating methods to support such models, and monetizing data to achieve new levels of productivity has made data analysis and AI top technological priority for manufacturing firms. The fourth Industrial revolution is defined by a convergence of technology that blurs the distinctions between the physical, digital, and biological realms. Artificial Intelligence and Data analysis is playing a pivotal role in achieving this target [1]. The industrial internet and smart factories are revolutionizing industries with intelligent manufacturing helping in boost productivity and quality of manufacturing. The augmented intelligence, which combines human judgement with AI powered data analytics, provides smart algorithms for fast, data driven decisions. The use of early detection or fault prediction inside production lines has recently been emphasised as a way to improve manufacturing processes [2]. AI and data analytics technologies have been used for predictive maintenance and unusual event prediction in production lines [3]. Smart manufacturing is a new type of production that incorporate sensors, computer platforms, communication technologies, control, simulation, data intensive modelling, and predictive engineering into manufacturing processes [4]. Data collected during a smart manufacturing process can be used to track down material origins, simplify equipment management, boost transaction efficiency, and build a flexible pricing structure [5]. AI is a widespread technology with applications span ranging from manufacturing and media to education and healthcare. According to their developmental history, AI is of four types- mechanical, analytical, intuitive and empathetic with level of AI increasing from mechanical to empathetic [6].

2. Industry 4.0: The fourth industrial revolution is known as industry 4.0. In recent years, the concept of industry 4.0 has gained popularity. Smart systems are used to provide more digitised systems and network integration in 4.0. The fourth industrial revolution is leaning towards the use of data produced for predictive maintenance and early error detection during the production process. The industry 4.0 involves the use of industrial internet of things (IIOT), smart manufacturing and cloud based manufacturing [7]. Manufacturing and service improvements based on Cyber-Physical Systems are two unavoidable developments and challenges for the manufacturing industry. Interacting with many surrounding systems that have a serious influence on machine output can lead to enhanced intelligence [8]. The fundamental requirements of Industry 4.0 include real-time data monitoring, tracking product status and positions, and storing instructions to regulate production operations. The future Manufacturing Execution Systems needs decentralization, vertical integration, and connectivity, cloud computing and advanced analysis [9]. Integration and networking is required spanning the entire value chain (horizontal axis) and at all organizational level (vertical axis). Small decentralised and digitalized production networks functioning autonomously and thus capable of effectively directing their activities in relation to environmental factors and strategic goals characterise the potential of production as foreseen by Industry 4.0 [10]. The basic components of industry 4.0 include innovative digital business models, digitization of product and service offerings and digitization and integration of digital and vertical value chains. The main principles of industry 4.0 are:

- Interoperability – Interchanging machines to perform the same task
- Decentralization – Decision making capability at the local level
- Virtualization – Use of virtual twins and simulation models
- Real-time capability – Data collection and analysis in real-time
- Modularity – Use of flexible and modular systems
- Service orientation – Aid the formation of the product-service system [11].

Integration at the horizontal level and vertical level is described as the most important aspect of industry 4.0. To achieve this various pillars of industry 4.0 have been proposed by researchers as shown in the diagram. With the developing technologies, the Internet is enabling a massive increase in productivity. The Internet of Things (IoT) connects every device, company, residence, and vehicle in an intelligent network in an Internet as network communications, energy, and transportation, with all of them incorporated in a single system [12].

3. Intelligent manufacturing and Smart manufacturing: An intelligent manufacturing system is a hybrid intelligent system that combines humans, cyber systems, and physical systems to achieve specified manufacturing objectives at a high level of efficiency. Machine intelligence technologies like intelligent sensing, autonomous cognition, intelligent decision-making and intelligent control [13]. Smart manufacturing emphasises the use of Information and communications

technology and advanced data analytics to optimize manufacturing processes at different levels of the product supply chain [14]. Fig 1 shows the intelligent manufacturing system technologies.

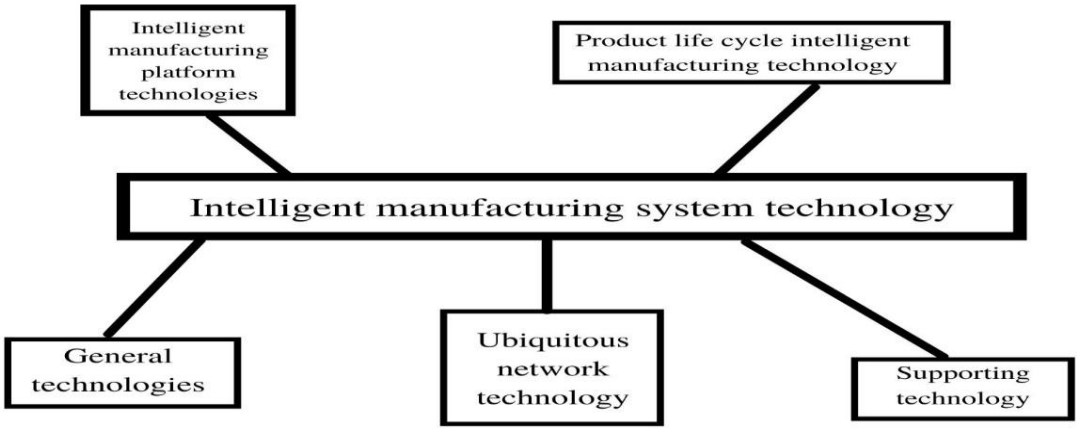


Fig: 1 Intelligent manufacturing system technologies

The word "smart manufacturing" implies a future state of manufacturing in which information from across the industry is transmitted and analyzed in real-time, resulting in manufacturing information that can be utilized to improve all areas of operation. Big data technology and processes are expanded to satisfy the demands of production in smart manufacturing, which may be regarded as a speciality of big data. Machine learning, simulation, cyber-physical systems (CPS) and the internet of things is one of the important technologies of smart manufacturing [15]. The term "smart manufacturing" refers to the widespread adoption and use of networked, information-based technologies in the manufacturing and supply chain industries. The importance of production equipment maintenance in smart manufacturing cannot be overstated. Smart manufacturing relies on real-time and internet-aware ubiquitous networks, as well as closely integrated and sophisticated data-driven analytics systems [16]. The implementation of smart manufacturing generally include four phases: 1) Data integration and contextualization 2) Simulation, modelling and analytics 3) process and product innovation [17]

4. Manufacturing and data analysis: For sustaining continuous improvement in manufacturing, it is vital to understand the origins of underlying causes of issues making data analysis a tool of great importance in manufacturing. The optimization of production processes in analytical process manufacturing is a unique problem from the perspective of information systems [18]. Data analytics (DA) technologies are supposed to evaluate data and provide decision-makers with actionable insight, which can solve many issues arising during manufacturing [19]. The data analytics based decision making will play a vital role in deciding the growth in competitiveness, productivity and innovation of an industry [20]. Data Analytics and associated applications are a collection of approaches and techniques for extracting useful information and insights from enormous volumes of data (Zhou Liu, Zhou 2016) [21]. Manufacturing system control has been improved using predictive analytics approaches. Statistical techniques or machine learning algorithms can be used for improved manufacturing performance by altering the system's present and future states [22]. The most essential big data aspect in producing high-quality solutions is data quality. For efficient solutions to problems during manufacturing, subject matter knowledge in analytics is necessary [23]. To upgrade the existing analytical capabilities and incorporate predictive analytics of the manufacturing system in terms of increase in volume, speed and type of data, Big data analytics can be used [24]. Due to the complexity of manufacturing processes, the business environment of suppliers, industrial roadmaps and standards the solution to these problems require various data collection and analysis approaches [25]. Data analytics finds applications in 1) manufacturing system control 2) Manufacturing quality control. 3) Fault diagnosis of manufacturing equipment. 4) Predictive maintenance of manufacturing equipment. For manufacturing companies, maintenance of process is crucial. Enterprises may use predictive modelling to prepare for maintenance and do cost savings during maintenance and by preventing breakdowns The data analytics platform for industry may include the following aplications: 1) Production data pattern analysis 2) Operation pattern analysis 3) Quality prediction 4) Predictive facility maintenance 5) Integrated 4M monitoring [14]. Fig:2 shows the core technologies used to achieve the above mentioned applications.

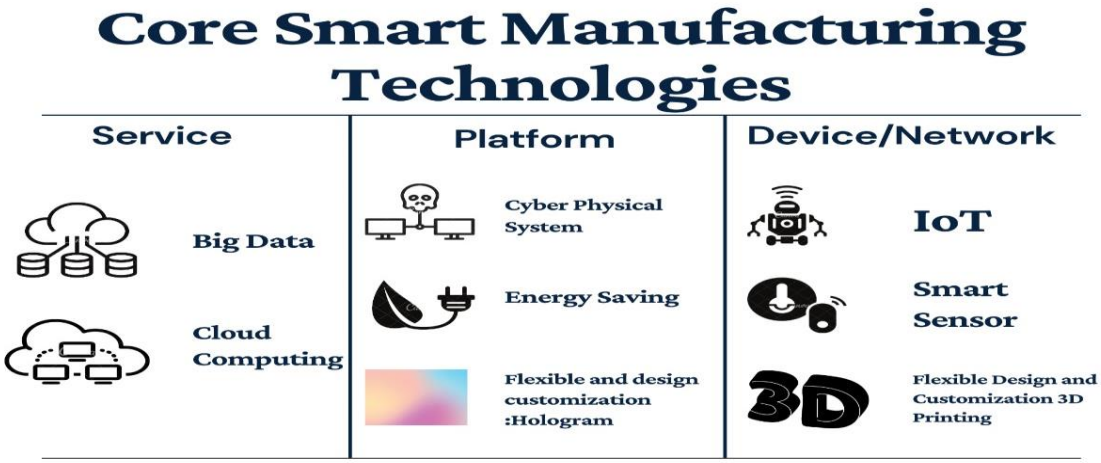


Fig: 2 Core smart manufacturing technologies

Manufacturing execution system and bog data analysis are important technologies for digitalization of the firm. Big data analysis relates the data collected at the firm's machine shop floor to the data collected from Enterprise resource planning level [27].

5. Manufacturing and AI: There is no denying that the industry is at the forefront of AI application. Manufacturers are using AI-powered analytics to increase efficiency, product quality, and employee safety, from major reductions in unplanned downtime to better-designed goods [28]. A deep-learning AI programme might be taught using a company's data, perform simulations of possible futures, and design algorithms to offer the desired business efficiency in the future. Artificial intelligence algorithms estimate market demand by looking for patterns that link geographical, social and macroeconomic aspects, weather patterns, political status, consumer behaviour, and other things. Manufacturers benefit greatly from this data since it allows them to manage personnel, inventory control, energy consumption, and raw material supply. Machines, sensors, controllers and labour records on the machine shop floor generate large amount of data continuously. This data can be in the form of environmental data from ambient sensors, process data, production operation data and data from quality inspection [29]. This large volumes of complicated industrial data, which has become a standard in industry, may be transformed into useful and meaningful information using AI [30]. The Machine learning and Deep learning technologies of Artificial intelligence provides data analysis at micro level in manufacturing such as the use of Artificial intelligence for the use of machine condition monitoring and fault diagnosis [31]. AI finds its application in 1) Modelling and performance analysis 2) Manufacturing decision making and control 3) Manufacturing applications of human-robot collaboration 4) Condition- based maintenance 5) manufacturing control 6) [29]. The AI/ ML based methods are able to attain targets like 1) prior recognition of quality defects. 2) Root cause diagnosis of quality issues [32]. Machine-learning technology and pattern-recognition software are at the heart of AI's foray into manufacturing, and they may hold the key to revolutionising factories in the near future. If firms want to take advantage of the growth prospects are given by this new era of intelligent manufacturing, ERP and manufacturing execution systems (MES) must be connected. The benefits of using AI in manufacturing includes round clock production, a secure operating environment, reduced operating costs, quick data-driven decisions.(article). The internet of things (IoT) with its elements like IoT products, IoT devices, IoT gateway devices, industrial IoT development kits, play a pivotal role in industrial automation [33].

6. Conclusion: With the emergence of Industry 4.0 the use of data analysis and AI is gaining momentum in the manufacturing industries. Due to the complexity of manufacturing processes, the business environment of suppliers, industrial roadmaps and standards, the solution to these problems require various data collection and analysis approaches. The need for data analytics in manufacturing arises due to the factors like (1) the demand for precise manufacturing; (2) cost pressures necessitating quick implementation of new technologies and constant progress; (3) process complexity and, in many cases, a lack of process visibility; and (4) process dynamics. Different AI tools have already been implemented at various hierarchical level of manufacturing firms. Due to large amount of data available from the sensors, controllers and manufacturing equipment AI tools have been extensively used for monitoring, diagnosis and prognosis of the manufacturing process. Data analytics and AI tools find their application in a variety of domains including, production, supply chain, maintenance and diagnostics, quality management, and energy management .Despite the issues regarding data collection and problem formation the future trend is undoubtedly towards more usage of data analytics tool and AI tools in manufacturing firms. AI and data analytics technology will be able to through its analytical techniques provide value added to companies with knowledge for improvements in manufacturing processes and future decision making

REFRENCES

[1] Schwab, K. (2016). The fourth industrial revolution, Word Economic Forum, Geneva.

[2] Carbery CM, Woods R, Marshall A(2019). A new data analytics framework emphasising preprocessing of data to generate insights into complex manufacturing systems. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science. 2019;233(19-20):6713-6726. doi:10.1177/0954406219866867

[3] Zhang, D., Xu, B., Wood, J(2016).: Predict failures in production lines. In: IEEE International conference on Big Data, pp. 2070–2074. Washington, USA (2016).

[4] Kusiak, A.(2018), Smart manufacturing, International Journal of Production Research, 2018 Vol. 56, Nos. 1–2, 508–517, https://doi.org/10.1080/00207543.2017.1351644.

[5] Y. Zhang, X. Xu, A. Liu, Q. Lu, L. Xu and F. Tao, "Blockchain-Based Trust Mechanism for IoT-Based Smart Manufacturing System," in *IEEE Transactions on Computational Social Systems*, vol. 6, no. 6, pp. 1386-1394, Dec. 2019, doi: 10.1109/TCSS.2019.2918467.

[6] Zhang, Hankun & Liu, S. & Morača, Slobodan & Ojstersek, Robert. (2017). An Effective Use of Hybrid Metaheuristics Algorithm for Job Shop Scheduling Problem. International Journal of Simulation Modelling. 16. 644-657. 10.2507/IJSIMM16(4)7.400.

[7] Lu, Y. , Morris, K. and Frechette, S. (2016), Current Standards Landscape for Smart Manufacturing Systems, NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.IR.8107 (Accessed March 16, 2022).

[8] Lee, Jay & Kao, Hung-An & Yang, Shanhu. (2014). Service Innovation and Smart Analytics for Industry 4.0 and Big Data Environment. Procedia CIRP. 16. 3–8. 10.1016/j.procir.2014.02.001.

[9] Hirsch-Kreinsen, H. (2016), "Digitization of industrial work: development paths and prospects", Journal for Labour Market Research, Vol. 49 No. 1, pp. 1-14

[10] Erol, Selim & Jäger, Andreas & Hold, Philipp & Ott, Karl & Sihn, Wilfried. (2016). Tangible Industry 4.0: a scenario-based approach to learning for the future of production. 10.1016/j.procir.2016.03.162.

[11] Carvalho, Nubia & Chaim, Omar & Cazarini, Edson & Gerolamo, Mateus. (2018). Manufacturing in the fourth industrial revolution: A positive prospect in Sustainable Manufacturing. Procedia Manufacturing. 21. 671-678. 10.1016/j.promfg.2018.02.170.

[12] Lu, Yang. (2017). Industry 4.0: A Survey on Technologies, Applications and Open Research Issues. Journal of Industrial Information Integration. 6. 10.1016/j.jii.2017.04.005.

[13] Ji.Z, Yanhong.Z, Baicun.W, Jiyuan.Z (2019) .Human–Cyber–Physical Systems (HCPSs) in the Context of New-Generation Intelligent Manufacturing[J].Engineering,2019,5(4):624-636.

[14] Zhong, Ray & Xu, Xun & Klotz, Eberhard & Newman, Stephen. (2017). Intelligent Manufacturing in the Context of Industry 4.0: A Review. Engineering. 3. 616-630. 10.1016/J.ENG.2017.05.015.

[15] Kang, Hyoung & Lee, Ju & Choi, Sangsu & Kim, Hyun & Park, J. & Son, Jiyeon & Kim, Bo & Noh, Sang Do. (2016). Smart manufacturing: Past research, present findings, and future directions. International Journal of Precision Engineering and Manufacturing-Green Technology. 3. 111-128. 10.1007/s40684-016-0015-5.

[16] Lee, Ju & Yoon, Joo & Kim, Bo Hyun. (2017). A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: An empirical case study of a die casting factory. International Journal of Precision Engineering and Manufacturing. 18. 1353-1361. 10.1007/s12541-017-0161-x.

[17] O'Donovan, P., Leahy, K., Bruton, K. *et al.* An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *Journal of Big Data* **2,** 25 (2015). https://doi.org/10.1186/s40537-015-0034-z.

[18] Shah, Devarshi & Wang, Jin & He, Q.. (2020). Feature Engineering in Big Data Analytics for IoT-Enabled Smart Manufacturing – Comparison between Deep Learning and Statistical Learning. Computers & Chemical Engineering. 141. 106970. 10.1016/j.compchemeng.2020.106970.

[19] Wuest, Thorsten & Weimer, Daniel & Irgens, Chris & Thoben, Klaus-Dieter. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. Production & Manufacturing Research. 4. 23-45. 10.1080/21693277.2016.1192517.

[20] Manyika, James & Chui, Michael & Brown, Brad & Bughin, Jacques & Dobbs, Richard & Roxburgh, Charles & Byers, Angela. (2011). Big data: The next frontier for innovation, competition, and productivity.

[21] Zhou, Keliang & Liu, Taigang & Zhou, Lifeng. (2015). Industry 4.0: Towards future industrial opportunities and challenges. 2147-2152. 10.1109/FSKD.2015.7382284.

[22] Li, Jing & Shi, Jianjun. (2007). Knowledge discovery from observational data for process control using causal Bayesian networks. Iie Transactions. 39. 681-690. 10.1080/07408170600899532.

[23] Davis, Jim & Edgar, Thomas & Porter, James & Bernaden, John & Sarli, Michael. (2012). Smart manufacturing, manufacturing intelligence and demand-dynamic performance. Computers & Chemical Engineering. 47. 145–156. 10.1016/j.compchemeng.2012.06.037.

[24] J. Moyne, J. Samantaray and M. Armacost, "Big Data Capabilities Applied to Semiconductor Manufacturing Advanced Process Control," in *IEEE Transactions on Semiconductor Manufacturing*, vol. 29, no. 4, pp. 283-291, Nov. 2016, doi: 10.1109/TSM.2016.2574130.

[25] Moyne J, Iskandar J. Big Data Analytics for Smart Manufacturing: Case Studies in Semiconductor Manufacturing. *Processes*. 2017; 5(3):39. https://doi.org/10.3390/pr5030039.

[26] Lechevalier, David & Narayanan, Anantha & Rachuri, Sudarsan. (2014). Towards a Domain-Specific Framework for Predictive Analytics in Manufacturing. Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014. 10.1109/BigData.2014.7004332.

[27] Frank, Alejandro & Dalenogare, Lucas & Ayala, Néstor. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. International Journal of Production Economics. 210. 10.1016/j.ijpe.2019.01.004.

[28] Kushmaro, P. (2018). CIO: 5 ways industrial AI is revolutionizing manufacturing. Available from: https://www.cio.com/article/3309058/. Accessed on 2019-03-07.

[29] Arinez, J. F., Chang, Q., Gao, R. X., Xu, C., and Zhang, J. (August 13, 2020). "Artificial Intelligence in Advanced Manufacturing: Current Status and Future Outlook." ASME. *J. Manuf. Sci. Eng*. November 2020; 142(11): 110804. https://doi.org/10.1115/1.4047855.

[30] Wang, Jinjiang & Ma, Yulin & Zhang, Laibin & Gao, Robert & Wu, Dazhong. (2018). Deep Learning for Smart Manufacturing: Methods and Applications. Journal of Manufacturing Systems. 48. 144-156. 10.1016/j.jmsy.2018.01.003.

[31] Liu, Ruonan & Yang, Boyuan & Zio, Enrico & Chen, Xuefeng. (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review. Mechanical Systems and Signal Processing. 108. 33-47. 10.1016/j.ymssp.2018.02.016.

[32] Wang, Guodong & Ledwoch, Anna & Hasani, Ramin & Grosu, Radu & Brintrup, Alexandra. (2019). A Generative Neural Network Model for the Quality Prediction of Work in Progress Products. Applied Soft Computing. 85. 10.1016/j.asoc.2019.105683.

[33] https://www.plantautomation-technology.com/articles/impact-of-internet-of-things-on-industrial-automation

CYBER SECURITY AND THE VULNERABILITY OF THE INDIAN BANKING SECTOR: A REVIEW PAPER

Amisha Tiwari, Abrar Ali Khan, Jashanpreet Singh Toor

ABSTRACT:—— With the digitalization and advancement of e-banking technology, the confidentiality of every user is put at stake. The total cases of frauds reported by banks/FIs increased by 28 per cent by volume and 159 per cent by value during 2019-20 as reported by the Reserve Bank of India (RBI). Financial Institutions are 300 times more vulnerable than any other sector. The banking sector needs a structural reformation and more stringent laws to strengthen fraud detection and consumer protection. The undermentioned paper reviews the evolving cyber threats and propose suitable solutions on collective protection strategies. With the amalgamation of the everchanging geo-political landscape and an informed consumer, cyberspace can be harnessed to its fullest potential.

Introduction: Embattled by new waves of infections and mutant strains of COVID-19, the global and domestic outlook had once again turned grim and overcast with extreme uncertainty and downside risks. This de facto has extensively affected cybercrime in banks as they have shown a tectonic increase in cases over the past year. Online banking or e-banking is an electronic payment system that enables customers of a financial institution to conduct financial transactions on a website operated by the institution, such as a retail bank, virtual bank, credit union or building society. The continuous metamorphosis of the banking sector has improved the customer base due to the instant availability of money in a single click. However, this convenience compromises the security of the user as it gives rise to cyber threats, which in turn attempts to infiltrate or disrupt a computer network/system. Cyber-attacks have become an easy option for cybercriminals to access confidential data through the Internet. Cybercriminals go where the money is, and banks have more money than other organizations. Typically, hackers are targeting customers' data and funds, as well as the bank's core systems. These attacks have become highly targeted from hacking the bank accounts of individuals, companies, governments and demanding heavy ransoms to decrypt the force-encrypted data.

Cyber Security in Banks

Privacy preserved Banking sector consists of three intangible factors: applications, which have data to share with authorized clients; secondly, clients who want data contained in the applications; and finally, the privacy control factor required to maintain records about the purposes. It has become possible for innumerable computers operating on different platforms to communicate with each other over the Internet because they adopt the same communication protocol, thereby compromising the privacy of every user. A bank runs multiple servers that store an enormous amount of information and details of various operations such as credit cards, ATMs, real-time gross settlements, ATMs and SWIFT (the global financial messaging service banks use to move funds), among others.

The main aim of these attacks is to take over the user's bank accounts and funds, so the attacker occupies the funds without proper user knowledge.

The emergence of cybercrimes over the years

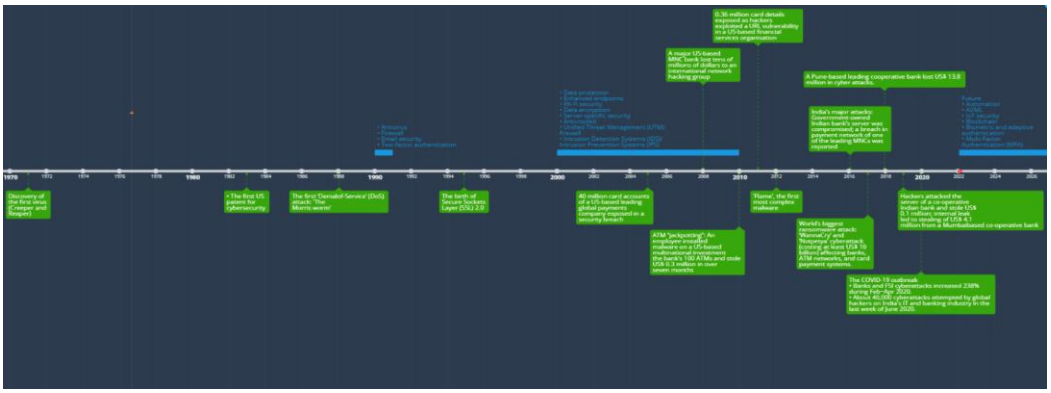


Figure 1: Timeline showing trends of cybercrime over the years

Significant vulnerabilities in the banking sector:

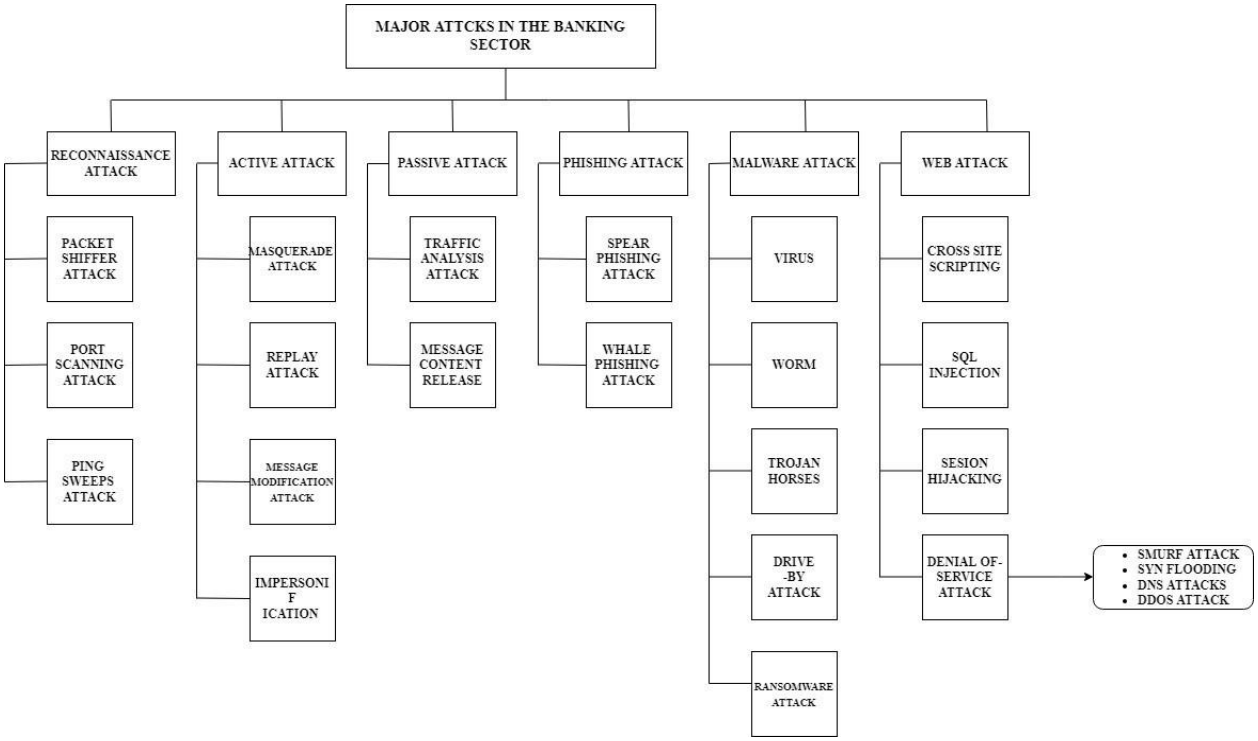


Figure 2: Flowchart showing various cyber-attacks affecting the banking industry

Phishing or Identity Theft: Hackers use it to steal sensitive information such as credit card numbers, usernames, passwords, PINs, bank account numbers, and personal information. Phishing is often carried out through email spoofing, which involves sending out bogus emails demanding the user's personal information.

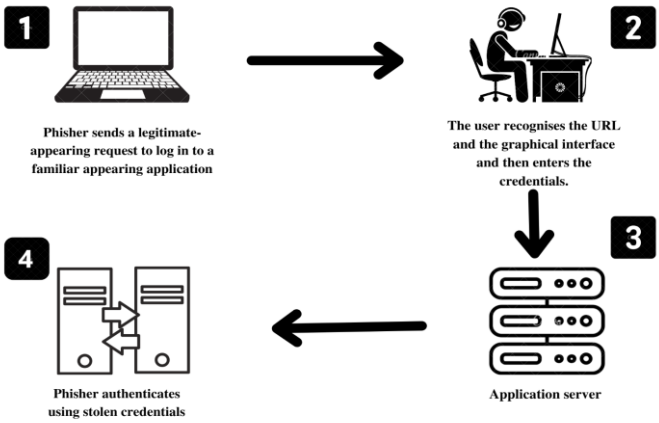


Figure 3: Block diagram depicting the mechanism of phishing

Keystroke Logging or Keylogging: Keylogging is a technique used by fraudsters to capture actual keystrokes and mouse events. Key loggers are "Trojan" software applications that target a computer's operating system and are "installed" by a virus. These are especially hazardous since the fraudster records the user ID and password, account number, and input.

Spyware: Spyware is the most commonly used method for stealing banking passwords for fraudulent purposes. Spyware collects information on the computer when exchanged between the computer and websites. It is frequently installed through bogus "pop up" adverts requesting software downloads.

Watering hole: "Watering hole" cyber fraud is regarded as a subset of phishing assaults. Malicious code is injected onto a website's public web pages in a watering hole, only frequented by a limited number of individuals. When a victim accesses a site containing malicious code in a watering hole attack, the information is tracked. In a phishing attempt, the target unwittingly gives up personal information. In contrast, the assailant waits for the victim to visit the location at a watering hole. The attackers typically compromise the site utilized for an assault months before the actual attack. The watering hole is a surgical assault approach in which hackers target only a limited number of individuals on the Internet, and it is less audible than phishing.

Credit Card Redirection and Pharming: In Pharming, attackers hijack a bank's URL so that when a customer checks in to the bank's website, he or she is led to another false website that is a clone of the bank's original website. Pharming is done over the Internet, and ATM skimming is another method.

DNS Cache Poisoning: Poisoning attacks against a DNS server make use of DNS software flaws. As a result, the server improperly validates DNS answers that assure they are from an authoritative source. Incorrect entries are cached locally by the server and served to other users who make the same request. Victims of a banking website are led to a server owned by criminals, who exploit it to deliver malware or trick bank clients into providing their credentials to a spoof of an actual website. An attacker can hijack clients if they spoof an IP address DNS entry for a bank website on a specific DNS server and replace them with the IP address of a server they control.

Hacking: The primary hazards associated with hacking include strategic risk, business risk, operational risk, security risk, privacy/security risk, legal risk, cross-border risk, reputational risk, liquidity risk, and many more. These hazards are closely interconnected, and events affecting one risk category might have repercussions for various other risk categories. Black-hat hackers operate unlawfully, white-hat hackers are ethical hackers employed on a contract basis by an enterprise to check computer system vulnerabilities, and grey-hat hackers seek to enhance the system and network security. They breach the system's security flaw without authorization in order to expose that flaw to the system's owners, and blue-hat hackers work for outside computer security consulting firms. More hackers include Script Kiddies, Elite Hackers, Hacktivists, and Phreakers.

Password Sniffing: Cybercriminals utilize this approach to break user passwords. Password sniffers are programs used by cybercriminals to monitor and collect network users' names and passwords when they check in to a website.

Denial of Service Attacks: It attempts to render a system or network inaccessible to its users, either temporarily or permanently. It happens when a computer receives more requests than it can process. This type of attack can cause systems to hang. Target sites are often those hosted by banks, credit card payment gateways, and others.

Banking Sector in Jeopardy

To remain relevant and competitive, banks must fulfil stakeholder expectations while averting risks and following regulatory obligations by enhancing their cyber defence activities. Banks can expect more sophisticated assaults as they invest more resources on digital platforms. These assaults will migrate from in-house devices to those housed on digital platforms accessed by many stakeholders, and they will directly target end-user computing environments. Banks will have to proactively address cyber threats from many security parts such as data, application, identity, infrastructure, and cloud and oversee end-user education and regulatory compliances. They must regularly address security gaps, develop a security roadmap, review and benchmark best practices, and make strategic investments in cybersecurity core areas based on business needs and risk appetite.

Conclusion: Since organizations began to use computers, cybercrime's complexity and financial consequences have increased. With the usage of credit and debit cards expanding daily and new technologies such as online wallets slowly gaining traction, financial transactions are at an all-time high. Banking sectors are vulnerable to multiple disruptions caused by various risks; numerous hazards are divided into various domains, such as cyber fraud, trade permanency development, and information security measures. Hacking, credit card fraud, money laundering, DoS attacks, phishing, salami attacks, ATM card cloning, and so on are examples of cybercrimes conducted in banks. Cyber dangers such as Pharming, phishing, and the tempted disclosure of private information such as identity theft are among the security concerns clients in the banking and financial industries have. Because of countless online business transactions and frantic network traffic, which creates massive data, only a portion of which relays to criminal behaviours, detecting cybercrime may be incredibly difficult. Cybercriminals are growing more adept, increasingly targeting consumers and public and private businesses. Cybercrime analysis has a very momentous responsibility on the law enforcement system in any country. Our law enforcement organizations must be adequately outfitted to defeat and prevent cybercrime. *Cybercrime* is a crime that is harder to detect and hardest to stop once it occurred causing a long-term negative impact on victims. With the increasing popularity of online banking, online shopping requires sensitive personal and financial data; it is a term that we hear in the news with some frequency. To protect ourselves from this crime, we need to know what it is and how it works against us.

References

1. Harshita, B. (2015). Metamorphosis of Banking Products-A Perception of Bank Employees. *The Journal of Internet Banking and Commerce*, *20*(2).
2. Singh, S., & Kumar, S. (2020). THE TIMES OF CYBER ATTACKS. *Acta Technica Corviniensis-Bulletin of Engineering*, *13*(3), 133-137.
3. ELECTRONIC CRIME IN INDIAN BANKING Pooja Pasricha1 Research Scholar, Depart
4. Kshetri, N. (2017). Cybersecurity in India: Regulations, governance, institutional capacity and market mechanisms. *Asian Research Policy*, *8*(1), 64-76.
5. Lekha, K. C., & Prakasam, S. (2017, August). Data mining techniques in detecting and predicting cyber crimes in banking sector. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)* (pp. 1639-1643). IEEE.

6. Gunjan, V. K., Kumar, A., & Avdhanam, S. (2013, September). A survey of cyber crime in India. In *2013 15th International Conference on Advanced Computing Technologies (ICACT)* (pp. 1-6). IEEE.
7. Kesharwani, S., Sarkar, M. P., & Oberoi, S. (2019). Growing Threat of Cyber Crime in Indian Banking Sector. *Cybernomics*, *1*(4), 19-22.
8. Bhasin, M. (2007). Mitigating cyber threats to banking industry. *The chartered accountant*, *50*(10), 1618-1624.
9. Singh, O., Gupta, P., & Kumar†, R. (2016). A Review of Indian Approach towards Cybersecurity. *International Journal of Current Engineering and Technology*, *6*(2), 644-648.
10. Godbole, T., Gochhait, S., & Ghosh, D. (2022). Developing a Framework to Measure Cyber Resilience Behaviour of Indian Bank Employees. In *ICT with Intelligent Applications* (pp. 299-309). Springer, Singapore.
11. Raghavan, A. R., & Parthiban, L. (2014). The effect of cybercrime on a Bank's finances. *International Journal of Current Research & Academic Review*, *2*(2), 173-178.
12. Mahapatra, C., VIPS, G., & Chopra, I. M. Towards Digital India Transformation: Pragmatic Implementation of 3 Dimensional Cyber Security Pyramid to Counter Cyber Attacks in India.
13. Kshetri, N. (2016). Cybercrime and cybersecurity in India: causes, consequences and implications for the future. *Crime, Law And Social Change*, *66*(3), 313-338. doi: 10.1007/s10611-016-9629-3
14. *RESERVE BANK OF INDIA ANNUAL REPORT 2020-21*. (2021).
15. Bamrara, D., Singh, G., & Bhatt, M. (2013). Cyber Attacks and Defense Strategies in India: An Empirical Assessment of Banking Sector. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2488413
16. Reddy, L., & Bhargavi, V. (2018). Cyber security attacks in banking sector: Emerging security challenges and threats. *American International Journal of Research in Humanities, Arts and Social Sciences*, *21*(1), 65-71.
17. Khatik, S. K., & Nag, A. K. (2015). PERFORMANCE MEASUREMENT SYSTEM IN INDIAN BANKING SECTOR IN CAMEL FRAMEWORK. *Delhi Business Review*, *16*(1).
18. More, D. M. M., & Nalawade, M. P. J. D. K. (2015). Online banking and cyber-attacks: the current scenario. *International Journal of Advanced Research in Computer Science and Software Engineering Research Paper*.
19. Kaur, J., & Ramkumar, K. R. (2021). The recent trends in cyber security: A review. *Journal of King Saud University-Computer and Information Sciences*.
20. Singh, Gaurav. (2021). CYBER SECURITY: A BOON TO DIGITAL INDIA.
21. MANIVANNAN, A. CYBER ATTACKS IN THE BANKING INDUSTRY.
22. Sirisha, T., & Kalyan, N. B. Jeopardy and Arrival Analysis of Equity With Reference To Indian Banking Sector.
23. KATHIRIYA, D. D., & PATEL, R. S. Heterogeneous Framework for Indian Cybercrime Cases.
24. Sarmah, A., Sarmah, R., & Jyoti Baruah, A. (2017). A brief study on cyber crime and cyber laws of India. *International Research Journal of Engineering and Technology (IRJET)*, *4*(6), 1633-1640.
25. Acharya, S., & Joshi, S. (2020). Impact of cyber-attacks on banking institutions in India: A study of safety mechanisms and preventive measures. *PalArch's Journal of Archaeology of Egypt/Egyptology*, *17*(6), 4656-4670.
26. Bamrara, D., Singh, G., & Bhatt, M. (2013). Cyber attacks and defense strategies in India: An empirical assessment of banking sector. *Gajendra and Bhatt, Mamta, Cyber Attacks and Defense Strategies in India: An Empirical Assessment of Banking Sector (January 1, 2013)*.
27. Kumbhar, V. M. (2011). Factors affecting the customer satisfaction in e-banking: Some evidences form Indian banks. *Management Research & Practice*, *3*(4).
28. Deloitte. (2020, November). *Cybersecurity in the Indian banking industry: Part 1*. https://in-ra-cybersecurity-in-the-indian-banking-industry-noexp.pdf

DATA MINING APPLICATIONS IN HEALTHCARE SECTOR: A REVIEW

Diksha Rattan[#1], Jasvir Singh[*2]
[#]*Computer Science Department, Punjabi University*
[1]Dikshrattan111@gmail.com
[2]Jassiccet@gmail.com

ABSTRACT—— In this paper, the focus was to compare a variety of techniques, approaches and different tools and its impact on the healthcare sector. The goal of data mining application is to turn that data are facts, numbers, or text which can be processed by a computer into knowledge or information. The main purpose of data mining application in healthcare systems is to develop an automated tool for identifying and disseminating relevant healthcare information. This paper aims to make a detailed study report of different types of data mining applications in the healthcare sector and to reduce the complexity of the study of the healthcare data transactions. Also presents a comparative study of different data mining applications, techniques and different methodologies applied for extracting knowledge from database generated in the healthcare industry. Finally, the existing data mining techniques with data mining algorithms and its application tools which are more valuable for healthcare services are discussed in detail.

Keywords— Data Mining, Knowledge Discovery Database, In-Vitro Fertilization (IVF), Artificial Neural Network, WEKA, NCC2.

I. INTRODUCTION

The purpose of data mining is to extract useful information from large databases or data warehouses. Data mining applications are used for commercial and scientific sides. This study mainly discusses the Data Mining applications in the scientific sides[1]. Scientific data mining distinguishes itself in the sense that the nature of the datasets is often very different from traditional market driven data mining applications. In this work, a detailed survey is carried out on data mining applications in the healthcare sector, types of data used and details of the information extracted. Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases. There are a large number of data mining applications are found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management. To find the useful and hidden knowledge from the database is the purpose behind the application of data mining. Popularly data mining called knowledge discovery from the data. The knowledge discovery is an interactive process, consisting by developing an understanding of the application domain, selecting and creating a data set, preprocessing, data transformation. Data Mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile and mobile computing. systems to produce reliable reports with respect to other information in purely financial and volume related statements. Data mining tools to answer the question that traditionally was a time consuming and too complex to resolve. They prepare databases for finding predictive information. Data mining tasks are Association Rule, Patterns, Classification and Prediction, Clustering. Most common modeling objectives are classification and prediction. The reason that attracted a great deal of attention in information technology for the discovery of useful information from large collections is due to the perception that we are data rich but information poor.Some the sample data mining applications are:

Developing models to detect fraudulent phone or credit- card activity
Predicting good and poor sales prospectus.
Predicting whether a heart attack is likely to recur among those with cardiac disease.
Identifying factors that lead to defects in a manufacturing process.

Expanding the health coverage to as many people as possible, and providing financial assistance to help those with lower incomes purchase coverage [2]. Eliminating current health disparities would decrease the costs associated with the increased disease burden borne by certain population groups.Health administration or healthcare administration is the field relating to leadership, management, and administration of hospitals, hospital networks, and health care systems[1,3]. In the Healthcare sector Government spends more money.

Proposal in draft NHP 2001 is timely that State health expenditures be raised to 7% by 2015 and to 8% of State budgets thereafter[21].

Health spending in India at 6% of GDP is among the highest levels estimated for developing countries.

Public spending on health in India has itself declined after liberalization from 1.3% of GDP in 1990 to 0.9% in 1999. Central budget allocations for health have stagnated at 1.3%ofthe total Central budget. In the States it has declined from7.0% to 5.5% of the State health budget.

This paper mainly compares the data mining tools deals with the health care problems. The comparative study compares the accuracy level predicted by data mining applications in healthcare. Infertility is on the rise across the globe and it needs the sophisticated techniques and methodologies to predict the end results of infertility treatments particulars IVF (in-vitro fertilization) treatments, since the cost of IVF procedure is on the rise. In this study, we have taken this issue and compare the different techniques of data mining applications for predicting the Success rate of IVF treatment with the accuracy

level. This comparative study could be useful for aspiring researchers in the field of data mining by knowing which data mining tool gives an accuracy level in extracting information from healthcare data.

II. REVIEW OF LITERATURE:-

A literature review is a text written by critical points of current including substantive find theoretical and methodological contributions to a particular topic. Literature reviews are secondary sources and do not report any new or original experimental work.

- It mainly discusses data mining and its applications with major areas like Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management[1].
- It presents how data mining discovers and extracts useful patterns of this large data to find observable patterns. This paper demonstrates the ability of Data mining in improving the quality of the decision making process in pharma industry. Issues in the pharma industry are adverse reactions to the drugs[2].
- It illustrates a hybrid prediction system consists of Rough Set Theory (RST) and Artificial Neural Network (ANN) for dispensation medical data. The process of developing a new data mining technique and software to assist competent solutions for medical data analysis has been explained. Propose a hybrid tool that incorporates RST and ANN to make proficient data analysis and indicative predictions. The experiments on spermatological data set for predicting excellence of animal semen is carried out. The projected hybrid prediction system is applied for pre-processing of medical database and to train the ANN for production prediction. The prediction accuracy is observed by comparing observed and predicted cleavage rate[20].
- It mainly examine the potential use of classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Using an age, sex, blood pressure and blood sugar medical profiles it can predict the likelihood of patients getting a heart disease[4].
- The various data mining approaches were discussed that have been utilized for breast cancer diagnosis and prognosis Decision tree is found to be the best predictor with 93.62% Accuracy on benchmark dataset and also on SEER data set[5].

III. DATA MINING

Data mining is the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. With the widespread use of databases and the explosive growth in their sizes, organizations are faced with the problem of information overload. The problem of effectively utilizing these massive volumes of data is becoming a major problem or all enterprises.

A. Definition:-

Data mining or knowledge discovery in database, as it is also known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes, and detecting anomalies[8].

B. Development of data mining:-

The current evaluation of data mining functions and products is the results of influence from many disciplines, including databases, information retrieval, statistics, algorithms, and machine learning[9].

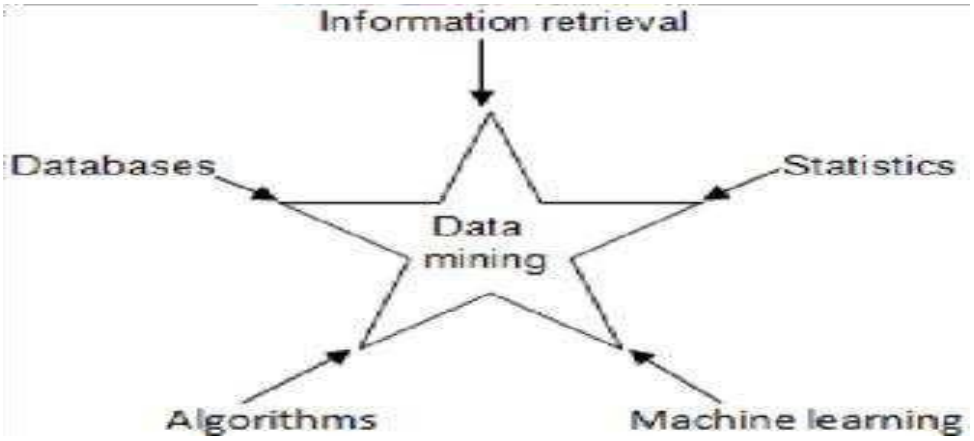


Fig. 1. Historical perspective of data mining

C. History of Data Base and Data Mining

Data mining development and the history represented in the Fig. 2. The data mining system started from the year of 1960s and earlier. In this, the data mining is simply on file processing. The next stage its Database management Systems to be started year of 1970s early to 1980s. In this OLTP, Data modeling tools and Query processing are worked. From database management system there three broad categories to be worked. First one is Advanced Database Systems, this evaluated year of Mid-1980s to present in this Data models and Application oriented process are worked. The Second part is Data

Warehousing and Data Mining worked since the year of the late 1980s to present. The third part is Web based Database Systems this started from 1990s to present and in this Web mining and XML based database systems are included. These three broad categories are joined and create the new process that's called New generation of the Integrated Information system is started in 2000.

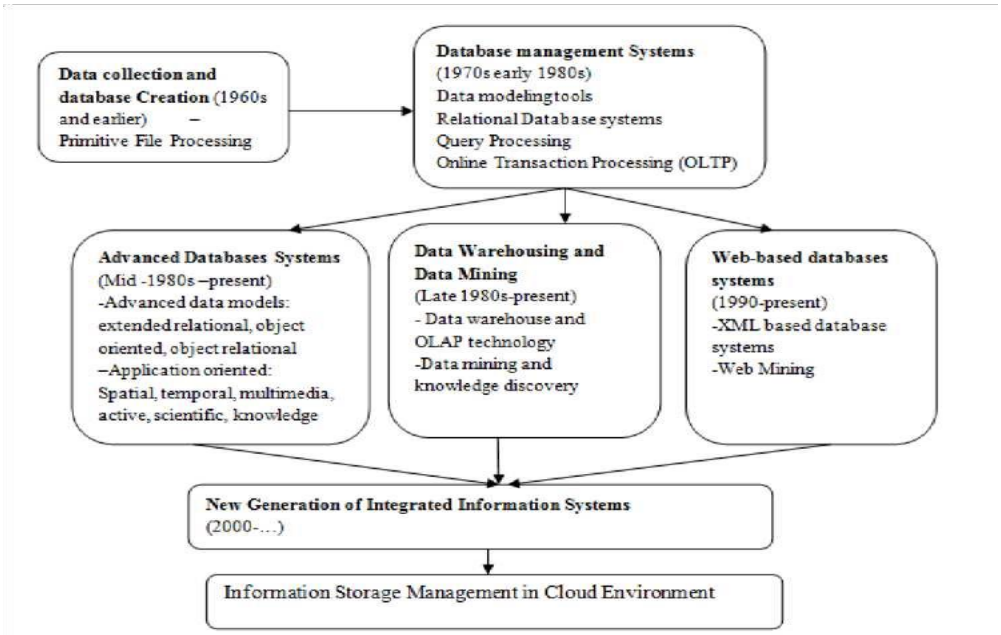


Fig. 2. History of Database Systems and Data Mining

D. Data Mining Application Areas

Data mining is driven in part by new applications which require new capabilities that are not currently being supplied by today's technology. These new applications can be naturally into two broad categories.

- Business and E-Commerce
- Scientific, Engineering and Health Care Data

E. Data Mining Tasks

Data mining tasks are mainly classified into two broad categories:

- Predictive model
- Descriptive model

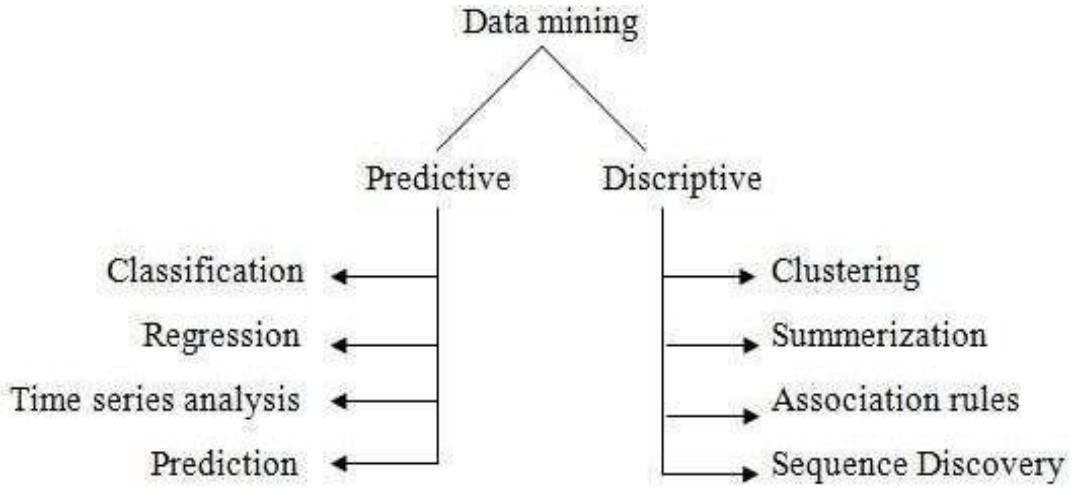


Fig 3.3 Data mining models and tasks

IV. DATA MINING APPLICATIONS IN HEALTHCARE SECTOR

Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into broad categories[1,10].

A. Treatment effectiveness

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

B. Healthcare management

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

C. Customer relationship management

Customer relationship management is a core approach to managing interactions between commercial organizations typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings.

D. Fraud and abuse

Detect fraud and abuses establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims.

E. *Medical Device Industry*

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless biosensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients[11Pharmaceutical Industry.

F. Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized[12]. Three layers of hospital management:

- Services for hospital management
- Services for medical staff
- Services for patients
- System Biology

Biological databases contain a wide variety of data types, often with rich relational structure.Consequently multirelational data mining techniques are frequently applied to biological data[13]. Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.

V. RESULTS OF COMPARATIVE STUDY

This chapter, a comparative study of data mining applications in healthcare sector by different researchers given in detail. Mainly data mining tools are used to predict the successful results from the data recorded on healthcare problems. Different data mining tools are used to predict the accuracy level in different healthcare problems. In this study, the following list of medical problems has been analyzed and evaluated.

- Heart Disease
- Cancer
- HIV/AIDS
- Blood
- Brain Cancer
- Tuberculosis
- Diabetes Mellitus
- Kidney dialysis
- Dengue
- IVF
- Hepatitis C

In the Table 1,the most important healthcare problems specifically in disease side and research results have been illustrated. The diseases are the most critical problems in human. To analyze the effectiveness of the data mining applications for diagnosing the disease, the traditional methods of mathematical / statistical applications are also given and compared. Listed eleven problems are taken for comparison with this work.

TABLE 1. DATA MINING APPLICATIONS IN HEALTHCARE

Graph chart formed by using this table with the values of health care problems, Data Mining tools and Accuracy Level is as illustrated in Fig. 2. In this chart, the prediction accuracy level of different data mining applications has been compared

S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, NCC2	Classification	Naïve	Probability	60
2	Cancer	WEKA	Classification	Rules. Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFIA		85
6	Tuberculosis	WEKA	Naïve Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

VI. COMPARATIVE STUDY OF IVF SUCCESS RATE PREDICTION

The section deals with the comparative study of three different data mining application for predicting the success rate of IVF treatment. The process of data mining applications, its advantages and results obtained are compared. The detailed study of selected works gives a broad idea about the application of data mining techniques. This study mainly compares the three different data mining applications carried out on the prediction of the IVF treatment success rate.

A. Application of rough set theory for medical informatics data analysis

The research work aims to analyze the medical data by applying Rough Set Theory of data mining approach. The data reduction process has been done using rough set theory reduction algorithm. Rough set is mainly used to reduce the attributes without compromising its knowledge of the original. To analyze the fertilization data, ROSETTA tool kit reduction algorithm is used in this work to produce the optimal reduct set without affecting the original knowledge. The treatment success rate is predicted and tabulated as depicted in Table 2.

TABLE 2. IVF SUCCESS RATE PREDICTED BY ROUGH SET

		Predicted		
Actual		SUCCESS	UN SUCCESS	
	SUCCESS	17	4	0.80952
	UN SUCCESS	26	10	0.27777
		0.395349	0.714286	0.47368

The actual and desired outputs are compared with each other. It also depicts that the success rate obtained after reducing the number of attributes is 47%.

B. Artificial neural network in classification and prediction

This research work is mainly aimed to predict and classify the IVF treatment results using Artificial Neural Network (ANN). The artificial neural network is constructed with multi-layer perception and back-propagation training algorithm, and constructed network is trained, tested and validated using patients' sample IVF data. This work finally compares the success rate between desired output which is field recorded data and actual output which is predicted output of neural network. In the Table 3, the comparison between desired and actual output of the neural network is illustrated.

TABLE 3. IVF SUCCESS RATE PREDICTED BY

Performance	DESIRED OUTPUT	ACTUAL NETWORK OUTPUT
MSE	0.209522132	0.212860733
NMSE	1.164459543	1.18301446
MAE	0.23114814	0.25780224
Min Abs Error	9.90854E-07	6.66044E-06
Max Abs Error	1.015785003	0.998857054
R	0.498099362	0.498099362
Percent Correct	73.07692308	75

This work finds the actual output using patients' IVF data by applying Artificial Neural Network. By comparing success rate, desired and actual output, the result obtained has a prediction accuracy of 73%.

C. Modeling an integrated methodology of neural networks and rough sets for analyzing medical data

This work is mainly aimed to develop a combined prediction system for analyzing medical data using Artificial Neural Network and Rough Set Theory. Two kinds of rules Deterministic and Non-deterministic are effected in the application of Rough set tool. For the rough set application, the software tool Neuro solution is used to predict the result. The performance of the combined technique of Artificial neural network and rough set theory is described in the Table 4.

TABLE 4. PERFORMANCE OF IVF SUCCESS RATE PREDICTION USING HYBRID TECHNIQUE

Performance	Unsuccess of treatment (0)	Success of treatment (1)
MSE	0.092835478	0.110601021
NMSE	0.378803726	0.451293836
MAE	0.14313612	0.191653959
Min Abs Error	0.002563409	0.005851654
Max Abs Error	1.055555499	1.055555556
R	0.789058201	0.789058201
Percent Correct	89.23076923	91.83673469

The prediction accuracy of this hybrid approach of combined use of ANN and RST is around 90%. These comparison results of three different data mining applications for predicting the success rate of IVF treatments are shown in Table 5 and Fig. 5.

TABLE 5. COMPARISON BETWEEN THREE DIFFERENT DATA MINING APPLICATIONS

	Rough Set	ANN	Rough Set & ANN (Hybrid)
Percentage of Accuracy in Estimating Success	47	73	90

The application of combined Rough Set and Artificial Neural Network yields better result when compared with other techniques. It is observed that the hybrid technique of combined use of two or more machine learning tool yields better results than the use of a single technique for mining information from the database.

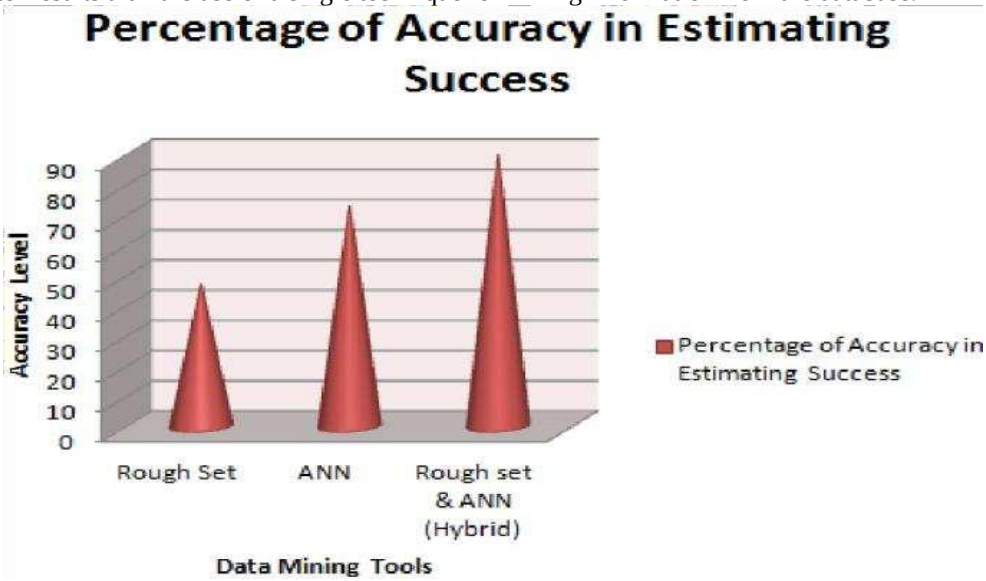


Fig. 5. The Success rate of Rough Set, ANN and Hybrid

VII. BENEFITS OF DATA MINING IN HEALTHCARE:

The use of data mining in the medical field is becoming more and more extensive due to the invaluable advantages it provides. The most notable examples of these advantages are described below.

A. Diagnosis accuracy increase

Having large amounts of historical data gives healthcare institutions the ability to assist their doctors with diagnosis in complicated cases. Of course, a "machine" should not replace doctors, but they can complete each other. For example, if a doctor studies a patient's social network page and reads comments to find the source or early signs of an illness, it may be considered as unethical behavior or a conflict of interests. However, nobody really cares when a data mining solution does exactly the same. If a patient ever complained about feeling bad and specified what exactly hurt, the software can add this information to the patient's history.

B. More effective treatment

Data mining can have a significant impact on the quality of treatment by giving doctors the additional information on patients' genetic heritage, previous cases of sickness or disease, shifts in activity and lifestyle, and a lot more. Such information may be categorized and stored in specially designed customer relationship management systems. Also, if a healthcare institution has legal access to certain databases that store data on previous treatment of a patient, the system can give doctors accurate information on which medications worked and which ones failed or made a person feel even worse.

C. Enhanced fraud detection:-

There will always be people who want to get prescription drugs without needing them, and data mining can be the tool for effective fraud prevention. All analytical and machine learning systems designed to detect fraudulent activity are completely useless without significant amounts of data. Thus, by having enough information on a new or returning patient, the software can tell a doctor whether a person truly needs a drug or tries getting it illegally. Thanks to such initiatives as the Healthcare Fraud Prevention Partnership (HFPP), doctors hope to reduce the number of fraudulent cases and give help only to those who really need it.

D. *Access to predictive analytics*

1) The COVID-19 pandemic is a vivid example of how important predictive analytics can be for the healthcare industry and the entire humanity. Data mining is the first stage of getting good-quality predictive software. Without significant amounts of information, there will be nothing to cluster and analyze in order to get predictive statistics on a certain topic for a certain period of time.In fact, information on epidemics, ecology, and more can be taken from trusted open sources from around the world like news, scientific journals, official reports, etc. That means that, in addition to the national database of electronic health records, there is a need for other extensive data sets with supplementary information to get a full picture of the current state of citizens' health and the healthcare system within a country. For example, a high-quality system that uses mined information can predict the percentage of cancer patients that will be in 2 years from now, and if it increases, a medical institution can start producing and testing new drugs, train more staff, consult with various healthcare organizations, etc.

E. *Better resource and management optimization*

2) Data mining and analytics go side by side, and a system that has plenty of information for analysis can significantly improve resource management of any healthcare institution. For example, a system can give advice on purchasing special equipment, hiring more doctors and medical staff of certain specialties.Also, the system can tell which drugs and medical procedures prove to be ineffective and can be replaced or removed with more effective ones. Pharmaceutical producers can use data mining applications to adjust and improve their medication or equipment development, test trials, and see what products and services will be the most demanded in the near and distant future.

F. *Drug quality assessment*

3) When a healthcare company produces medication or medical equipment, it is vital to be aware of even the slightest flaws of the product. In this case, data mining implementation by the company is highly important because the respective specialists can get more information on the product's impact on people's health.Sometimes people who agree to test drugs may intentionally or unintentionally hide information that is very important for the drug quality assessment. In contrast, software that has accurate data on current and past health conditions, genetic hereditary, living conditions, and the like can provide specialists with an extensive image of the entire drug creation process, thus helping to create medicine that works and won't turn into lawsuits.

G. Disaster prevention

4) Disasters in healthcare can be local (such as confidentiality breach or interrelated deaths within one medical facility) or global (such as pandemics). Depending on the goal and software complexity, it can provide the managerial staff of a healthcare company with predictions that can be used to improve weak spots; thus, solving the problem before it occurs. For example, a healthcare facility can detect future equipment malfunction or point on certain problems among the medical staff (stealing, slight but repetitive wrong diagnostics, etc.).

H. *Optimal health insurance price policy*

5) Health insurance has always been a big challenge for healthcare companies and their clients. When using data mining, insurance providers can have an extensive picture of the health condition of the person who applies for insurance. Of course, to get desirable health insurance for less money, people can hide information on their serious illnesses, real income, etc. Data mining can eliminate or at least significantly reduce losses of insurance companies by giving the analytical system information legally collected from various sufficiently trusted sources.

VIII. CONCLUSION:-

This paper aimed to compare the different data mining application in the healthcare sector for extracting useful information.The prediction of diseases using data mining apllications is a challenging task but it drastically reduces the human effort and increases the diagnostic accuracy.Developing efficient data mining tools for an application could reduce the cost and time constraint in terms of human resources and expertise. Exploring knowledge from the medical data is such a risk task as the data found are noisy, irrelevant and massive too. In this scenario, data mining tools come in handy in exploring of knowledge of the medical data and it is quite interesting. It is observed from this study that a combination of more than one data mining techniques than a single technique for diagnosing or predicting diseases in healthcare sector could yield more promising results. The comparison study shows the interesting results that data mining techniques in all the health care applications give a more encouraging level of accuracy like 97.77% for cancer prediction and around 70% for estimating the success rate of IVF treatment.

REFERENCES

[1] HianChyeKoh and Gerald Tan, ―Data Mining Applications in Healthcare‖, journal of Healthcare Information Management – Vol 19, No 2.

[2] JayanthiRanjan, ―Applications of data mining techniques in pharmaceutical industry‖, Journal of Theoretical and Applied Technology, (2007).

[3] RubanD.Canlas Jr., MSIT., MBA , ― Data mining in Healthcare: Current applications and issues‖.

[4] K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, ―Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks‖ International Journal on Computer Science and Engineering (2010).

[5] ShwetaKharya, ―Using Data Mining Techniques ForDiagnosis And Prognosis Of Cancer Disease‖, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.
[6] EliasLemuye, ―Hiv Status Predictive Modeling Using Data Mining Technology‖.
[7] Arvind Sharma and P.C. Gupta ―Predicting the Number of Blood Donors through their Age and Blood
[8] Group using DataMining Tool‖ International Journal of Communication and Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.
[9] Arun K Punjari, ―Data Mining Techniques‖, Universities (India) Press Private Limited, 2006.
[10] Margaret H.Dunham, ―Data Mining Introductory and Advanced Topics‖, Pearson Education (Singapore) Pte.Ltd.,India. 2005.
[11] PrasannaDesikan,Kuo-Wei Hsu, JaideepSrivastava, ―Data Mining For Healthcare
[12] Management‖, 2011SIAM International Conference on Data Mining, April, 2011
[13] Mobile Data Mining for Intelligent Healthcare Support
[14] ShusakuTsumoto and Shoji Hirano,―Temporal Data Mining in Hospital Information Systems‖.
[15] David Page and Mark Craven, ―Biological
N. AdityaSundar, P. PushpaLatha and M. Rama Chandra, ―Performance Analysis of Classification Data
[16] Mining Techniques Over Heart Disease Data Base‖, International Journal of Engineering Science & Advanced Technology, (2012).
[17] HardikManiya, Mosin I. Hasan and Komal P. Patel, ―Comparative study of Naïve Bayes Classifier and
[18] KNN for Tuberculosis‖, International Conference on Web Services Computing (ICWSC) 2011
[19] Proceedings published by International Journal of Computer Applications® (IJCA).
[20] Andrew Kusiak , Bradley Dixonb and ShitalShaha, ―Predicting survival time for kidney dialysis patients: a data mining approach‖, Computers in Biology and Medicine 35 (2005) 311–327.
[21] B.Renuka Devi, Dr.K.NageswaraRao, Dr.S.PallamSetty and Dr.M.NagabhushanaRao,‖ Disaster Prediction System Using IBM SPSS Data Mining Tool‖, International Journal of Engineering Trends and
[22] Technology (IJETT) - Volume4 Issue8- August 2013ISSN: 2231.
[23] Leah Passmore, Julie Goodside, Lutz Hamel, LilianaGonzalez, T Ali Silberstein And James Trimarchi, Assesing Decision Tree Models For Clinical In-Vitro Fertilization Data‖, Technical Report TR03-296
[24] SaangyongUhmn, Dong-Hoi Kim , Jin Kim , Sung Won Cho and Jae Youn Cheong, ―Chronic Hepatitis Classification using SNP data and Data Mining Techniques‖, Frontiers in the Convergence of Bioscience and Information Technologies 2007.
[25] Durairaj, K.Meena, ―A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks‖, International Journal Of Innovative Technology & Creative Engineering (ISSN: 2045- 8711) VOL.1 NO.7 JULY 2011.
[26] R. Srinivasan, Health care in India – Vision 2020 Issues and Prospects.

INTELLIGENT SERVICE ORIENTED ARCHITECTURE (SOA) FOR STATE-OF-THE-ART IOT-DDOS DEFENSE AND RESEARCH CHALLENGES

Manish Snehi[#1], Abhinav Bhandari[#2]
[#]*Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India*
[1]snehi.manish@outlook.com
[2]abhinavbhandari@pbi.ac.in

ABSTRACT— The Internet of Things (IoT) is rapidly surfacing that spans all spheres of life, and the applications of IoT devices have reaped enormous advantages. However, IoT devices are less secure due to limited computational capabilities, making them easy for malware for various attacks. Distributed Denial of Service (DDoS) attacks focused on the Internet of Things (IoT) are among the most devastating attacks that have gained attention as the number of intelligent devices is ever increasing. While several security protocols and defensive solutions have been developed to improve the security scenario in IoT networks, the frameworks are tightly coupled. The heterogeneity of network traffic is frequently increasing, making it challenging to keep the defense solutions updated for zero-day attacks. Additionally, the tight coupling makes scaling, upgrading, and migrating defense systems challenging. The state-of-the-art solutions cannot address the increased need for on-demand services and security concerns inherent in today's Internet. This article conducted an in-depth investigation of the research gaps in IoT-based DDoS defense solutions and presented a novel service-oriented architecture (SOA) for security frameworks that can be supplied to tenants and customers of tenants to mitigate these attacks. Service-oriented architecture aims to group application components into a network of loosely connected services to construct flexible, dynamic business processes and agile applications that transcend enterprises and computer platforms. The proposed conceptual model of Security SOA uses a Software-Defined Network (SDN)-based security mechanism to identify and mitigate DDoS attacks in the Internet of Things (IoT) networks. Programmable networks facilitate the implementation of security as service interfaces for defensive systems.

KEYWORDS— Distributed Denial of Service, Internet of Things, Network Function Virtualization, Service Oriented Architecture, SOA, Software defined network.

INTRODUCTION

The Internet of Things marked an evolution in the technology industry. The amalgamation of “Internet” and “Things” has spawned a global network of interconnected computer networks. The IoT network utilizes the Internet protocol suite (TCP/IP) to provide services to billions of users worldwide. The IoT networks range from local to worldwide, connected by various electrical, wireless, and optical networking technologies. The presence of IoT in the sphere of a variety of application areas (as shown in Fig. 1) has uniquely addressed the communication of physical objects with varying degrees of processing, sensing, and actuation by connecting the physical objects to the cyber layer. These Intelligent physical devices interoperate and communicate via Internet as their common platforms. The high adaptability of IoT devices has led to an ecosystem in eclectic application realms. The following are a few examples of IoT application areas: Smart Agriculture, Smart Grids, Smart Cities. A report from Gartner forecasted the IoT revenue growth to 21 billion dollars by 2022, and the forecast for the number of connected devices by 2030 is 25.44 billion dollars [1] [2].

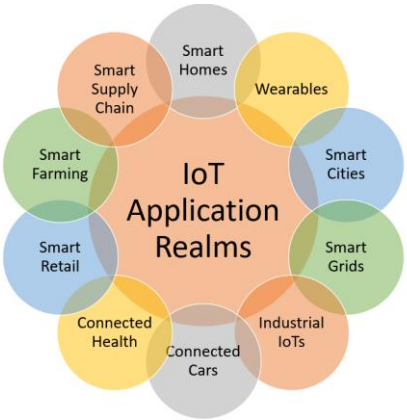


Fig. 1 - IoT Application Realms

Due to significant interest and evolution in IoT, futuristic IoT technology has various de facto standards are available for usage. The flip side of the coin is that the universally adapted IoT devices are specific purpose devices, less secure, and have limited computational capabilities. The dearth of adequate security on IoTs forms a slew of issues for network operators, and applications service providers of major enterprises. Implementing security protocols on the devices would undoubtedly help defend against automated attacks; the effectiveness varies between manufacturers, the communication protocols, computational capability, and device operation mode. However, the inefficient security policies for IoT devices

make the less secure and computationally weak IoT devices vulnerable to cyber-attacks [3], [4]. A report from a preeminent security firm, Imperva, says that IoT-based DDoS attacks are the most severe security affairs [5]. IoT-based DDoS attacks had gained the attacker's and researcher's attention when a series of IoT-based DDoS attacks employed the Mirai botnet [6]. The initial attacks hit unprecedented rates of 1.2 Tbps. The Command and Control (C&C) architecture make the IoT botnets (such as Mirai and Hajime) highly hazardous because of its capability to scan and infect the botnets to form the botnet zombie's army. Fig. 2 (a) and (b) presents the IoT-DDoS Attack stages and modus operandi for the attacks.

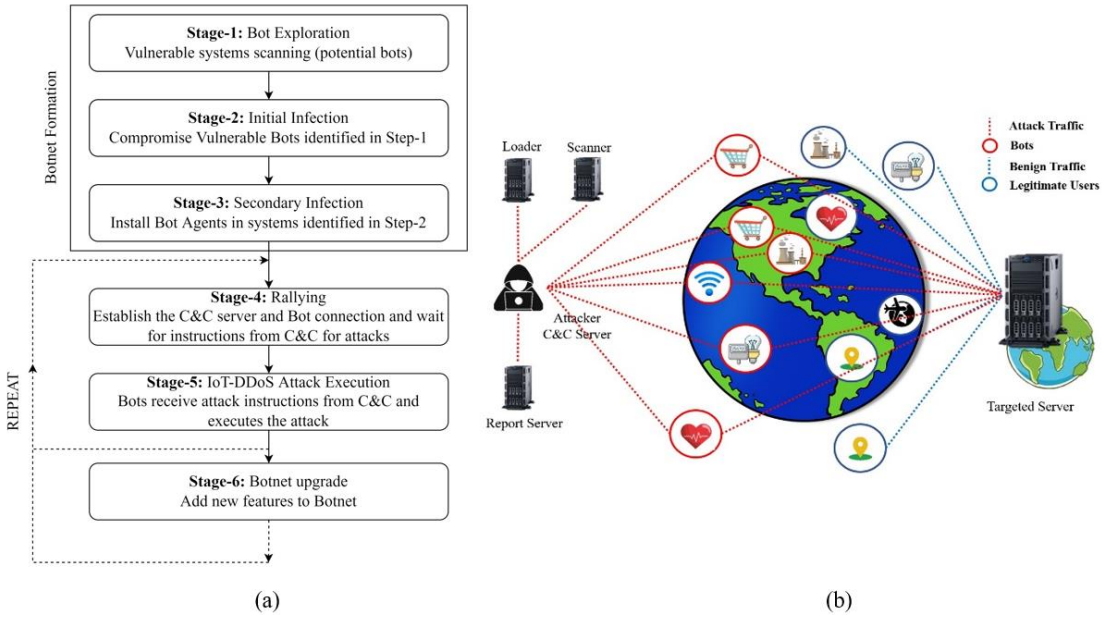


Fig. 2 – (a) IoT-DDoS Attack stages and (b) modus operandi

There have been significant efforts in developing defense solutions against IoT-based DDoS attacks. The state-of-the-art software-defined networks (SDN) have given a complimentary state of affairs to nourish IoT-DDoS solution development. SDN paradigm, by its nature, provides flow-level isolation and visibility in a low-cost and scalable manner. SDN offers programmable networks and has the global visibility of the overall network. The network is highly reconfigurable, flexible to fault-tolerant, and load changes. The companion technology "Network Function Virtualization (NFV)" offers scalable and autonomous architecture to the additional networked services. There have been significant research efforts in developing modern-day intelligent defense solutions by leveraging machine learning and deep learning models. However, the available solutions are inadequate for upgrades. Moreover, the solutions are tightly coupled, hard to upgrade, and have several limitations.

The article proposes service-oriented architecture (SOA) for the security solution and tosses the term "Security-as-a-Service" interface for IoT-DDoS defense service. The interface leverages Fog Computing as a defense layer to bring the security service out of the cloud to offer the reduced latent SOA.

PP. Paper Contributions

The paper has following contributions to the research community:

1. The paper presents the literature survey for the contemporary smart and intelligent IoT ecosystems, analyses the available defense solutions to presents the research gaps.
2. The paper puts the SDN and NFV technologies together to offers the novel Security-As-A-Service interface architecture for IoT-DDoS attacks defense solutions. The offered interface addresses the issues of tight coupling of defense solutions, promises to reduces the installation and maintenance cost. Furthermore, the offered security interface resolves the scalability and compatibility issues.
3. The article leverages the Fog Computing layer to bring the security interface out of the cloud to reduce the latency and provide the futuristic security interface.

QQ. Paper Organization

The paper organization for rest of the article is as follows: Section 0 presents the modern-day intelligent and smart IoT based DDoS solutions. Section 0 offers the analysis of the modern-day IoT-DDoS solutions and extracts the research gaps. Section 0 presents the contemporary security interface for futuristic defense solutions. Lastly, the paper concludes with the recommendations and prospects in section 0.

ASSOCIATED WORK

Recent researchers have made significant contributions towards the detection of IoT-based DDoS attacks. The early detection of IoT-based DDoS traffic is described in [7]. The general approach towards detecting IoT-based DDoS attacks is

to detect the attacks based on pattern matching and anomaly detection. The researchers in [8] have investigated the attacks using anomaly methods.

The authors of [9] established a methodology for detecting DDoS traffic originated by IoT devices. The model is built on the Deep Autoencoding technique. The experiment has demonstrated that it can detect 100% of DDoS traffic cases. The designed detection approach achieves promising results in detection accuracy for devices that are currently accessible. However, when applied to a new set of devices, the model's indeterminate behaviour is a key concern given the proliferation and variety of devices in the IoT ecosystem.

Research [10] is established exclusively on data compiled by emulation of the IoT devices in the controlled laboratory. The authors have leveraged the SDN capabilities to carry out DDoS detection at the edge of the network.

The researchers in [11] emphasize the boost in the effect of IoT devices on the vigour of DDoS network traffic. The authors of [12], [13] highlight the requirement to investigate the prospect of detecting IoT-based DDoS traffic.

Future studies should develop novel methods for identifying DDoS traffic generated by IoT devices. The IoT devices continue to grow in homes and businesses, as does their exploitation to create IoT botnets. Machine-to-Machine traffic generated by such devices has specific characteristics separate from human-generated traffic. Table XVII presents summary for the research dealing with the issue of IoT-based DDoS traffic detection systems in the IoT ecosystems.

TABLE XVIII
SUMMARY OF LITERATURE SURVEY

Ref.	Experimental Setup	DDoS Detection Level	ML / DL Model	Accuracy
[7]	Smart Homes	Edge Detection	KNN, SVM, Decision Tree, ANN	99%
[9]	Smart homes	Edge Detection	Deep Autoencoders	100%
[10]	IoT (General)	Edge Detection	Software-defined Networking (SDN)	Not specified
[14]	IoT (General)	Edge Detection	CNN, RNN	98%
[15]	Intrusion Detection System	KDD Cup'99 Dataset	Deep Neural Networks	99%
[16]	Simulation	SDN	SDN, Stacked Autoencoders	99.82%
[17]	In-house IoT Network Setup	Edge Detection	Artificial Neural Networks (ANN)	99.4%
[18]	Simulation	Fog Computing, NSL-KDD Dataset	Deep Learning based	99.27%

RESEARCH CHALLENGES

Researchers have made significant efforts to develop security solutions for Smart Applications. The developed solutions offer contemporary intelligent mitigation frameworks against security breaches. Large enterprises leverage trending modern-day technologies to migrate the physical infrastructure to offer the Anything-As-A-Service. The offered services provide on-demand, scalable, reliable, and efficient solutions at Cloud Computing. However, the defense solutions against the vicious cyber-attacks are still tightly coupled to the system, which spawn the following research challenges (As shown in Fig. 3):

RR. Tight Coupling

The defense solutions are tightly coupled to the host systems. In the ever-evolving technological era, technophiles employ contemporary technologies to breach into the systems. Hence, there is a constant need to upgrade systems with the latest defense policies. Tight coupling makes the upgrade process a nightmare.

SS. Less Cost Effective

The tightly coupled solutions are less cost-effective as the solution deployment includes the software and hardware cost. Moreover, any upgrades come with additional field visits, hardware, and software upgrade costs.

TT. Scalability Issues

The offered defense solutions are not scalable. Hence, computational upgrades require vertical scaling of the hardware infrastructure in memory and computational power. Tightly coupled solutions are hard to scale horizontally.

UU. Less Adoptability

The available defense solutions are delivered as an integrated solution and delivered as black boxes. The third parties involved in the defense solution development lack the standardization process. Hence, it requires high training due to a lack of standardization.

VV. On-demand not possible

As the solutions are available as an integrated piece of software, on-demand services are not possible. The end-users are forced to avail of all the services available in the software package. Furthermore, the lack of scalability is a bottleneck in on-demand expansion or computational load reduction.

WW. Compatibility Issues

The offered defense frameworks are decentralized, and upgrades require manual intervention. Hence, the upgrades due to compatibility issues are a nightmare.

Fig. 3 - Research Challenges in modern-day IoT-DDoS defense solutions

PROPOSED ARCHITECTURE AND PROSPECTS

A moderately novel strategy to security is built on the Security-as-a-Service (SECaaS) notions. Some enterprises chose to outsource the system security, and the demand underwent a need for such services. With the increasing demands of third-party security providers, Managed Security Services (MSS) appeared on the business stack. MSS is the aspect that directs the evolution of the SECaaS. As investigated in the literature survey, the coupling of security at the cloud can be decoupled and migrated to the Fog layer for more robust defense solutions. The network bandwidth is conserved as most of the computational activities are performed at Fog, and results are transferred to the cloud platform.

The Security-as-a-Service (SECaaS) interface is proposed at the Software-defined Networking enabled Fog Computing layer (as shown in Fig. 4). The Fog Computing layer (a.k.a. Cloud Edge) is responsible for latency reduction, network bandwidth conservation, and uninterrupted service to the cloud platform. The conceptual model of the SECaaS interface is proposed to be deployed at the distributed cluster of the SECaaS server. The SECaaS cluster hops onto a big-data technology stack for real-time network traffic processing. Furthermore, the core of the SECaaS implementation is the machine and deep learning algorithms that make the system intelligent.

The SDN controller queries the SECaaS interface for ingress new-flow entries to detect the IoT-DDoS attacks. The SDN controller mitigates the attack by executing appropriate security policies. SDN controller, further, transfers the legitimate traffic to the cloud.

The proposed service architecture is not limited to IoT-DDoS attacks but various attack scenarios. Enterprises can add more services and offer security services in an on-demand model. Moreover, the services can be charged per hit model (a.k.a. pay-as-you-go model) to comply with cloud fundamentals.

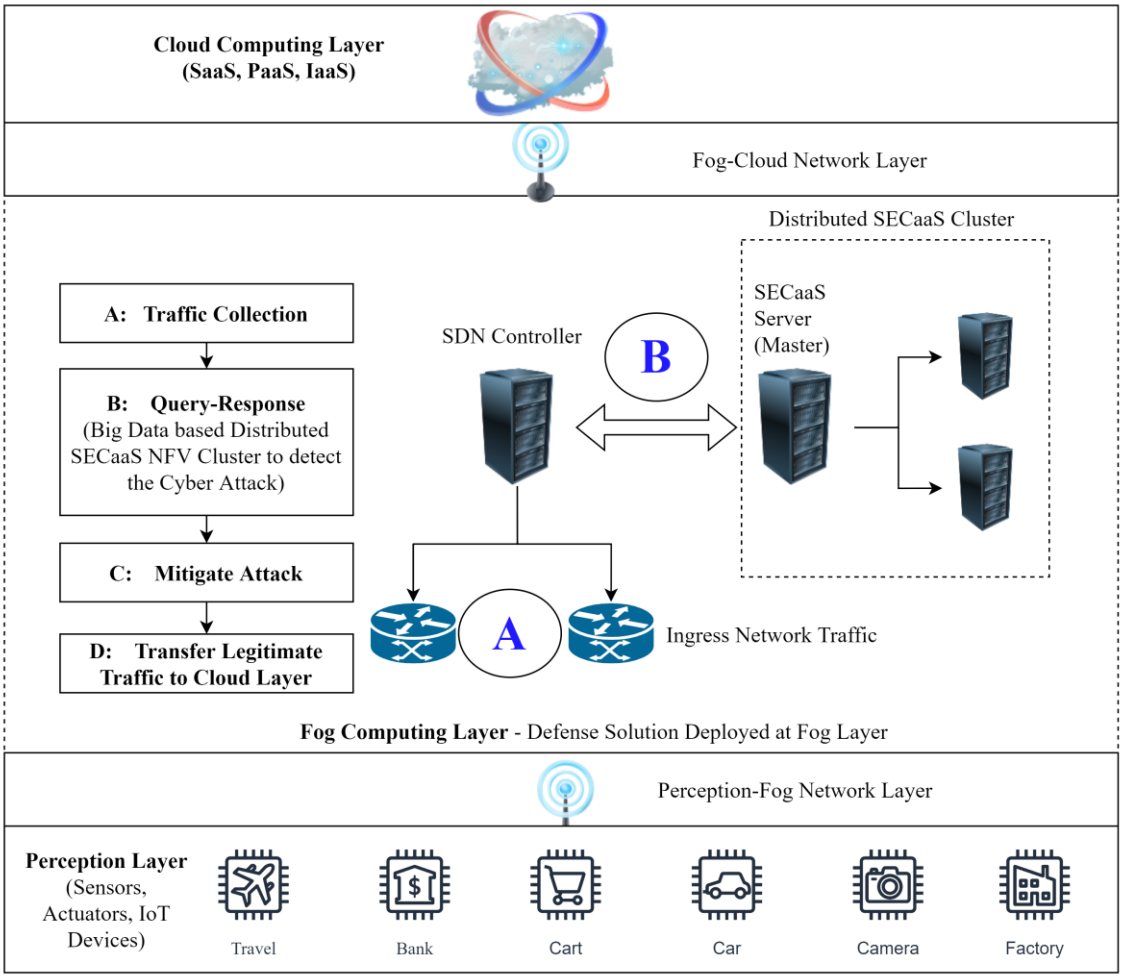


Fig. 4 – A Novel Fog-Layer-based Real-time Security-as-a-Service (SECaaS) interface for IoT-DDoS Defense

This logical segmentation of SOA architecture is motivated by the need to distinguish:

- fundamental service capabilities given by a middleware architecture and traditional SOA from more complex service capability required for dynamically constructing services
- professional services distinct from those focusing on systems, and
- Composition of services as a result of systems integration

CONCLUSION

The article reviews the literature on modern smart and intelligent IoT ecosystems and analyses available defense solutions to identify research gaps. Furthermore, the paper combines SDN and NFV technologies to present a novel Security-as-a-Service interface architecture for IoT-DDoS defense solutions. The proposed architecture makes the most of Fog's capabilities to offer an efficient security interface. It addresses the research gaps by (a) Decoupling security solutions from the host system, (b) offering a distributed and horizontally scalable approach, (c) offering on-demand security as a service, and (d) resolving the compatibility issues. The authors propose using intelligent machines and deep learning algorithms in the security interface to address zero-day attacks in the future work.

ACKNOWLEDGMENT

The authors would like to thank their mentors, colleagues, and the esteemed journal for publishing the research.

REFERENCES

[1] STAMFORD, “Gartner Says Global Government IoT Revenue for Endpoint Electronics and Communications to Total $21 Billion in 2022,” 2021. https://www.gartner.com/en/newsroom/press-releases/2021-06-30-gartner-global-government-iot-revenue-for-endpoint-electronics-and-communications-to-total-us-dollars-21-billion-in-2022 (accessed Nov. 22, 2021).

[2] M. Snehi and A. Bhandari, “Vulnerability retrospection of security solutions for software-defined Cyber–Physical System against DDoS and IoT-DDoS attacks,” *Computer Science Review*, vol. 40, p. 100371, May 2021, doi: 10.1016/j.cosrev.2021.100371.

[3] J. Snehi, A. Bhandari, M. Snehi, U. Tandon, and V. Baggan, “Global Intrusion Detection Environments and Platform for Anomaly-Based Intrusion Detection Systems,” 2021, pp. 817–831.

[4] J. Verma, A. Bhandari, and G. Singh, "A Meta-analysis of Role of Network Intrusion Detection Systems in Confronting Network Attacks," in *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2021, pp. 506–511, doi: 10.1109/INDIACom51348.2021.00090.

[5] Imperva.com, "Q2 2017 Global DDoS Threat Landscape," 2017. https://www.imperva.com/resources/resource-library/reports/q2-2017-global-ddos-threat-landscape (accessed Nov. 20, 2021).

[6] M. Snehi and A. Bhandari, "Apprehending Mirai Botnet Philosophy and Smart Learning Models for IoT-DDoS Detection," in *2021 8th International Conference on Computing for Sustainable Global Development (INDIACom)*, 2021, pp. 501–505, doi: 10.1109/INDIACom51348.2021.00089.

[7] M. Ozcelik, N. Chalabianloo, and G. Gur, "Software-Defined Edge Defense Against IoT-Based DDoS," in *2017 IEEE International Conference on Computer and Information Technology (CIT)*, Aug. 2017, pp. 308–313, doi: 10.1109/CIT.2017.61.

[8] M. H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita, "Network Anomaly Detection: Methods, Systems and Tools," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 303–336, 2014, doi: 10.1109/SURV.2013.052213.00046.

[9] A. Sivanathan *et al.*, "Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics," *IEEE Transactions on Mobile Computing*, vol. 18, no. 8, pp. 1745–1759, Aug. 2019, doi: 10.1109/TMC.2018.2866249.

[10] M. Lopez-Martin, B. Carro, A. Sanchez-Esguevillas, and J. Lloret, "Network Traffic Classifier With Convolutional and Recurrent Neural Networks for Internet of Things," *IEEE Access*, vol. 5, pp. 18042–18050, 2017, doi: 10.1109/ACCESS.2017.2747560.

[11] T. Sciences, "ANALYSIS OF THE IoT IMPACT ON VOLUME OF DDoS ATTACKS," *XXXIII Simpozijum o novim tehnologijama u poštanskom i telekomunikacionom saobraćaju – PosTel 2015*, 2015.

[12] J. Costa Gondim, R. de Oliveira Albuquerque, A. Clayton Alves Nascimento, L. García Villalba, and T.-H. Kim, "A Methodological Approach for Assessing Amplified Reflection Distributed Denial of Service on the Internet of Things," *Sensors*, vol. 16, no. 11, p. 1855, Nov. 2016, doi: 10.3390/s16111855.

[13] D. H. Summerville, K. M. Zach, and Y. Chen, "Ultra-lightweight deep packet anomaly detection for Internet of Things devices," in *2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC)*, Dec. 2015, pp. 1–8, doi: 10.1109/PCCC.2015.7410342.

[14] Y. Meidan *et al.*, "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," *IEEE Pervasive Computing*, vol. 17, no. 3, pp. 12–22, Jul. 2018, doi: 10.1109/MPRV.2018.03367731.

[15] A. Divekar, M. Parekh, V. Savla, R. Mishra, and M. Shirole, "Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives," *Proceedings on 2018 IEEE 3rd International Conference on Computing, Communication and Security, ICCCS 2018*, pp. 1–8, 2018, doi: 10.1109/CCCS.2018.8586840.

[16] R. M. A. Ujjan, Z. Pervez, K. Dahal, A. K. Bashir, R. Mumtaz, and J. González, "Towards sFlow and adaptive polling sampling for deep learning based DDoS detection in SDN," *Future Generation Computer Systems*, vol. 111, pp. 763–779, 2020, doi: 10.1016/j.future.2019.10.015.

[17] B. B. Gupta, R. C. Joshi, and M. Misra, "ANN based scheme to predict number of zombies in a DDoS attack," *International Journal of Network Security*, vol. 14, no. 2, pp. 61–70, 2012.

[18] A. Abeshu and N. Chilamkurti, "Deep Learning: The Frontier for Distributed Attack Detection in Fog-To-Things Computing," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 169–175, 2018, doi: 10.1109/MCOM.2018.1700332.

ENVISIONING INTELLIGENT NIDS: FEATURE ENGINEERING TECHNIQUES FOR PRE-PROCESSING OF THE REAL-TIME NETWORK TRAFFIC DATA

Jyoti Verma[#1], Abhinav Bhandari[#2], Gurpreet Singh[#3]
[#1]Deptt. of Computer Science and Engineering
Punjabi University Patiala, India,
[#2]Deptt. of Computer Science and Engineering
Punjabi University Patiala, India,
[#3]Deptt. of Computer Science and Engineering
Punjab Institute of Technology Rajpura, India
1jyoti.snehiverma@gmail.com
2bhandarinitj@gmail.com
3myselfgurpreet@gmail.coom

ABSTRACT— With the global adoption of cloud-based services increasing at an exponential rate, the complexity, dimensions, and possibility of attacks on cloud-based networks have increased as well. Distributed security breaches in the area of Internet of Things (IoT) networks have exposed the inefficiency of presently offered network intrusion detection systems (NIDS). Computer security is reliant on intrusion detection systems to a large extent. Proposed research dedicated to advancing NIDS techniques has developed in the context of network disruptions. Numerous learning-based techniques can improve their detection accuracy by removing correlated, recurring, and irrelevant features. According to previous research, the researchers have strived to increase classification results on NIDS datasets by incorporating Machine Learning (ML), Dimensionality Reduction (DR), and Deep Learning (DL) techniques. NIDS datasets, on the other hand, vary in terms of their selected features, design, and attack types. As such, this article examines the breadth of these techniques using a wide range of datasets. The purpose of this paper is to expedite a survey of existing frameworks by utilizing benchmark datasets to determine unconventional attacks and shed insight into the current condition of NIDS. This article summarises extensive research on identifying various types of attacks and the issues that arise as a result of them using machine learning classifiers and deep learning algorithms. Additionally, this paper evaluates the performance of existing NIDS by classifying attribute features using machine learning algorithms methods. The authors discussed feature selection techniques for impeding anomaly detection in this article, as well as how to pre-process authentic internet traffic data and envisage virtualized real-time Intelligent NIDS utilizing feature engineering techniques.

KEYWORDS— Deep Learning, DL, Machine Learning ML, Feature Engineering, Dimensionality reduction, Feature Selection

INTRODUCTION

Cloud computing is the fastest-growing field in the Information Technology sector, owing to its low expenses, ease of use, and resource efficiency [1][2]. Cloud computing makes use of technologies such as computing services, grid computing, and virtualization. Outsider attacks frequently expose Spoofing, Address Resolution Protocol spoofing, Routing Information Protocol attacks, and nefarious network traffic injection. Network infrastructure security measures such as firewalls are more efficient at defending against a broad range of outsider attacks. It is incapable of defending against both insider and sophisticated outsider attacks. Utilizing intrusion detection systems is advantageous. It automates the detection of intrusions [3]. IDS continuously monitors network, system, and host activity after producing and sending reports to a centralized unit or system administrator. Cloud computing's size, complexity, and vulnerability have increased. The existing Network Intrusion Detection System (NIDS) is incapable of detecting real-time attacks caused by breaches in IoT network security. Cloud security is critical for establishing user trust. Cloud-based NIDS are depicted in Figure 1. Due to NIDS inadequacies and the loss of sensitive data, a new area of research is being committed to improving NIDS technologies [4]. Numerous machine learning techniques can benefit from the elimination of superfluous, repetitive, and meaningless attributes. The majority of researchers attempted to improve classification results on NIDS datasets by combining several Dimensionality Reduction, Machine Learning, and Deep Learning techniques. These datasets have a distinct set of features, intrusion types, and network architectures. The rigor of these techniques is evaluated in this article using a variety of datasets. There are multiple sorts of intrusion detection systems available, and this article explains how they start comparing and how they can be improved. This section discusses the use of machine learning classifiers and deep learning algorithms to detect various types of attacks and issues that arise as a result of them. Additionally, this study makes use of machine learning and deep learning techniques to assess the effectiveness of existing NIDS systems. This article discusses the feature selection for network traffic-based anomaly detection, and also how to pre-process real-time network traffic information [5].

LITERATURE REVIEW

NIDS monitor and analyze network traffic in detecting and preventing security issues and intrusions. As a consequence, developing a successful NIDS requires considerable expertise in high-performance computing as well as the acquisition and analysis of a large amounts number of datasets. The NIDS models have been validated using a variety of publicly

available datasets. The KDD99, ISCX2012, NSL-KDD, UNSW-NB15, CICIDS2017, KYOTO 2006+, and CICDDoS2019 datasets are all publicly available and can be used for NIDS. The training/testing datasets can be created from the same data set. Attacks are categorized into four classes in the dataset. Denial of service (DOS) attacks happen when an unauthorized user is turned down access to the system or connections, such as a computer or network. R2L attacks occur when attackers attempt to compromise a user account or through another host. A U2R attack occurs when an attacker tries to log in with a constrained user account to gain root access. A probe occurs when an attacker searches for unlabeled data or vulnerabilities on a computer or network.

Table 1: Summary of Literature Survey on various NIDS

References	Feature Learning	ML/DL Algorithm	Data sets	Performance
[1]	UFL	STL, Soft-max regression	NSL-KDD	Accuracy=98%
[2]	ZSL	DT	KDD Cup 99	Accuracy=99.94%
[4]	Wrapper approach, GA, LR	C4.5, RF, and NBTree.	KDD Cup 99 and the UNSW-NB15	Accuracy=99.91% with KDD99 and 94.65% with the UNSW-NB15 dataset.
[5]	Random undersampling.	SMOTE	KDD Cup 99, UNSW-NB15, and NSL-KDD	-
[6]	Filter based Approach	DT,RFHT and KNN	KDD Cup 99 and UNSW-NB15	Accuracy=99.8393 with KDD 99
[7]	MDAE	LSTM	NSL-KDD, UNSW-NB15, and CICIDS 2017	Accuracy=94 %
[8]	Long Short Term Memory	SVDD	KDD99	Accuracy=98.0% and 99.8%, respectively.
[9]	Bayesian corsets	BLR, SVM	CICIDS2017	
[10]	Rule-based	Xgboost and lightgbm	QiAnXin DataCon datasets and GeekPwn	Acceptable accuracy
[11]	Hyperparameters tuning	FNN, LSTM	KDD Cup 99, temporally-correlated-attacks dataset	Accuracy of FNN= 88±1% LSTM= 99.54±0.03%
[12]	Spatial and temporal features	CNN+RNN Neural network, LuNet	NSLKDD and UNSW-NB15	Accuracy= 82.78%
[13]	Autoencoder	DNNs	CICIDS 2017 and ISCX IDS 2012	False-Positive Rate of 0.00013
[14]	Feature embedding	RNN,LSTM	UNSW-NB15	Binary classification accuracy of 99.72%
[15]	Principle Component Analysis	Naïve Bayes SVM,RF, ANN	KDD Cup 99 and the NSLKDD	Average f1-score of 0.9452
[16]	CART	RF	KDD Cup 99 and UNSW-NB15	Accuracy= 99.97%
[17]	Maximum and minimum normalization method	PSO,Xgboost	NSL-KDD	PSO-Xgboost is 13% higher than RF
[18]	SMOTE	DNN and CNN	NSL-KDD, UNSW-NB15, and Bot-IoT	Accuracy= 99.45%
[19]	PSO,GWO, FFA ,GA	SVM and J48	UNSW-NB15	Accuracy= 79.077%–90.119%
[20]	IG	SMOTE DBN	KDD CUP 99, NSL–KDD	Accuracy=90.02, 988.4 % resp.

Abbreviations:
Random Forests (RF);Self-taught Learning (STL);Decision trees (DT);Hidden naïve Bayes (HNB); Hoeffding Tree (HT),K-Nearest Neighbors (KNN); Feedforward neural network (FNN); Support Vector Data Description (SVDD);Bayesian logistic regression (BLR);Support Vector Machine (SVM); Long-short term memory (LSTM);Deep Neural Networks (DNNs);Convolutional Neural Networks(CNN);Particle Swarm optimization (PSO);Energy-based Flow Classifier (EFC);Deep Belief Network (DBN);Unsupervised Feature Learning (UFL);Zero-Shot learning (ZSL);Genetic Algorithm (GA); Logistic Regression (LR) ;Multimodal deep auto encoder (MDAE);Classification and Regression Trees (CART);Grey wolf optimizer (GWO);Firefly optimization (FFA);Genetic algorithm (GA);Information gain (IG)

In 2016, researchers concentrated on host-based intrusion detection systems and envisaged NIDS based on older datasets such as NSL-KDD and KDD99, as well as supervised machine learning algorithms. The researchers who analyzed big data using Network Log and Memory Log techniques proposed a system that incorporates feature engineering and classification, as well as their integration with Apache Spark's big data cloud computing. In 2017, the majority of researchers recommended deep learning for NIDS, with a particular emphasis on signature-based NIDS. No real-world attack scenarios have been tested for NIDS, according to the literature. Studies have been scarce into cloud services and efficient data mining concepts. The literature survey revealed that auto-encoder and multilayer perceptron were the most

commonly used models. The researchers proposed an enhanced model based on the backpropagation algorithm and stochastic gradient descent methods for artificial neural networks. Additionally, NIDS was proposed for cloud use to recognize DoS attacks that degrade resource availability [6]. In 2018, researchers shifted their focus away from cloud computing and big data and toward anomaly-based NIDS solutions, shifting away from supervised popular machine learning algorithms and toward unsupervised popular machine learning algorithms. Canonical coefficient of determination and logistic regression analysis, for example, are both techniques for reducing the size of features in the methodologies they suggested. Before 2019, researchers concentrated on NIDS based on signatures and anomalies [7]. In 2019, the focus moved to machine learning algorithms that could be used to detect both types of NIDS. Researchers enhanced the efficiency of network intrusion detection systems using advanced cybersecurity data analysis research, with a particular emphasis on big data handling and deep learning techniques. By 2020, a few researchers have effectively applied an artificial intelligence-based intrusion detection system, utilizing both Supervised and Unsupervised learning techniques. The researchers proposed Convolutional neural networks and Long-term memory networks in their experiments. Numerous researchers have used data mining algorithms such as clustering algorithms and Apriori, as well as machine learning/deep learning methods on cloud-based NIDS Solutions, to solve problems in 2021 [8].

ENVISIONING INTELLIGENT NIDS

A NIDS supervises both outbound and inbound network activity to detect unexpected intrusion attempts. Because solution infrastructure and services are virtualized and modern IT organizations are trying to migrate to the cloud, an efficient NIDS remedy is required to defend the cloud from a variety of intrusions. Due to the large volume of raw data generated by high-dimensional data sources, it is difficult for the researcher of deep learning methods to collaborate with high-dimensional cloud computing. Figure 1 illustrates the various steps necessary to create an Intelligent NIDS. The research will address a critical issue in the context of cloud computing by addressing the following: the creation of a framework for distributed intelligent intrusion detection systems.

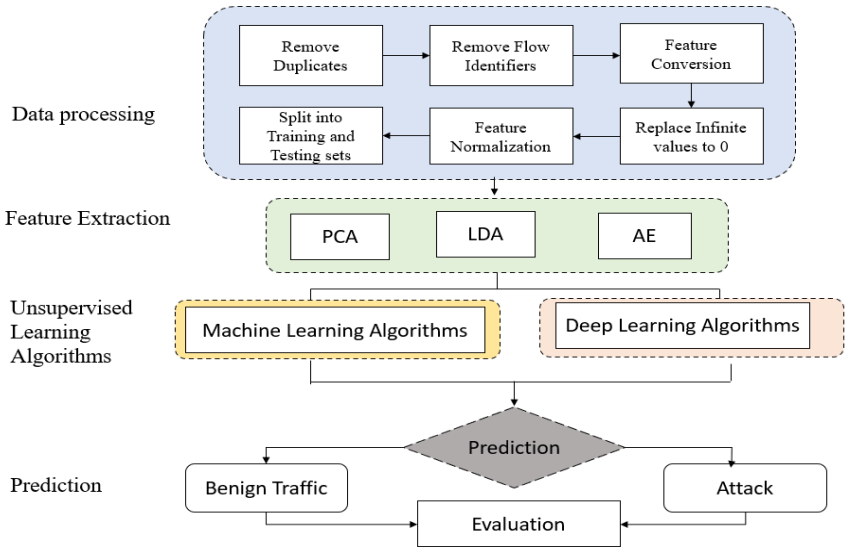


Figure 1: Envisioning Intelligent NIDS

The procedure consists of multiple phases or phases. The first step is to eliminate attack samples from the training data set to prepare the autoencoder exclusively on normal samples. The data is then mapped and scaled using upper and lower limits from the training set. Feature extraction from the input data increases the accuracy of the model. By removing redundant data, this phase of the framework reduces the data dimension. By combining and converting the original feature set, features are extracted. The inputs are given the training to be as close to the training set's log patterns as possible. The difference between the input and output of an input sample dictates its attack classification. Validation data is composed of benign and intrusion logs during the training phase. The Area Under Curve (AUC) score is used to evaluate the model's performance. Throughout the training phase, the weights of the better model and the AUC score are recorded. Following training, the framework weights with optimizing scores are used. The distance between the input and output of an attack is used to ascertain whether the traffic is benign or not. The trained model's result is evaluated to that of the testing data by ranging the threshold level and assessing the quality of ML/DL -NIDS on testing data. Model evaluation can be performed by contrasting predicted labels to the initial labels in test data (i.e., accuracy, True positive rate, False positive rate, and MCC, among others). In machine learning, feature extraction is critical because it has a significant impact on prediction accuracy [10].

A. Feature Engineering Techniques for Pre-processing of the Real-time network traffic data

By reducing the number of input features, data reduction enhances computational efficiency and accuracy. This simplifies the model and makes it easier to interpret for both the machine learning algorithm. The first stage in machine learning is data pre-processing. It helps clean, format, and arrange raw data in preparation for the development of machine learning

models. Pre-processing information for learning algorithms requires expertise in both data science and feature engineering. The term "data engineering" refers to the process of converting unstructured data into structured data. The data is then fine-tuned using feature engineering to produce the features anticipated by the insightful learning method. The process of repurposing existing features paves the way for feature engineering. Feature engineering is the process of evaluating various techniques on a variety of datasets to determine their impact on model performance. The stages in Data Pre-processing are depicted in Figure 1. Importing the necessary libraries is the first step. NIDS are capable of analyzing data from a wide variety of datasets. Delete the entire column or substitute some critical data for a dataset's missing value. Choose a training set and inspect it for missing values. Due to the model's and framework's reliance on numerical calculations and equations, categorical variables must be transformed into numerical data. Following transformation, the set of data is partitioned into training and testing segments. Before testing the predictive accuracy of the learning model on the test dataset, the learning method attempts to apprehend any correlations in the training data. The dataset is divided into 80 and 20 percent training and testing set, respectively. Feature scaling is a term used in machine learning to refer to the process of expanding features to a certain scale. The dataset includes variables with a range of possible values. To facilitate processing, all nominal data points are discretized. Both labeled and unlabeled variables are acquired from the same origin during implementation. Classification of attributes from a set of data that contribute most to the target variable or output, either automatically or manually. This is an important element in identifying network intrusions and determining the types of attacks. The purpose of feature selection is to reduce the size of the dataset while raising detection performance. Principal Component Analysis (PCA), Auto-encoder, and Linear Discriminant Analysis (LDA) techniques are used for dimensionality reduction [26].

1. Principal Component Analysis

It is a technique for reducing the dimensionality of a dataset by generating additional statistically independent variables with the greatest variance. By extracting a subset of features that adequately describes the dataset, the number of parameters in the feature space can be reduced. After standardizing the data, it extracts the covariance matrix's eigenvectors and eigenvalues. It seeks out a lower-dimensional subset of features that adequately describes the data [26][27].

2. Linear Discriminant Analysis

It overtures to model differences between samples assigned to groups by utilizing location and covariance estimators. The method's objective is to maximize the ratio of the variance between groups to variance within groups. While this ratio reaches the maximum value, the data points within each group exhibit minimal scatter and the gatherings exhibit the greatest separation. LDA overtures to model differences between samples assigned to groups by utilizing location and covariance estimators. The method's objective is to maximize the ratio of the variance between groups to variance within groups. Whenever the ratio reaches its maximum value, the samples inside every group exhibit the least dispersion and the groups exhibit the greatest separation [28].

3. Auto Encoder

It is a special type of artificial neural network capable of understanding to compressing data. They are unsupervised learning techniques that make use of neural network models to learn representations. Autoencoders are given training as an input reconstruction part of a larger model. The lower threshold of data besides training the neural network to recognize the most salient features of an image. Autoencoders are regularly used as a model for learning or automatically extracting features. Autoencoders can be trained to complexly depict data in higher dimensions [29][30].

B. Performance of NIDS using ML and DL Techniques

Machine Learning is a AI term that refers to all techniques and algorithms that facilitate computers to learn automatically through the use of statistical equations to extract knowledge from large datasets. It incorporates Decision Trees, KNNs, SVMs, FNNs, and ANNs into its architecture. Deep Learning is a subfield of machine learning in which numerous hidden layers are used to simulate a deep network. Due to their sophisticated structure and capacity to self-significant points from a dataset, these techniques perform better machine learning. CNN, LSTM, RNN Networks, and stacked auto-encoders are some of the techniques that are frequently used. NIDS benchmarking quantifies the properties of NIDS through the use of a variety of metrics. The evaluation criteria are intended to ensure a fair assessment of the framework's efficiency. Confusion matrixes are a widely used technique for labeling effects. The uncertainty matrix forms the basis for several different metric measurements. A confusion matrix summarises a classifier's actual and predicted classifications and contains information about the actual and estimated classifications. Accuracy Metrics can be used to evaluate the performance of models. The various components of the uncertainty matrix are True Positive, True Negative, False Positive, and False Negative.[31] Model fitting quantifies the generalizability of machine learning algorithms to similar data. Accurate results derived from a well-fitted method model fitting are central to machine learning. If the model does not fit the data correctly, the predictions will be inaccurate and useless. The evaluation of models is a critical step in the development of machine learning models. It assists in determining the optimal model to depict the dataset and the model's performance [32].

DISCUSSION

The NIDS's success is highly dependent on the datasets chosen by the researchers. A significant dataset is used to train machine learning models, thereby increasing their accuracy. ML is not well suited for large datasets unless they are labeled, which is costly and time-consuming. While machine learning has long been used in network intrusion detection,

these techniques proceed to be afflicted by a lack of labeled data, high operational costs, and low accuracy. DL methods are advantageous for large datasets because they facilitate the discovery of patterns in raw data. To effectively detect zero-day attacks, NIDS must be updated regularly with the new dataset. Due to the large set of data and deep learning algorithms, the computational and time prerequisites of the learning process increase. The NIDS framework can become more efficient in detecting intrusions as it is trained. However, due to their efficiency in attempting to learn from large raw datasets, deep learning-based NIDS methods have gained popularity and acceptance. In Figure 2, a review of the literature reveals that the majority of proposed solutions test models using older datasets, such as Knowledge discovery Cup'99 and NSL-KDD. The performance of some proposed solutions is lower for newer datasets than for older sets of data. Most techniques are inefficient at detecting attacks when the training data contains fewer samples. All such minority attack classes face an imbalance of power that must be addressed. In the last six years, researchers have concentrated their efforts on developing deep learning tools for designing NIDS systems. Prominently, half of the proposed techniques are DL-only, 15% are hybrids that combine machine learning and deep learning methodologies, and only 35% are ML-only. Deep learning models are complex and require a substantial amount of computational power. With the introduction of the GPU, DL-based techniques have increased in popularity for NIDS design. The most frequently used deep learning methods are AE, DNN, CNN, and RNN. Methods based on machine learning, such as RF and SVM, aid and improve algorithms based on deep learning. DT, KNN, and Naive Bayes are less frequently used machine learning algorithms. To propose NIDS solutions, autoencoders and their different versions are used. This reduces the complexity of the model and, consequently, the training time. Several of the methods proposed made use of various deep learning algorithms. These techniques increased detection accuracy at the expense of increased complexity and computational resources. Additionally, data may require formatting before use by an algorithm. A one-dimensional variable must be converted into a two-dimensional matrix before CNN can be used for NIDS. The most frequently used performance metrics are Detection Accuracy and Recall. A higher Accuracy and Detection rate is considered for network security. Accuracy, recall, and F-measure should be required performance metrics for a typical machine learning/deep learning-based NIDS to detect intrusions. 70% of the time, the NSL-KDD, KDD Cup'99, and UNSW-15 were used. Despite their age, researchers have begun to use these sets of data due to their breadth of coverage. Historically, network architecture was quite distinct. User privacy threats to big data and sensor networks are rapidly evolving. In practice, a model trained and validated using the most recent data outshines older data-based models or systems. The suggested Cloud-based real-time intelligent NIDS identifies all of the issues raised above and utilizes feature engineering techniques to pre-process real-time network traffic data. The graphs in Figure 2 are based on a review of the literature on various NIDS.

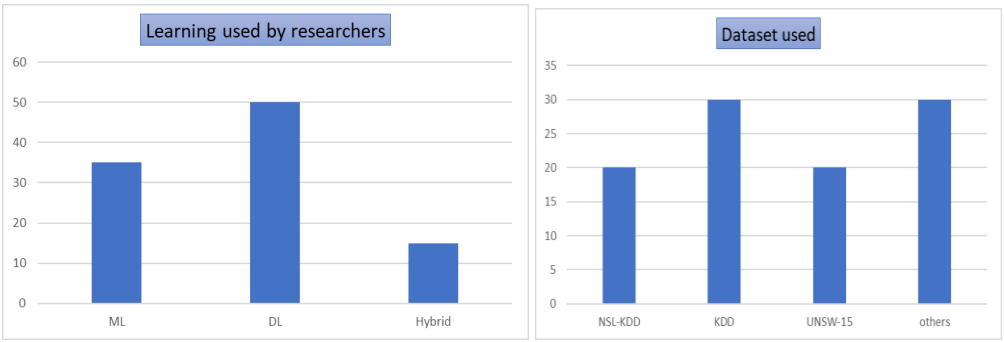


Figure 2: Graph-based on Literature Survey on various NIDS

CONCLUSIONS

This paper examines the performance of the existing NIDS that employ Machine Learning/Deep Learning strategies for attribute feature extraction and classification. This article discusses feature selection models and techniques for network traffic anomaly detection. The purpose of this research is to suggest a method for pre-processing real-time network traffic data using feature engineering techniques. We considered combining various dimensionality reduction, machine learning, and deep learning methods on NIDS datasets for research. The primary objective of this research is to propose a Virtualized real-time intelligent NIDS on the dataset based on feature engineering to predict the framework having the highest precision and accuracy with the fewest false positives.

REFERENCES

[1] H. F. Atlam, R. J. Walters, and G. B. Wills, "Fog computing and the internet of things: A review," *Big Data Cogn. Comput.*, vol. 2, no. 2, pp. 1–18, 2018, doi: 10.3390/bdcc2020010.

[2] S. Jyoti, S. Manish, and G. Rupali, "Virtualization as an Engine to Drive Cloud," *High Perform. Archit. Grid Comput.*, no. 09988701479, pp. 62–66, 2011, doi: https://doi.org/10.1007/978-3-642-22577-2_9.

[3] F. A. Khan and A. Gumaei, *A Comparative Study of Machine Learning Classifiers for Network Intrusion Detection*, vol. 11633 LNCS. Springer International Publishing, 2019.

[4] P. Wu and H. Guo, "LuNet: A Deep Neural Network for Network Intrusion Detection," *2019 IEEE Symp. Ser. Comput. Intell. SSCI 2019*, pp. 617–624, 2019, doi: 10.1109/SSCI44817.2019.9003126.

[5] N. Thanh Van, T. N. Thinh, and L. T. Sach, "a Combination of Temporal Sequence Learning and Data Description for Anomalybased Nids," *Int. J. Netw. Secur. Its Appl.*, vol. 11, no. 03, pp. 89–100, 2019, doi: 10.5121/ijnsa.2019.11307.

[6] M. Snehi and A. Bhandari, "Vulnerability retrospection of security solutions for software-defined Cyber–Physical System against DDoS and IoT-DDoS attacks," *Comput. Sci. Rev.*, vol. 40, p. 100371, May 2021, doi: 10.1016/j.cosrev.2021.100371.

[7] H. Azwar, M. Murtaz, M. Siddique, and S. Rehman, "Intrusion Detection in secure network for Cybersecurity systems using Machine Learning and Data Mining," *2018 IEEE 5th Int. Conf. Eng. Technol. Appl. Sci. ICETAS 2018*, pp. 1–9, 2019, doi: 10.1109/ICETAS.2018.8629197.

[8] A. Singhal, "Intrusion Detection Systems," *Adv. Inf. Secur.*, vol. 31, pp. 43–57, 2007, doi: 10.4018/978-1-59904-168-1.ch007.

[9] Q. Niyaz, W. Sun, A. Y. Javaid, and M. Alam, "A deep learning approach for network intrusion detection system," *EAI Int. Conf. Bio-inspired Inf. Commun. Technol.*, 2015, doi: 10.4108/eai.3-12-2015.2262516.

[10] J. L. R. Pérez and B. Ribeiro, "Attribute learning for network intrusion detection," *Adv. Intell. Syst. Comput.*, vol. 529, pp. 39–49, 2017, doi: 10.1007/978-3-319-47898-2_5.

[11] H. A. Mahmood, "Network Intrusion Detection System (NIDS) in Cloud Environment based on Hidden Naïve Bayes Multiclass Classifier," *Al-Mustansiriyah J. Sci.*, vol. 28, no. 2, p. 134, 2018, doi: 10.23851/mjs.v28i2.508.

[12] C. Khammassi and S. Krichen, "A GA-LR wrapper approach for feature selection in network intrusion detection," *Comput. Secur.*, vol. 70, pp. 255–277, 2017, doi: 10.1016/j.cose.2017.06.005.

[13] A. Divekar, M. Parekh, V. Savla, R. Mishra, and M. Shirole, "Benchmarking datasets for Anomaly-based Network Intrusion Detection: KDD CUP 99 alternatives," *Proc. 2018 IEEE 3rd Int. Conf. Comput. Commun. Secur. ICCCS 2018*, pp. 1–8, 2018, doi: 10.1109/CCCS.2018.8586840.

[14] H. He, X. Sun, H. He, G. Zhao, L. He, and J. Ren, "A Novel Multimodal-Sequential Approach Based on Multi-View Features for Network Intrusion Detection," *IEEE Access*, vol. 7, pp. 183207–183221, 2019, doi: 10.1109/ACCESS.2019.2959131.

[15] F. M. Zennaro, "Analyzing and Storing Network Intrusion Detection Data Using Bayesian Coresets: A Preliminary Study in Offline and Streaming Settings," *Commun. Comput. Inf. Sci.*, vol. 1168 CCIS, pp. 208–222, 2020, doi: 10.1007/978-3-030-43887-6_16.

[16] S. Lu *et al.*, "New Era of Deeplearning-Based Malware Intrusion Detection: The Malware Detection and Prediction Based On Deep Learning," pp. 1–30, 2019, [Online]. Available: http://arxiv.org/abs/1907.08356.

[17] J. Gao *et al.*, "Omni SCADA Intrusion Detection Using Deep Learning Algorithms," *IEEE Internet Things J.*, pp. 1–1, 2020, doi: 10.1109/jiot.2020.3009180.

[18] G. C. Fernandez and S. Xu, "A Case Study on using Deep Learning for Network Intrusion Detection," *Proc. - IEEE Mil. Commun. Conf. MILCOM*, vol. 2019-Novem, no. i, 2019, doi: 10.1109/MILCOM47813.2019.9020824.

[19] H. Gwon, C. Lee, R. Keum, and H. Choi, "Network Intrusion Detection based on LSTM and Feature Embedding," 2019, [Online]. Available: http://arxiv.org/abs/1911.11552.

[20] S. Sapre, P. Ahmadi, and K. Islam, "A Robust Comparison of the KDDCup99 and NSL-KDD IoT Network Intrusion Detection Datasets Through Various Machine Learning Algorithms," 2019, [Online]. Available: http://arxiv.org/abs/1912.13204.

[21] Z. Chkirbene, S. Eltanbouly, M. Bashendy, N. Alnaimi, and A. Erbad, "Hybrid Machine Learning for Network Anomaly Intrusion Detection," *2020 IEEE Int. Conf. Informatics, IoT, Enabling Technol. ICIoT 2020*, pp. 163–170, 2020, doi: 10.1109/ICIoT48696.2020.9089575.

[22] H. Jiang, Z. He, G. Ye, and H. Zhang, "Network Intrusion Detection Based on PSO-Xgboost Model," *IEEE Access*, vol. 8, pp. 58392–58401, 2020, doi: 10.1109/ACCESS.2020.2982418.

[23] M. Mulyanto, M. Faisal, S. W. Prakosa, and J. S. Leu, "Effectiveness of focal loss for minority classification in network intrusion detection systems," *Symmetry (Basel).*, vol. 13, no. 1, pp. 1–16, 2021, doi: 10.3390/sym13010004.

[24] O. Almomani, "SS symmetry Detection System Based on PSO , GWO , FFA and," 2020.

[25] H. Jia, J. Liu, M. Zhang, X. He, and W. Sun, "Network intrusion detection based on IE-DBN model," *Comput. Commun.*, vol. 178, no. January, pp. 131–140, 2021, doi: 10.1016/j.comcom.2021.07.016.

[26] B. V. Snehi J., Bhandari A., Snehi M., Tandon U., "Global Intrusion Detection Environments and Platform for Anomaly-Based Intrusion Detection Systems," *Springer, Singapore*, 2021, doi: DOI https://doi.org/10.1007/978-981-16-0733-2_58.

[27] M. Panda, A. Abraham, and M. R. Patra, "A hybrid intelligent approach for network intrusion detection," *Procedia Eng.*, vol. 30, no. 2011, pp. 1–9, 2012, doi: 10.1016/j.proeng.2012.01.827.

[28] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.

[29] Z. Ahmad, "Network intrusion detection system : A systematic study of machine learning and deep learning approaches," no. August, pp. 1–29, 2020, doi: 10.1002/ett.4150.

[30] R. K. Rahul, T. Anjali, V. K. Menon, and K. P. Soman, "Deep Learning for Network Flow Analysis and Malware Classification," *Commun. Comput. Inf. Sci.*, vol. 746, no. November, pp. 226–235, 2017, doi: 10.1007/978-981-10-6898-0_19.

[31] Y. Xin *et al.*, "Machine Learning and Deep Learning Methods for Cybersecurity," *IEEE Access*, vol. 6, no. c, pp. 35365–35381, 2018, doi: 10.1109/ACCESS.2018.2836950.

[32] J. Snehi, A. Bhandari, V. Baggan, and M. Snehi, "Diverse Methods for Signature based Intrusion Detection Schemes Adopted," no. 2, pp. 44–49, 2020, doi: 10.35940/ijrte.A2791.079220.

METEOROLOGICAL PREDICTIONS USING DIGITAL IMAGE PROCESSING: RESEARCH CHALLENGES AND KEY OPPORTUNITIES

Rhythm Naswa, Dr. Navdeep Kanwal
Punjabi University
rhythmnaswa51@gmail.com
navdeepkanwal@gmail.com

ABSTRACT— India is a country of varied seasons and different parts of the country faces innumerable severe weather phenomena such as Cyclones, Tsunami, Thunderstorms, Hailstorms, Heavy rains, Snowfall, Fog etc. in different parts of the year. Meteorology is a science that deals with atmosphere and its phenomena including weather and climate.India Meteorological Department provides different meteorological services in India such as Cyclone Warning, City Forecast, Highway Forecast, Tourism Forecast, Issuance of severe weather warning for heavy rains, hail, floods, snowfall, fog etc. With recent advancements in technology trends, satellites and Doppler weather radars have proved to be important instruments in sensing and scanning the atmosphere to provide various weather-related information. In this paper, we tend to study Meteorological Predictions using Digital image processing in satellite and Doppler weather radar images and the Research Challenges and Key opportunities associated with it. Various thresholding, segmentation, edge detection techniques can be used on satellite and DWR images to detect different meteorological features say for example detecting a cyclone, center of eye of the cyclone, its track and landfall point, and classification of cyclone based on impact, differentiating between rain, hail, and snow etc. Loss of human life and property due to a lot of natural disasters such floods, cyclones, hailstorms etc. can be prevented with accurate and on-time weather warnings. The paper is aimed at using digital image processing algorithms on satellite and radar images to make accurate meteorological predictions which could be used for betterment of the people of the nation by minimizing the loss of life and property in case of occurrence of natural disasters.

KEYWORDS— Meteorology, satellite images, radar images, digital image processing, weather prediction

I. INTRODUCTION

Conventionally, Meteorology is the branch of science which studies the atmosphere of our planet; its structure, composition, and properties; the physical processes closely related to the Earth's surface, water, and air; various weather phenomena; weather and climate; and the future state of the atmosphere (Lindzen et al. 1990; Ackerman and Knox 2007; Ahrens and Henson 2016). It can also be defined as the study of various atmospheric elements such as temperature, air pressure, relative humidity, wind direction and speed, cloud cover etc. in order to make predictions relating to weather and climate. This is an extremely crucial field of science as it prepares a city for any kind of extreme weather conditions and thus prevent disasters.India, being a country with diverse demographic features experiences various meteorological events such as Floods, Droughts, Hailstorms, Thunderstorms, Snow, Cyclones, Heavy Rain, Fog etc. in its different parts. The role of meteorology becomes crucial in accurate and timely prediction of the occurrence of these weather events and issuing forecasts and nowcasts to the general public. The Government of India, under the vision of Honorable Prime Minister Shri Narendra Modi formed National Smart Cities Mission in 2015, with an approach to develop smart cities in India by making them citizen friendly for quality living and sustainable in all ways. The Smart Cities can be abbreviated as Sustainable Management Action Resource Tools for Cities [1]. The concept of smart city is incomplete without an aspect of safe, smart and sustainable environment for the citizens wherein "smart meteorology" shall play a key role [2]. Climatology of a city is a collective data of its past 30-50 years of its weather events. Meteorology deals with both weather and climate, therefore, where weather prediction models are useful for the future of smart cities, past climatology of the city is useful to design and develop urban features in a smart city. India Meteorological Department provides climatological data for 100 such smart cities shortlisted in India at its official website [3].

II. NEED FOR RESEARCH IN METEOROLOGY

Meteorology is a multidisciplinary field of science that directly affects human life in a number of ways. Firstly, precise and well-timed weather predictions assist in leading a sustainable life and preventing disasters such as Cyclones, Snowstorms, Thunderstorms etc. thus saving human life and property. Secondly, Agro-meteorology is another important sub area under meteorology that largely affects the population of country like India where Agriculture is the primary source of livelihood for about 58% population, the share of agriculture in GDP is 19.9 % (2020-21) and where rainfall is the primary source of water for irrigating the fields. Due to the topological features of India, the rains here are not just seasonal but irregular, thus, very commonly being affected by "Floods" and "Droughts". Thirdly, Aviation Meteorology and Marine Meteorology play a key role in smooth plying of air and sea transport that is affected by weather events such as Cumulonimbus Clouds, Reduced visibility due to Fog, Cyclones in the coastal areas etc. In addition to above, Meteorology serves a prospective domain for researcher for example, Flash Flood warnings, Monitoring of Stubble burning, Monitoring and prediction of Ozone and Air Quality levels, Tourism Forecasts, Pilgrimage Forecasts, Power grid failure warnings during severe thunderstorms and many more.

III. EXISTING WEATHER SATELLITE AND RADAR PRODUCTS IN METEOROLOGICAL PREDICTIONS

Weather satellites are man-made remote sensing equipment that revolves around the Earth to monitor and detect weather and climate. It could be in polar or geostationary orbits.Presently, the three meteorological satellites in the geosynchronous orbit are Kalpana-1, INSAT-3A and INSAT-3D. [4] These satellites provide huge volumes of data to facilitate weather forecasting, rainfall estimation etc.; to detect cloud top temperatures, cloud motion, water vapour content in the atmosphere, formation of cyclones and its track. These satellites also facilitate reception and transmission of meteorological data from in-situ instruments placed across vast and remote areas. The three main types of satellite image products are [5] (a) **Visible Imagery**- formed by visible rays of the spectrum and this visible imagery can be viewed during day time as the clouds reflect light from the sun. Useful in detecting formation of thunderstorms at earlier stages. (b) **Infrared imagery**- In Infrared satellite images, the clouds can be seen in day as well as night because these satellites does not use sunlight, rather the clouds are detected by sensors from the heat radiated from them. Clouds can be easily identified as they are colder than land and water. These image products are useful in detecting the intensity of thunderstorms on the basis of cloud top temperatures and identifying fog and low clouds during night. (c) **Water Vapor imagery**– this satellite image product indicates the content of moisture present in upper atmosphere and useful in identifying where heavy rains and thunderstorm development is possible.

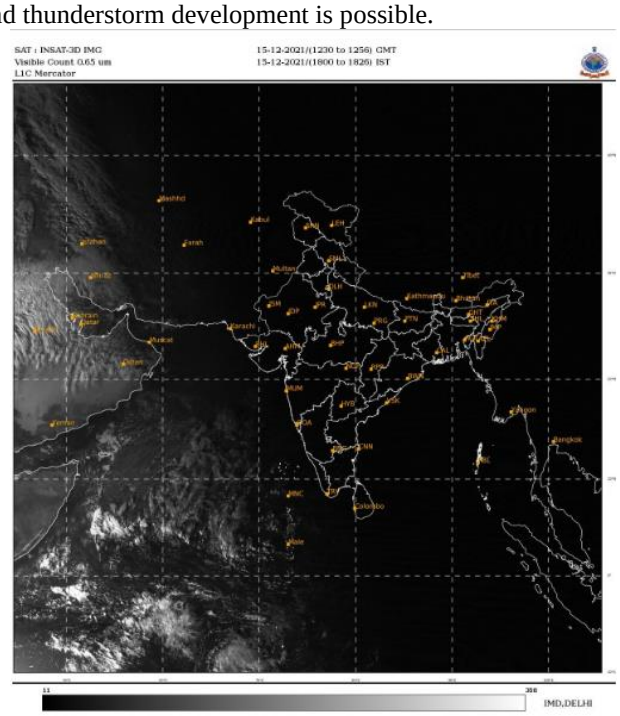


Fig. 1 A sample image of Visible Imagery of INSAT 3D Satellite

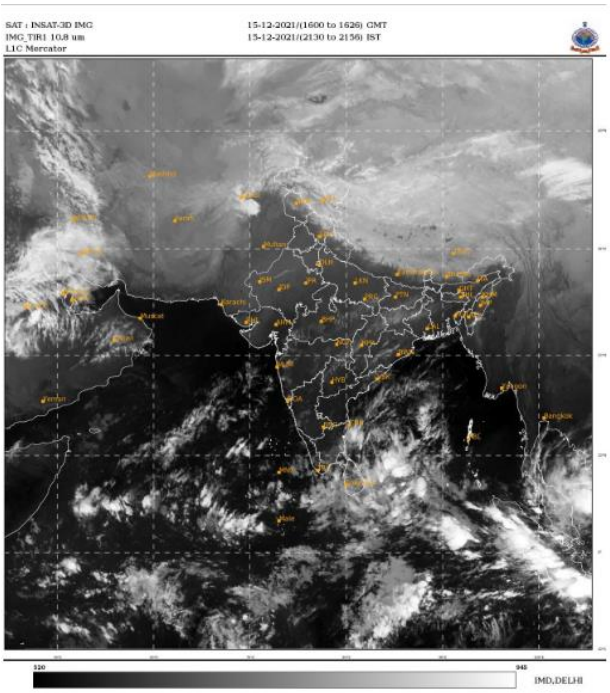


Fig. 2 A sample of Infrared Imagery of INSAT 3D Satellite

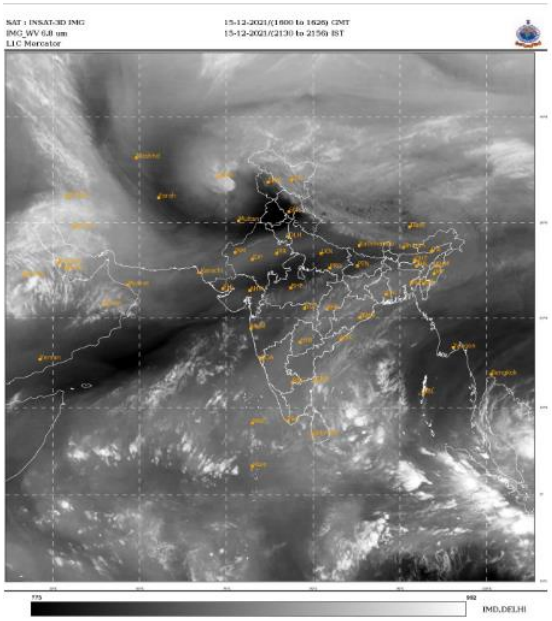


Fig. 3 A sample of Water Vapour Imagery of INSAT 3D Satellite

Doppler Weather Radar is another indispensable remote sensing equipment used to scan the atmosphere for clouds and precipitation. [6]. It is useful in differentiating the detected particles to be rain, hail, snow or some insect. The data from weather radars is used in detecting the structure of storm, its intensity and direction of motion. Various Base DWR products are (a) **Reflectivity** – this product is prepared by the energy reflected back from the atmosphere and indicates the content of water. Useful in identifying the precipitation particle to be rain or hail and also the reflectivity values indicate the hail size; (b) **Velocity** – this base product provides the details of cloud motion towards the radar or away from it and with what speed. Various products derived from these base radar products are useful for rainfall estimation, aviation, issuing severe weather warning etc.

IV. DIGITAL IMAGE PROCESSING IN METEOROLOGY

Quick visualization and analysis of data and products provides accurate weather assessments.Digital Image Processing on satellite and radar images helps in automatic and quick analysis of the displayed information. Development of thunderstorm monitoring [7] can be done using data from both satellite and radar images. Use of K-Means algorithm on satellite visible range images is done to identify different types of clouds [8]. Use of Deep learning techniques on satellite images for nowcasting has been seen as an efficient step [9]. An author developed a machine learning framework to provide reliable flood forecasts by assimilating satellite observed precipitation in to hydrologic models [10]. The researchers have also worked upon detecting formation of Tropical cyclones using edge detection and feature extraction techniques [11]. Work has also been done in addressing the problem of semiautomatic detection of center of cyclone using region detection and pattern matching in Synthetic Aperture Radar images [12]. Another important piece of work has been done by the author to identify split and merge processes in convective systems to propose radar based centroid tracking algorithm for severe weather surveillance [13].

V. RESEARCH CHALLENGES

Firstly, the accuracy of weather predictions depends greatly upon the observational data collected from different parts of the country. The observational data can be collected manually, or through Automatic Weather Stations, Automatic Rain Gauges, Radiosondes/ Pilot Balloon instruments. This observed data is the input to Numerical Weather Prediction Models for processing and thereby making weather predictions and issuing severe weather warnings. Neither manual nor automatically gathered observations are immune to errors, thus making it a challenge.

Secondly, as has lately been observed, some meteorological events occur and diminish in a smaller time frame such as flash floods, landslides, hailstorms. Therefore, the dissemination of warnings for such events is time critical to save human lives and public property. Creating algorithms and weather prediction models with lesser runtime to issue timely warnings to QRT teamsis another meteorological research challenge.

Thirdly, availability of authentic and appropriate datasets is pivotal factor in quality of the research work. In India, skilled manpower and state-of-the-art equipment costing hundreds of Crores of rupees are deployed for collecting, processing and archiving weather and climate data by the India Meteorological Department, the central, authentic Government agency operating in India since 1875 [14]. Limited access to the authentic source of meteorological data makes it a research challenge.

Lastly, certain disasters such as Earthquakes and Tsunami cannot be predicted in advance but cause havoc in the event of occurrence. Development of an approach to predict such events using machine learning, IoT and video surveillance etc. to analyze and classify the consequential impact areas poses a critical research challenge and a crucial one to take step forward towards building smart and sustainable cities.

VI. KEY OPPORTUNITIES

Meteorology is a wide domain with numerous research opportunities for budding researchers. Crucial sectors such as Disaster Management, Agriculture, Aviation, Tourism, Transport and many more that greatly affect the lives and property of the nation. Digital Image Processing on satellite and radar images opens the research opportunities not just in detecting severe weather events such as cyclones, thunderstorms, fog, heavy rains, snow or hail but also, offers a great source of information in detecting stubble burning, direction of locust attack, predicting flash floods etc.

REFERENCES

[1] S. M. a. R. Ramaswamy, "The State of Art: Smart Cities in India: A Literature Review Report," *International Journal of Innovative Research and Development,* pp. 115-119, 2013.

[2] M. K. S. M. G. J. M. Harish Kumar, "Moving towards smart cities: Solutions that lead to the Smart City Transformation Framework," *Technological Forecasting and Social Change,* p. 119281, 2020.

[3] I. M. Department, "Climatology of Smart Cities," 20 November 2021. [Online]. Available: https://mausam.imd.gov.in/imd_latest/contents/index_smart_cities1.php.

[4] "RAPID: Gateway to Indian Weather Satellite Data - ISRO," 20 November 2021. [Online]. Available: https://www.isro.gov.in/rapid-gateway-to-indian-weather-satellite-data.

[5] U. D. o. Commerce, "Three types of satellite imagery," 20 November 2021. [Online]. Available: https://www.weather.gov/mrx/sattype.

[6] M. R. Kumjian, "Weather Radars," *Remote Sensing of Clouds and Precipitation* , pp. 15-63, 2018.

[7] V. A. S. A. S. B. Aida A. Adzhieva, "Development of thunderstorm monitoring technologies and algorithms by integration of radar, sensors, and satellite images," in *Remote Sensing of Clouds and the Atmosphere XXII*, Warsaw, Poland, 2017.

[8] B. G. G. B. Sanjay Goswami, "Hexalevel Grayscale Imaging and K-Means Clustering to Identify Cloud Types in Satellite Visible Range Images," in *Proceedings of International Conference on Computational Intelligence and Computing*, Jhansi, India, 2021.

[9] G. C. E. M. Vlad-Sebastian Ionescu, "DeePS at: A deep learning model for prediction of satellite images for nowcasting purposes," *Procedia Computer Science*, pp. 622-631, 2021.

[10] A. L. S. N. V. R. Roderick Lammers, "Prediction models for urban flood evolution for satellite remote sensing," *Journal of Hydrology*, p. 127175, 2021.

[11] M. D. JV Bibal Benifa, "Recognizing Tropical Cyclone Formation from Satellite Image Data," *Artificial Intelligence and IoT: Smart Convergence for Eco-friendly Topography*, pp. 131-150, 2021.

[12] S. Jin, S. Wang, X. Li, L. Jiao, J. A. Zhang and D. Shen, "A Salient Region Detection and Pattern Matching-Based Algorithm for Center Detection of a Partially Covered Tropical Cyclone in a SAR Image," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 280-291, 2017.

[13] T. R. M. C. L. Annadel Moral, "A radar-based centroid tracking algorithm for severe weather surveillance: identifying split/merge processes in convective systems," *Atmospheric Research*, pp. 110-120, 2018.

[14] I. M. Department, "History of Meteorological Services in India," 21 November 2021. [Online]. Available: https://mausam.imd.gov.in/imd_latest/contents/history.php.

A REVIEW on COMPUTER VISION APPLICATION in FARM ANIMAL MANAGEMENT

Navdeep Singh[#1], Charanjiv Singh Saroa[#2]
[#] *Department of Computer Science and Engineering, Punjabi University Patiala, Patiala*
[1] writetonavdeepsingh@gmail.com
[2] charanjiv_saroa@yahoo.com

ABSTRACT— In today's era, dairy farming is increasingly opting for a modern form in which the health of dairy animals can be improved and well maintained, farmer's incomes can be increased and their living standards raised. Meanwhile, computer vision techniques are being progressively applied in the dairy industry to pinpoint the health status of dairy animals. Technologies that emerge under computer vision can also help to protect various dairy animals from discomfort. Automated veterinary diagnostic tools can help animals diagnose diseases and their causes and identify appropriate treatments and respond accordingly. Machine learning is a key driver of productivity and efficiency in various industries. Therefore, the animal health sector is no different. In the coming years, more and more technology will be used in the agricultural animal health sector. Exploring data through machine learning can help us form norms and make decisions about the future.

KEYWORDS– Dairy Animal, Computer Vision Technique, Dairy Industry, Automated Veterinary Diagnostic Tools, Machine Learning.

I. INTRODUCTION

Computer vision technology can be used to accurately determine the behavioral activity and health status of farm animals. This review covers technologies for developing an automated system for modern dairy industries. With the increase in population, the need for products like milk, eggs and meat has also increased. Majority of the population in India is dependent on livestock. Farm animals have a very close and deep relationship with humans. To meet this demand, the animals are stuffed in small spaces and unhygienic conditions, thereby spreading various diseases and infections. Animal health systems ensure that farm animals are healthy, disease free and well cared for. Technologies developed under computer vision could help protect farm animals from disease. Animals like cow, buffalo, goat and sheep play an important role in dairy farming. Automated veterinary diagnostic tools can assist animals in diagnosing diseases and their causes and responding accordingly by recommending appropriate treatments. Machine learning could help veterinarians treat animal diseases without the risk of infection. With the help of technology, veterinarians can perform surgery with stitches and provide fracture treatment with very high accuracy. It can also help in treatment with very high accuracy. It can also help treat cancer by operating on infected parts of animals. Machine learning is a major driver of productivity and efficiency across all industries and the animal health industry is no different. Machine learning has many roles to play aiding in disease prevention, diagnosis and treatment. More and more technology will be applied in the livestock health sector in the coming years. Apart from their contribution to human power, machines have also come to contribute to the brain power of humanity. These methods help us to make assumptions about the future and help us to make decisions by analyzing huge amount of data for any application. Machine learning is a new method of application in veterinary medicine and has recently received a lot of attention and Very little effort has been done in this extent. This review summarizes the methods of machine learning.

II. SYSTEM REVIEW

Machine Learning Techniques

Machine learning is a sub-set of artificial intelligence that delivers solutions to many challenging problems such as image, video and voice recognition by learning from patterns. Machine learning consists of various technologies for programming computers. In machine learning, we outline a model with various variables, and a computer program is trained to enhance the parameters of that model, either through image, audio, or video data, or through training such as certain measured data or past experiences. As shown in Figure 1, data in the form of image, audio, video or any measurement data is used as input for feature extraction / data segmentation. After that, preprocessing is applied to smooth the data using filters. Trained model under machine learning is used to get the desired and meaningful information in the form of output.

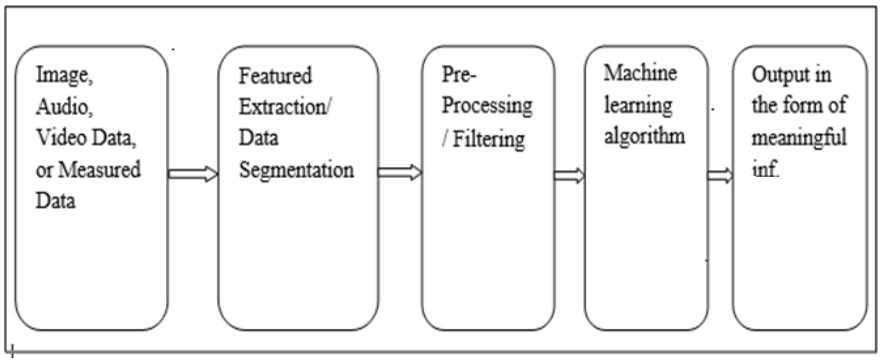


Figure: 1 General steps of machine learning

Machine learning uses two methods: Learning under supervision. This comprises training the model with known inputs and outputs values to forecast imminent results. Unsupervised learning comprises studying the hidden patterns and internal constructions of input statistics. As shown in Figure:2 supervised learning resolve the problems pertained to classification and regression and unsupervised learning resolve the problems pertained to clustering using relevant algorithms.

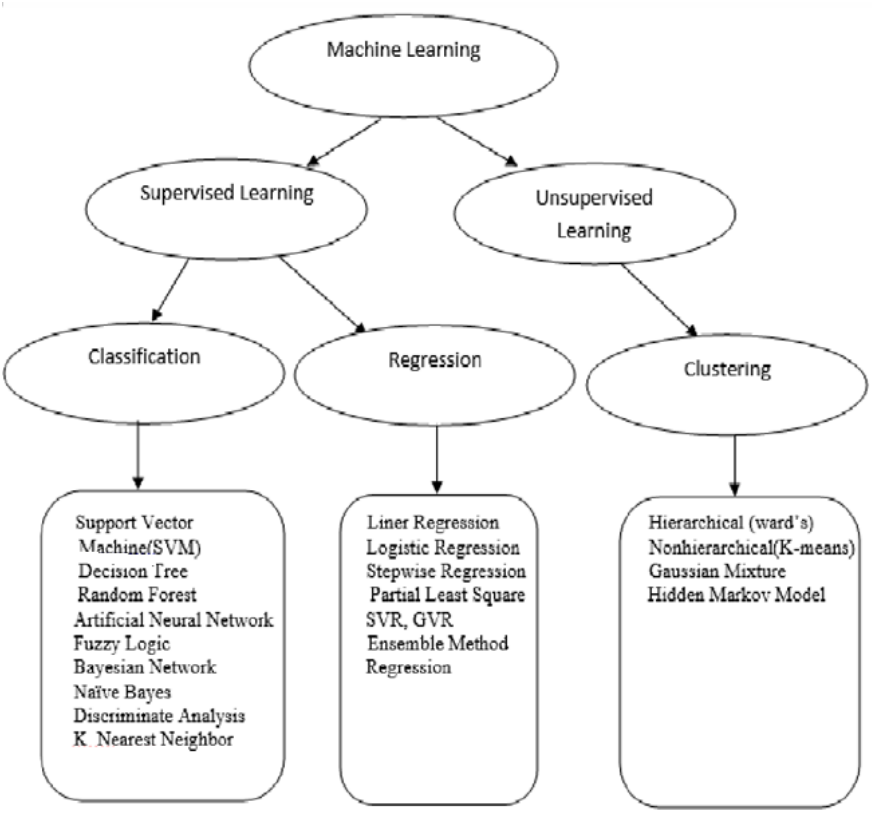


Figure: 2 Techniques of machine learning and commonly used algorithms

III. AIM AND METHODOLOGY

One of the areas of interest based on this review is the application of computer vision using machine learning technology, an application of artificial neural networks (ANNs) in animal health management. ANNs run in parallel and receive input through several layer processors. The first layer takes raw input data and processes it through a node with its own set of linked information and rules. The processor then transmits this as output to the nested layer. Each adjacent processor and node layer takes output from the previous layer and processes it further. A modest learning model implemented by a neural network is the method of putting weights to the input stream and prioritizing the model most likely to be correct. In other words, prioritizing the higher-weight and higher-weighted input current means it has a greater influence on another current (Volno, 2012). This study uses ANN to provide output for animal disease prediction. Therefore, it helps predict diseases based on location and species. The basic definition can be found in the 2011 article by Krenker et al. ANN integrates data based on how the human brain processes data to convert it into an information. Billions Neurons cells in the human brain accepts electric signals from each other and process data to drive it into an information.

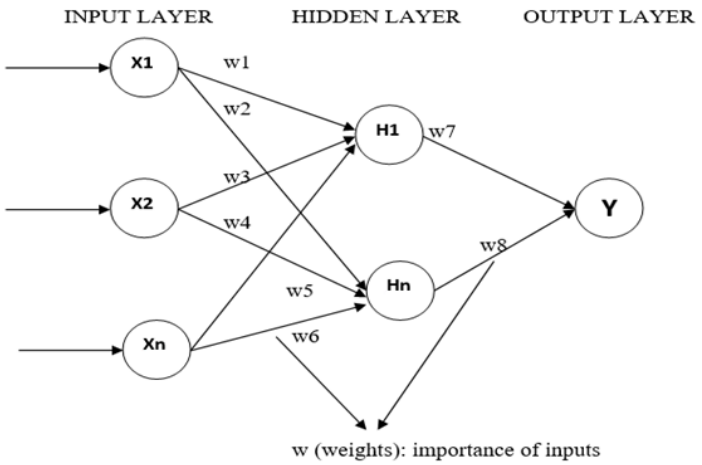
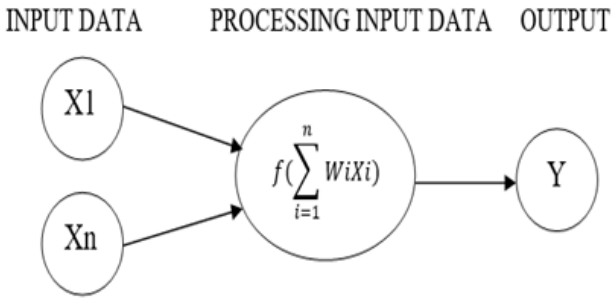


Figure: 3 Artificial Neural Network

As Shown in Figure: 3 ANN uses simple elements to join multiple nonlinear layers. It works in parallel and influence the biological nervous system. It comprises of an input layer, lots of hidden and output layers. These layers are interconnected by knobs or nodes or neurons, and each hidden layer uses the output of the preceding layer as input. ANN architecture classifications are generally feedforward neural networks (such as single-layer perceptron, multi-layer perceptron and radial basis function networks) or feedback or recurrent neural networks. (e.g., competitive networks, Kohonen's self-organizing map Hopfield networks).



It can be simple as equation

$Y=f(\sum_{i=1}^{n} w_i x_i)$

Figure: 4 The function f is the overall calculation of the neural network

IV. IMAGE PROCESSING

Image processing area refers to the processing of images by a computer system. It is used as a way of working on an image to improve an image or find useful information about it. There are three types of processing: low-level processing (the property is that both input and output are images), mid-level processing (the feature is that it has an input image, but the output is derived from these images). Objects (high-level processing) (including "making sense" of a group of recognized objects). Image processing is growing rapidly in numerous required areas today. It was recognized as an essential investigation extent in engineering and computer science. It includes two image processing approaches, one is analog and second is digital. Analog image treating is used for hard copies such as printed material and images that are continuous in nature. The analog image is a 2D image. The digital image contains a limited number of elements. Each element has a special place and value. These elements are pixels/stacks.

Using these methods, as shown in table: 1 Machine Learning implementation in the field of veterinary using image processing.

The Relevant Studies of Image Processing

Author	Year	Method	Objective	Results
Bozkurt et al.[6]	2013	ANN	They sought to determine the carcass properties of Brown Swiss and Holstein breed in the fattening field system.	They reported that body size and heart circumference were the best predictors for estimating bio-weight, according to digital image analysis and models derived from ANN. They reported that carcass length is the best predictor for estimating the weight of hot carcasses.
Mcevoy et al.[7]	2013	ANN PLSDA	They tried to identify the area containing the hip joint on the dog's radiograph.	Veterinary images can be used for educational purposes in classification and grouping, he said.
Bilgin et al.[8]	2011	Image Processing Method	They tried to show that Kangal breed dog's nostrils were different from one another.	According to the statistical data obtained, the resulting values were very different and far from each other. They said this was due to the uniqueness of the image.
Slosarz et al.[9]	2011	ANN	They sought to estimate the fat content of lamb muscle cells.	There was a significant relationship between body weight and the age of a pre-slaughtered lamb, but there was a weak link between body weight and internal fat content, he said.

ANN: Artificial neural network, PLSDA: Linear partial least squares discriminant analysis

Table: 1 *Machine learning (ML) implementation in the field of veterinary using image processing.*

V. LEARNING

Suppose you may take a series of images wherein each image comprises a category of objects, and you want the neural network to automatically recognize the objects in the image. Label the image to get training data for your network. Afatonovic-Kustrin & Beresford (2000) describes ANN's collection of useful information by finding hidden patterns, objects, and associations in data and learning through past experiences. The authors stated that ANN learns by improving the connectivity of internal units, reducing prediction errors and achieving the desired level of accuracy. After training and testing your model, you can enter new information into your model. Backpropagation, also known as the generalized delta rule, refers to the way artificial neural networks are learned based on statistics. It uses a monotonous process that involves 6 stages. (1) Then the individual status records is delivered to the input, the output is delivered to the hidden layer, and the first customary of joining weights is multiplied. (2) The arriving signal is collected, transformed to output, and delivered to the second linked weight matrix. (3) The arriving signal is collected, transformed and the network output is produced. (4) The output assessment is subtracted from the known or recognized value in this state and the error term is give back over the network. (5) Linked weights are tuned in amount to their role to the error. (6) Modified or tuned connection weights kept for the next session, the next status entry queued for the next session.

In ANN, image processing is usually represented as a matrix. Each element of the matrix contains color information for pixels with intensity of red, green, and blue (RGB). The matrix is used as an input for the neural network. Small image dimensions help you easily and quickly find vector dimensions and determine the number of input vectors. The transfer function used may be a sigmoidal function. The training ratio has values between [0.1], so it is acclaimed that the error be less than 0.1.

The input of the neural network algorithm only accepts a one-dimensional array as input, so the two-dimensional matrix representation must be converted to a one-dimensional array. Each input neuron to the algorithm denotes color data and each output neuron resembles to an image.

The procedure for image processing in ANN is as follows: Image preprocessing: Preprocessing includes converting to grayscale, reducing noise by applying filters, smoothing the image, restoring and enhancing the image. The preprocessed output will be an image of the same size as the input, but in an expanded version. **Feature Extraction or Data Reduction:** Each image contains many different pixel values. The necessary values are extracted from all these pixel values as attributes and these attributes are given to the input window. The feature can be extracted by image compression or by edge detection. **Segmentation:** it comprises defining a region of interest (ROI) by dividing the image into segments. **Classification:** This step classification objects or images according to their respective categories.

VI. CONCLUSION

Artificial intelligence and machine learning are associated with the studies that enable a machine to act like a human brain by its learning based on examples or past experiences to give the desired output. Machine learning has already been implemented in various fields like Military & Defense, Transportation, Industrial Inspection, Automation and Medical. However, it is new to the field of dairy farm animal management. It has been observed that there are lots of opportunities in the field of farm animal management since farmers are still using the conventional methods to manage their dairy farm animals. In the medical field, machine learning is being used on a large scale for the diagnosis of many diseases in real-time. Therefore, machine learning can be implemented to manage farm animals to determine their physical diseases as well as psychological conditions. Previous studies have shown that machine learning can be used in this field of animal health care to produce predictions about the future so that precautions can be taken to avoid loss in the form of farm livestock. An appropriate animal health care mechanism can be built to eliminate the risk aspects. In this review paper, related work in the field of the animal health sector associated with machine learning has been reviewed. The objective of this review paper is to show that scientific researches can be performed by using machine learning to resolve the problems related to dairy farm animals.

REFERENCES

[1] Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques. 2nd edn., Morgan Kaufmann Publishers, 29-30, 2005.

[2] Bhardwaj Rohan and Ankitha R. Nambiar: A Study of Machine Learning in Healthcare, *IEEE 41st Annual Computer Software and Applications Conference*, 2017.

[3] J. Sukanya: Applications of Big Data Analytics and Machine Learning Techniques in Health Care Sectors, *International Journal of Engineering and Computer Science*, vol. 6, pp. 21963-21967, 2017.

[4] Mishra Apoorva and Shukla Anupam: From Machine Learning to Deep Learning Trends and Challenges, *CSI Communications*, December 2018.

[5] K. Rajalakshmi, S. ChandraMohan and S. Dhinesh Babu: Decision Support System in Healthcare Industry, *International Journal of Computer Applications*, vol. 9, 2013.

[6] Bozkurt Y, Aydoğan T, Tüzün CG: Determination of performance and carcass characteristics of brown and peregrine breed animals reared in open-feedlot system by digital image processing and artificial neural network method, Tübitak Project, Project no: 111O269, 2013.

[7] Mcevoy FJ, Amigo JM: Using machine learning to classify image features from canine pelvic radiographs: Evaluation of partial least squares discriminant analysis and artificial neural network models. Vet Radiol Ultrasound, 54, 122-126, 2013. DOI: 10.1111/vru.12003

[8] Bilgin E, Ceylan M, Yalçın H: A digital image processing based bioidentification application from planum nasale of Kangal dogs. IEEE 19th Signal Processing and Communications Applications Conference (SIU), 20- 22 April, Antalya, 275-278, 2011. DOI: 10.1109/SIU.2011.5929640

[9] Slósarz P, Stanisz M, Boniecki P, Przybylak A, Lisiak D, Ludwiczak A: Artificial neural network analysis of ultrasound image for the estimation of intramuscular fat content in lamb muscle. Afr J Biotechnol, 10, 11792-11796, 2011.

[10] Ghotoorlar SM, Ghamsari SM, Nowrouzian I, Ghotoorlar SM, Ghidary SS: Lameness scoring system for dairy cows using force plates and artificial intelligence. Vet Rec, 170, 126, 2012. DOI: 10.1136/vr.100429

[11] Shortliffe EH, Cimino JJ, (editors): Biomedical Informatics: Computer Applications in Health Care and Biomedicine. 3rd edition. New York: Springer; 2006.

[12] Sornmo L, Laguna P: Bioelectrical Signal Processing in Cardiac and Neurological Applications. London: Academic Press Inc; 2006.

[13] Bhardwaj Rohan and Ankitha R. Nambiar: A Study of Machine Learning in Healthcare, *IEEE 41st Annual Computer Software and Applications Conference*, 2017.

[14] J. Sukanya: Applications of Big Data Analytics and Machine Learning Techniques in Health Care Sectors, *International Journal of Engineering and Computer Science*, vol. 6, pp. 21963-21967, 2017.

[15] Amasyalı MF: New machine learning methods and their applications to drug design. PhD Thesis, Yıldız Technical Univ. Science Science Inst., 2008.

[16] Bal M, Sever H, Kalıpsız O: Modeling the symptom-disease relationship by using rough set theory and formal concept analysis. World Acad Sci Eng Technol (WASET), 26 (12): 517-521, 2007.

[17] K. Shailaja; B. Seetharamulu; M. A. Jabbar: Machine Learning in Healthcare: A Review. IEEE, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)

[18] Tan, J& sheps ,B.: (1998) Health decision support systems, Jones &Bartlett publishers

[19] Wickramasinghe, N. &Geisler: E. (2008).Encyclopedia of health care information systems, information science

[20] R. Nambiar, A. Sethi, R. Bhardwaj and R. Vargheese: A look at challenges and opportunities of Big Data analytics in healthcare, *IEEE Big Data Conference*, 2013.

[21] T. Daveport: Industrial-Strength Analytics with Machine Learning, *The Wall Street Journal*.

[22] J. Brownlee: What is Machine Learning, A Tour of Authoritative Definitions and a Handy One-Liner You Can Use

[23] D. Page: Challenges in Machine Learning from Electronic Health Records, *MLHC*, 2015.

[24] C.M. Bishop: Neural networks for pattern recognition, England: Oxford University, 1995.

[25] K. Shailaja, B. Seetharamulu, M. Jabbar: Machine Learning in Healthcare: A Review, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)

REVIEW OF DAIRY ANIMAL PHYSIOLOGICAL PARAMETERS AND CRITICAL DISEASE DETECTION METHODS

Er. Atul Gupta, Research Scholar,
Department of Computer Science and Engineering, Punjabi University, Patiala
Er. Karandeep Singh, Asstt. Professor,
Department of Computer Science and Engineering, Punjabi University, Patiala

ABSTRACT— Physiological parameters of dairy animals and key illness detection strategies are reviewed in this paper. In current situation, due to the dependence of human life on dairy animals, so it is set to examine dairy animal health on usual basis and routine health check-up of a dairy animal cost too a large amount, which a quantity of farmers may not be capable to afford. Moreover this, if the health of a dairy animal is not taken care of accurately and diagnosed well-timed, it can be life threatening to the dairy animal. Therefore, farmers can constantly monitor physiological parameters of dairy animals such as body temperature, Heart Rate, Behavior and critical disease like Mastitis and FMD.

The review covers several methods to monitor the aforesaid physiological parameters and critical disease (Mastitis and FMD (Foot and Mouth disease) through IOT Bases Sensor technology, Electronic nose system (with sensor technology), Micro-Controller and Machine Learning Algorithm (Neural Networks) system and Electrical conductivity method (ECM) of milk.

KEYWORDS– IOT Bases Sensor technology, Electronic nose system, Neutral Network, Electrical Conductivity

INTRODUCTION

Dairy production is one of the world's fastest-growing companies and a critical component of the global financial system. Through a boost in milk production in the world, here is an increase in the figure of dairy animals too. One of the key challenges, which are faced by the dairy industry, is the healthcare of Dairy animals. Its only cannot be answerable for the health examine of various dairy animals; but, farmers also require to improving and contributing by taking care of their animals.

Because of the steady increase in air temperature in the troposphere, livestock farmers are currently looking at Dairy Animal health problems in the region of the planet. The impact of temperature changes on dairy animals' health are dangerous, leading to diseases including foot and mouth disease and mastitis.

Due to the current state of individual life's reliance on dairy animals, it is necessary to monitor dairy animal physical status on a regular basis. Moreover, constant health check-ups of a dairy animal cost too much, which a few farmers may not be capable to meet the expenses. Moreover, if the health of a dairy animal is not taken care accurately and not to timely diagnose, it can be life threatening to the dairy animal. Therefore, farmers can constantly monitor physiological parameters of dairy animals such as body temperature, Heart Rate, Behavior and critical disease like Mastitis and FMD.

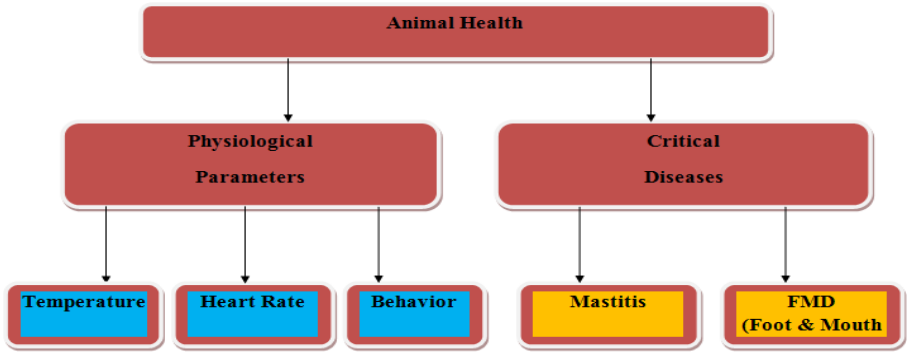


Figure – 1 Animal Health Management based on Physiological parameters and Critical Diseases

Physiological Parameters of Dairy Animals

Temperature

A significant characteristic is the ability to measure body temperature. The entire antibiotic treatment decisions are based on body temperature. The standard body temperature of Dairy Animals ranges from 37 to 39 degree C. If the body temperature of the animal more than the given range, its indicate abnormality.

Because all antibiotic treatment decisions are based on body temperature, compute body temperature is a crucial attribute. A number of factors, including the type of thermometer used, the depth of insertion, and the investigator's skills, might affect the assessment of body temperature in dairy animals (days in milk, time of the day, during the time of milk).

A healthy resting cow's normal core body temperature is 101.5 degrees Fahrenheit (38.6 degrees Celsius), with a fever being detected when the temperature exceeds 103.0 F. Table 1 shows the normal temperatures of dairy animals.

Kind of Animal	Temperature		Rate Per Minute
	C	F	Pulse
Buffalo	38.3	101	40-50
Cow	38.5	101.4	50-60
Goat	39.8	103.8	70-90
Sheep	39.1	102.4	70-90
Pig	39.1	102.4	70-80
Chicken	41.7	107.2	128-140
Camel	36.3	97.4	32-50
Cat	38.5	101.4	100-130
Dog	38.8	102	70-00
Elephant	36.3	97.4	22-35
Mare	37.7	100	38-45
Rabbit	39.5	103.2	-
Man	36.8	98.4	60-90

Body Temperature and Heart Rate Range (Table 1)

Heart Rate

To Measuring Heart rate (number of heart beats per minute) in the dairy animals is also a key feature. In Dairy Animals, the approximate heart rate ranges from 48-84 beats per minute. An adult cow's heart rate ranges from 48 to 84 beats per minute. This can be assessed using a stethoscope and listening to the left side of the cow's chest behind the cow's elbow. Table 1 shows the normal heart rate of dairy animal.

Standard Body Temerature and Heart Rate of Dairy Animals

Dairy Animal	Heart Rate(Bpm)	Body Temeprature(Celcius)
Cattle	48-84	37-39

Table 2: Dairy Animals' Average Body Temperature and Heart Rate

Behavior

When dairy animals become ill, their behaviour changes, which can be linked to explicit illnesses and behavior. For example, cattle with severe lameness spend fewer times at the feed and eat lesser. Sick animals will be lazier, isolate themselves and lose their appetite. One can also detect changes in behavior for some illnesses before there are any clinical signs. This in turn affects the reproductive cycle and hence the dairy output of the cow.

Critical Diseases of Dairy Animals

Mastitis Disease

Mastitis is puffiness of the mammary gland and swelling of the udder. This disease reduces the milk giving capability of dairy animal (cows). Mastitis is reason by a bacterial infection in the udder tissue of the cow. One of the reasons how it is spread while an improper way is employed to milk the cows.

In rising countryside, it is not easy for a farmer to execute farm automation. Hence, they normally milk the cow with their hands. This way approach with the risk of a germ-infested environment for the movement as the farmer's hands may be carrying virus/bacteria which could get the cow's udders infected. This infection in the mammary gland of the cow causes Mastitis. For the reason that of Mastitis, the quality of milk is get of inferior quality.

There are three stages in which Mastitis can be detected

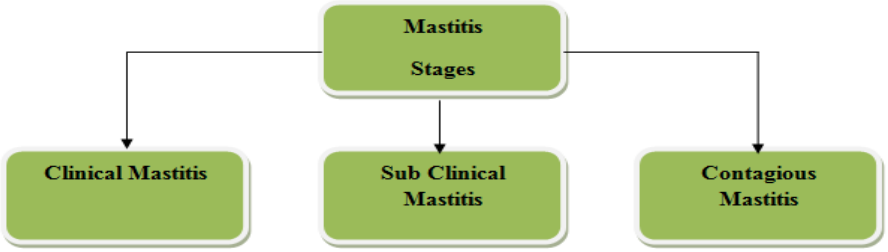


Figure – 2 Stages of Mastitis

Stages of Mastitis

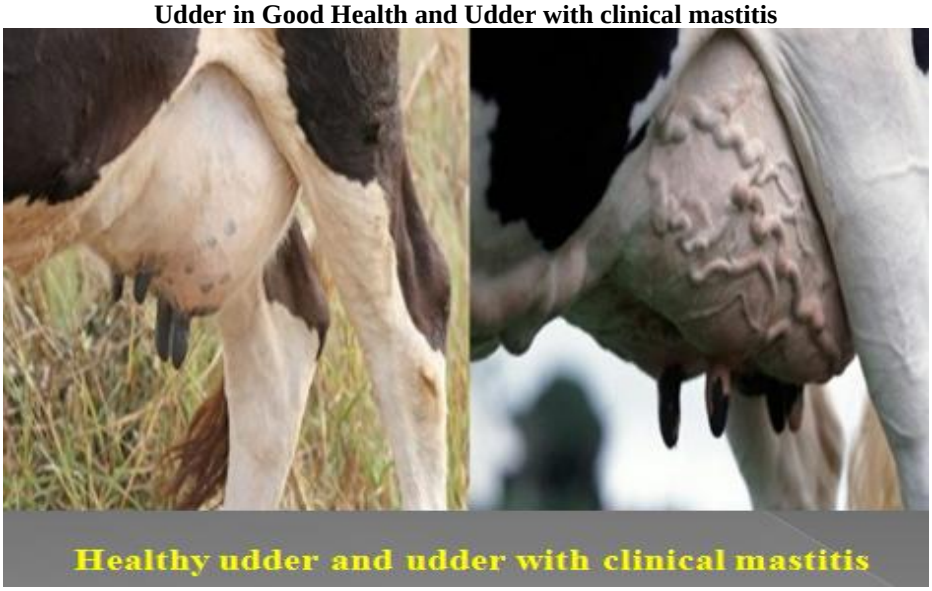
Clinical Mastitis: A farmer may notice a few signs that an animal is diseased, such as this case occur clots in milk/Flakes; the udder may swell a spot, as the temperature rises, so does the amount of milk produced.

Sub Clinical Mastitis: Farmer cannot forecast this case as the cow shows very fewer indication like slow rise in temperature, increase in Somatic cell count (SSC) in the milk, Decrease in Milk production. A Dairy animal's start avoidance the food.

Contagious Mastitis: That's the severe condition of Mastitis where the swelling of the udder is hits the highest point, fast reduction in the quantity of milk and their will be Somatic cell counts are more than 300,000 Cells/ml.

Fig.1. (a) and (b) shows the infected mammary gland of female cattle suffering from Mastitis disease.

Udder in Good Health and Udder with clinical mastitis



FMD (Foot and Mouth Disease)

FMD (Foot and Mouth Disease) threatens dairy animals in a number of developing countries. It's a virus that causes symptoms such as fever, blisters in the mouth and foot, and lameness. If this viral disease is not treated promptly, it weakens dairy animals and reduces their milking capacity, reducing their efficiency. When dairy animals are housed in unsanitary conditions and drink dirty water, FMD spreads.

If FMD is not treated in a timely manner, it affects the efficiency of the cattle by weakening them and reducing their ability to produce milk. That is, sick animals cannot be restored to their previous ability to produce milk.

When cattle are maintained in contaminated facilities and drink dirty water, FMD spreads FMD usually appears between 3-6 days after the infection has taken hold.

Blisters in Foot & Mouth

LITERATURE REVIEW

Berry, R.J., et al. (2003) used infrared thermography to track daily and within-day variations in udder temperature in dairy cows (IRT). The initial measurement and prediction of the change in udder surface temperature will ideally serve as a foundation for the creation of a mastitis early detection system in the future.

According to M. Janzekovic et al., higher average electric conductivity than 6.5mS/cm was associated with an increase in the number of somatic cells in milk in 80% of cases (2009) .The application of ECM to the diagnosis of subclinical mastitis.

Specific sensor technologies have been established as a crucial technique of monitoring animal health, according to Miss. Amruta Helwatkar and her colleagues (2014). Temperature, Accelerometer, and Microphone are the three key sensors used in this study to assess the health quotient of cattle.

A different approach was taken by Lien.Cheng et al. (2016), who assessed the EC in milk. Utilizing alone EC to detect mastitis has historically been less accurate than using a combination of EC and other data. This framework enables the development of a simple and cost-effective mastitis detector that provides real-time detection results without the need for historical EC records.

Sweta Jha and her colleagues (2017) identified specific sensor technologies as a valuable tool for monitoring animal health. Several cow illnesses have been thoroughly investigated, with the symptoms analyzed. Sensors that could measure behaviour were used to map these symptoms.

On the basis of electro-chemical data as well as milk quality characteristics of normal and mastitis milk, Panchal. Indu. et al. (2017) built and validated numerous connectionist models to identify healthy vs mastitis murrah buffaloes. Connectionist models were compared to classical multiple linear regression models in terms of performance.

Smart computing and sensing technologies for domestic, wild, and farm animal welfare were established by Roger Rozario A.P et al. (2018). These technologies are used to determine whether an animal is healthy, pain-free, and stimulated in the surroundings in a favorable way.

Anand, M. J., et al. (2019) used a gas sensor to identify mastitis. Milk samples can be classified using the E-nose system based on headspace volatile readings. The classification results were enhanced by using PCA and LDA.

Vyas Shivank et al. (2019) developed the Internet of Things (IOT) in the identification of Mastitis and FMD, which will have a significant impact on the reduction of these diseases, thanks to the employment of Neural Networks and smart sensors. As a result, the lower quality of cow milk will be reduced, resulting in lower dairy processing costs.

Dr. Kirti Wankhede and Manisha Pathakal developed a contemporary way for controlling animal fitness using biosensors. For the identification of many infectious diseases in livestock, nano biosensors and advanced molecular biology diagnostic techniques are being developed.

With the help of a Raspberry Pi3, a body temperature sensor, a heartbeat sensor, and a rumination sensor, Seema Kumari and Dr. Sumit Kumar Yadav built a prototype of an IOT based smart animal health monitoring system that can monitor body temperature, heartbeat, and rumination in real time.

Snehal.S.Kharde and Meenakshi. M. Kharde created a sophisticated health monitoring system that monitors the health parameters of cows such as body temperature, humidity, and respiration using sensors. Sensors are connected to an Arduino UNO, which displays the graph on the I Chart app through a Wi-Fi module.

Adithya Sampath and P Meena created the LORA IOT wearable device, which monitors the temperature, activity, and position of a cow and communicates the information to a LORA gateway using the LORA protocol. The farmers and veterinarians can use the cloud analytics performed on the sensor data received at the gateway to compile a report with actionable information.

Literature Review

S.No.	Publication Title	Author(s)	Tools and Techniques / Technology	Salient Findings
1	Daily variation in the udder surface temperature of dairy cows measured by infrared thermography: Potential for mastitis detection	Berry. R.J. et al. (2003)	Infrared thermography Technique	Infrared thermography was used to track the daily and intraday variations in udder temperature in dairy cows (IRT). The initial measurement and prediction of udder surface temperature variation will ideally serve as a foundation for the future development of a mastitis early detection system.
2	Mastitis detection based on electric conductivity of milk	M. Janzekovic. et al. (2009)	Electric conductivity Method (ECM)	Higher average electric conductivity than 6.5mS/cm was shown to be associated with an increase in the number of somatic cells in milk in 80 percent of cases. ECM as a diagnostic tool for subclinical mastitis
3	Sensor Technology For Animal Health Monitoring	Miss. Amruta Helwatkar et al.(2014)	Sensor Technology	To create a specific sensor technology as a key tool for monitoring animal health. Temperature, Accelerometer, and Microphone are the three key sensors used in this study to assess the health quotient of cattle.
4	Online detection of dairy cow subclinical mastitis using electrical conductivity indices of milk	Lien.Cheng et al. (2016)	Electric conductivity Method (ECM)	A different way is to measure the EC in milk. Utilizing simply EC for mastitis detection has traditionally been less accurate than using a combination of EC and other data. This framework enables the development of a simple and cost-effective mastitis detector that provides real-time detection results without the need for historical EC data.
5	E-Cattle Health Monitoring System Using IOT	Sweta Jha et al.(2017)	Sensors Technology and Embedded System and IOT and Cloud Computation.	To develop particular sensor technology as an important tool for animal health monitoring. Several cow diseases have been examined in depth, with symptoms analysed. These symptoms were matched to sensors that could track the behaviour.
6	Mastitis detection in Murrah buffaloes with intelligent models based upon electro-chemical and quality parameters of milk	Panchal. Indu. et al. (2017)	Connectionist models and Error Back propagation algorithm	To distinguish between healthy and mastitis mastitis, several connectionist models have been designed and validated. Electro-chemical measures, as well as milk quality parameters of normal and mastitis milk, were used to evaluate Murrah buffaloes. Connectionist models were compared to classical multiple linear regression models in terms of performance.
7	Recent Advances in IOT Based wireless sensors for Cattle Health Management	Roger Rozario A.P et al.(2018)	IOT Based Wireless Sensor Technology	Domestic, wild, and farm animal welfare using smart computing and sensing technology. These technologies are used to determine whether an animal is

				healthy, pain-free, and stimulated in the surroundings in a favorable way.
8	Detection of Sub-Clinical Mastitis Using Prototype Electronic -Nose	Anand. M. J. et al. (2019)	Electronic Nose System With Sensor Technology	Mastitis was detected using a gas sensor. Milk samples can be classified using the E-nose system based on headspace volatile readings. The classification results were enhanced by using PCA and LDA.
9	FMD and Mastitis Dsease Detection in Cows Using Internet of Things (IOT)	Vyas Shivank et al. (2019)	Micro-Controller and Machine Learning Algorithm (Neural Networks)	Because of the employment of Neural Networks and smart sensors, the use of the Internet of Things (IOT) to detect Mastitis and FMD will have a significant impact on the reduction of these diseases. As a result, the poor quality of milk provided by cows will be reduced, lowering dairy processing costs.
10	Use of IOT in Animal Husbandry	Dr. Kirti Wankhede, Manisha Pathakala	IOT and Nano Bio-Sensors Technique	Biosensors are used in a modern way to control animal fitness. For the identification of many infectious diseases in livestock, nano biosensors and advanced molecular biology diagnostic techniques are being developed.
11	Development of IOT Based Smart Animal Health Monitoring System using Raspberry Pi	Seema Kumari, Dr. Sumit Kumar Yadav	IOT and Raspberry Pi Wi-Fi Technology	With the help of a Raspberry Pi3, a body temperature sensor, a heartbeat sensor, and a rumination sensor, develop a prototype of an IOT-based smart animal health monitoring system that is capable of real-time monitoring of body temperature, heartbeat, and rumination.
12	Advance Cattle Health Monitoring System Using Arduino and IOT	Meenakshi .M, Snehal. S. Kharde	Arduino UNO, Sensor and IOT Based Technology	Developed an innovative health monitoring system that monitors a cow's health parameters such as body temperature, humidity, and respiration using sensors. Sensors are connected to an Arduino UNO, which uses a Wi-Fi module to display the graph on the I Chart app.
13	A Novel Approach to Cattle Health Monitoring for Maximizing Dairy Output using LORA IOT Technology	Adithya Sampath, P Meena	LORA IOT Technology and Cloud Computation	The temperature, activity, and position of a cow are monitored by a LORA IOT wearable device and sent to a LORA gateway. The farmers and veterinarians can use the cloud analytics performed on the sensor data received at the gateway to compile a report with actionable information.

FUTURE PROPOSED WORK

To design and develop an IOT (Internet of Things)-based Health Management System for Dairy Animals in order to detect the bacterial disease Mastitis early and prevent Dairy Animals from becoming life-threatening.

To reduce financial loss in dairy farming owing to infertility in dairy animals caused by Mastitis, a bacterial illness

CONCLUSION

Various dairy animal physiological metrics as well as critical disease (mastitis and FMD) approaches are discussed in this paper. These methods are based on Internet of Things (IOT) sensor technology, an electronic nose system (with sensor technology), a microcontroller and machine learning algorithm (Neural Networks) system, and a milk electrical conductivity method (ECM). The several important diseases (Mastitis and FMD) approaches for dairy animal disease detection are analyzed in this study.

In the future, an Internet of Things (IOT)-based Health Management system will be built with minimal changes to detect crucial diseases in dairy animals such as mastitis and FMD.

REFERENCES

[1] Shivank Vyasa, Vipin Shuklab, Nishant Dosh "FMD and mastitis Disease Detection in Cows Using Internet of things", Procedia Computer Science 160 (2019) 728–733.

[2] Indu Panchal, I.K. Sawhney, A.K Sharma, M.K. Garg and A.K. Dang"Mastitis detection in murrah buffaloes with intelligent models based upon electro-chemical and quality parameters of milk", Indian J. Anim. Res., 51 (5) 2017: 922-926.

[3] Dr. Kirti Wankhede, Manisha Pathakala "Use of IOT in Animal Husbandry", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, p-ISSN: 2278-8727, PP 40-44, www.iosjournal.org**.**

[4] Sweta Jha, Amruta Taral, Komal Salgaonkar, Vaishnavi Shinde, Shraddha E-Cattle Health Monitoring System Using IOT", Journal of Network Communication & Emerging Technique(JNCET), Volume 7, Issue 8, August(2017), www.jncet.org.

[5] Seema Kumari, Dr. Sumit Kumar "IOT Based Smart Animal Health Monitoring System using Raspberry Pi", Special issue based on Proceeding of 4th International Conference on Cyber Security (ICCS) 2018.

[6] Meenakshi .M, Snehal. S. Kharde"Advance "Cattle Health Monitoring System using Arduino and IOT", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, ISSN: 2320-3765, ISSN: 2287-8875, Volume 5, Issue 4, and April 2016.

[7] Miss. Amruta Helwatkar, Daniel Riordan & Joseph Walsh" Sensor Technology For Animal Health Monitoring", Proceeding of 8th International Conference on Sensing Technology, Sep. 2-4, 2014, Liverpool, UK.

[8] Roger Rozario.A.P, Pravinthraja.S, Arjuman Banu.S, Nandhini.M "Recent Advances in IOT Based wireless sensors for Cattle Health Management", International Journal of Institutional & Industrial Research ISSN:2456-1274, Vol. 3, Issue 1, Jan-April 2018, pp. 78-80.

[9] Adithya Sampath, P Meena"Cattle "A Novel Approach to Cattle Health Monitoring for Maximizing Dairy Output using LORA IOT Technology", International Journal of Advanced Research in Computer Engineering, ISSN: 2278-1021, ISSN:2319-5940, Volume 8, Issue 6, June 2019.

[10] Anand M J, V. Sridhar, Ramasamy Ravi "Detection of sub-clinical mastitis using prototype electronic -nose", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019.

[11] Cheng-Chang Lien , Ye-Nu Wan , Ching-Hua Ting "Online detection of dairy cow subclinical mastitis using electrical Conductivity indices of milk", Engineering in Agriculture, Environment and Food 9 (2016) 201-207.

[12] M. Janzekovic , M. Brus , B. Mursec , P. Vinis , D. Stajnko , F. Cus "Mastitis detection based on electric conductivity of milk", Journal of achievements in material & manufacturing Engineering(JAMME), Volume-34, Issue-1, May - 2009.

[13] R. J. Berry, A. D. Kennedy, S. L. Scott, B. L. Kyle, and A. L. Schaefer "Daily variation in the udder surface temperature of dairy cows measured by infrared thermography: potential for mastitis detection", Canada T4L 1W1. Received 14 Feburary 2003, accepted 16 June 2003.

[14] C. J. Rutten, A. G. J. Velthuis, W. Steeneveld, and H. Hogeveen, "Invited review: Sensors to support health management on dairy farms," J. Dairy Sci., vol. 96, pp. 1952–1928, 2013.

[15] W. Steeneveld, L. C. van der Gaag, W. Ouweltjes, H. Mollenhorst, and H. Hogeveen, "Discriminating between true-positive and falsepositive clinical mastitis alerts from automatic milking systems.," J. Dairy Sci., vol. 93, no. 6, pp. 2559–68, Jun. 2010.

[16] T. T. F. Mottram, H. R. Whay, S. G. Vass, and Birte Lindstrom Nielsen, "Patent US6270462 - Apparatus for animal health monitoring - Google Patents," 2001.

[17] P. Løvendahl and M. G. G. Chagunda, "On the use of physical activity monitoring for estrus detection in dairy cows.," J. Dairy Sci., vol. 93, no. 1, pp. 249–59, Jan. 2010.

[18] T. Godsk and M. B. Kjærgaard, "High classification rates for continuous cow activity recognition using low-cost GPS positioning sensors and standard machine learning techniques," pp. 174–188, Aug. 2011.

[19] G. Tielens, "Device for animals and mode of operation of the detection methodology of the device to monitor, to report and to alarm changes of the intra - abdominal pressure by using communication technology," 19-Mar-2008.

[20] M. Futagawa, T. Iwasaki, M. Ishida, K. Kamado, M. Ishida, and K. Sawada, "A Real-Time Monitoring System Using a Multimodal Sensor with an Electrical Conductivity Sensor and a Temperature Sensor for Cow Health Control," Jpn. J. Appl. Phys., vol. 49, no. 4, p. 04DL12, Apr. 2010.

[21] T. Mottram, J. Lowe, M. McGowan, and N. Phillips, "Technical note: A wireless telemetric method of monitoring clinical acidosis in dairy cows," Comput. Electron. Agric., vol. 64, no. 1, pp. 45–48, Nov. 2008.

[22] H. Hogeveen, C. Kamphuis, W. Steeneveld, and H. Mollenhorst, "Sensors and clinical mastitis--the quest for the perfect alert." Sensors (Basel)., vol. 10, no. 9, pp. 7991–8009, Jan. 2010.

AN INSIGHT ON SOFTWARE VULNERABILITY DETECTION USING CODE CLONES- PAST AND FUTURE TRENDS

Gurpreet Singh[1], Dhavleesh Rattan[2]
Computer Science and Engineering, Punjabi University, Patiala
[1]gurpreet.1887@gmail.com, [2]dhavleesh@gmail.com

ABSTRACT—— There is a rapid growth in number of open-source software (OSS) in the last few years. This growth has linear to quadratic pattern [22]. Source-Forge recorded growth in number of open-source projects from 136 K to 430 K during October 2009 to March 2014 and GitHub reported the creation of 10 million repositories in December 2013 [9]. According to Black Duck Corp about 66% of the commercial applications had been reported with known vulnerabilities in it [15].

During software development, clones can occur in software intentionally or unintentionally. Developers tend to clone fragments of software during development to save efforts and expedite the development process. The use of application programming interfaces (APIs) can lead to generation of unintentional clones. Large systems contain 20-30% cloned code [19]. If cloned code is buggy or vulnerable, then cloning practice led to bug propagation. However, security fix of specific vulnerability often does not spread to code at other locations. Vulnerable code reuse in open software is serious threat to software security [15].

The vulnerability prevalence problem is a serious issue and this problem can't be simply solved with multiple patch management mechanism, as they often don't cover all vulnerability occurrences. While it may sound easy task to locate code reuse, but it is actually unmanageable because of large number of programs [12]. This can be witnessed by the fact that despite the presence of 13 automated patching mechanism, at least 86% median fraction of computers are unpatched at the time of exploits are available [16]. The lack of standard data set for this kind of research is one of the main challenges. There is lack of single code clone detection algorithm that is suitable for all kinds of vulnerabilities, as each vulnerability has its own characteristics that should be taken in to account. However, it is not known which code clone detection algorithm will be effective for which vulnerabilities [12]. So, there is need of efficient and scalable approach for detecting code clones having software vulnerability. The existing techniques are not able to detect all instances of vulnerable code clones. Different approaches suffer from high false positive rate and not scalable to large software systems due to high time complexity.

1) Introduction: Programmers often use copied code fragments with or without modifications in software development for reuse existing code. Such copied code fragments are considered as code clones. The activity of using copied code fragment is called code cloning. These code clones are similar or same code fragments that tend to appear in different locations of source code in software. During software development code clones are introduced to save organization's development time, development cost and to avoid errors. Because, these code fragments are well tested and used in previously developed software. Rattan et al. [19] reviewed an extensive amount of research on code clones and reported number of reasons for existence of code clones but many studies [4], [5], [6] have proved propagation of bugs, increase in maintenance cost, resource requirements [20] due to code clones, which reflect negative effect of code clones on software development. If there is a need to change any clone fragment, then it became crucial to consistently update all replicated similar clone fragments [16]. Inconsistent changes to clone fragments are also one of the main reasons for bug propagations in software development and increase in maintenance overhead [19], [20] in terms of cost, time, and efforts. Thus, it become an important job to locate all related clone fragments in software. Information about code clones help developers to fix bugs in better way as compared to they have no information about code clones [2].

2) Software Vulnerability: A software vulnerability can be viewed as a flaw, weakness or even an error in specification, development, or configuration of software that can be exploited by a hacker/attacker in order to compromise a system [10]. In simple terms, loophole or bug in software code can be treated as vulnerability in computer software. Not every software bug is a vulnerability. The bug must be exploitable to be considered as a vulnerability. Most of bugs cause lethal disruption in normal processing of software but can't be leveraged to compromise a system. Every software vulnerability has life a cycle [24]. The life cycle of vulnerability starts with its detection by vendor, a hacker or any third party. High security risk is associated with a vulnerability if it is first discovered by hacker. Next phase starts when it is publicly disclosed by vendor, a hacker, or a third party. The security risk further escalated by public disclosure of vulnerability because hacker community get active in developing and launching zero-day attack [1]. On the other side vendor has to release the patch (patches are the additional pieces of code developed to address problems in software [23]) as soon as possible to avoid the exploitation of vulnerability. Many users of the affected software don't install the released patch, might be they are unaware about the security patch. Life cycle ends when all users of affected software install the security patch to fix a vulnerability. An attacker or hacker can exploit vulnerability at any time during its entire life cycle. Each vulnerability has been assigned with unique name called CVE ID. CVE Numbering Authorities (CNAs) [44] are official to allot name to vulnerabilities affecting software. Vulnerability id start with prefix "CVE" followed by year in which it was discovered then followed by unique integer number. E.g., CVE-2014-0160 is a heartbleed vulnerability found in SSL protocol. The Common Vulnerabilities and Exposures (CVE) [45] service is maintaining list of public vulnerabilities. Which is considered as trusted, inclusive and independent site for reporting and tracking public vulnerabilities [4]. Large numbers of vulnerabilities are reported every year. Many vendors don't publicly disclose every vulnerability. Many of the experts

think that the actual figure is significantly higher than publicly known figures of vulnerabilities [4]. It is most important to assess the severity level of vulnerabilities and to prioritize them, which will help the vulnerability management processes. CVSS (The Common Vulnerability Scoring System) [43] do the job by capturing principal characteristics of vulnerability and compute a numeric score reflecting vulnerability severity. Overall score is computed based on formula that depends on several metrics like base score metrics, temporal score metrics, environmental score metrics. CVSS score varies from 0 (none) to 10 (critical). Quantitative severity rating scale is further mapped to qualitative severity rating scale as shown in below table 1.

Table 1: Vulnerability severity scale

Rating	CVSS Score
None	0.0
Low	0.1 - 3.9
Medium	4.0 – 6.9
High	7.0 – 8.9
Critical	9.0 – 10.0

2.1 Vulnerability Example: Heartbleed (CVE-2014-0160)

Heartbleed is serious bug in the OpenSSL cryptographic software library [9]. Several types of systems (including, web server, websites, software applications and operating system distributions) has been affected with this vulnerability. CVE-2014-0160 is the Common Vulnerabilities and Exposures designation for this vulnerability [47]. Heartbleed is two-way vulnerability, thus both server and client can be compromised. This Vulnerability was identified in 2014 by the team of security engineers from Google. All started with introduction of Heartbeat extension to TLS (Transport Layer Security) in 2012 as a phase of development in OpenSSL standards. To keep session alive and connected, Heartbeat messages are sent continuously to server as shown in Figure 1.4. The server will respond to packet by quarrying its own memory of the number of bytes mentioned in the packet and send it to client. Thus, vulnerability allow the attackers to compromise the memory of system which contain sensitive data including user authentication credentials and secret keys secured by the vulnerable OpenSSL. Vulnerability come into existence due to minor mistake in implementation of OpenSSL Heartbeat extension as there is no bound check on the size of packet to be acknowledged to client/server data including user authentication credentials and secret keys secured by the vulnerable OpenSSL. Vulnerability come into existence due to minor mistake in implementation of OpenSSL Heartbeat extension as there is no bound check on the size of packet to be acknowledged to client/server.

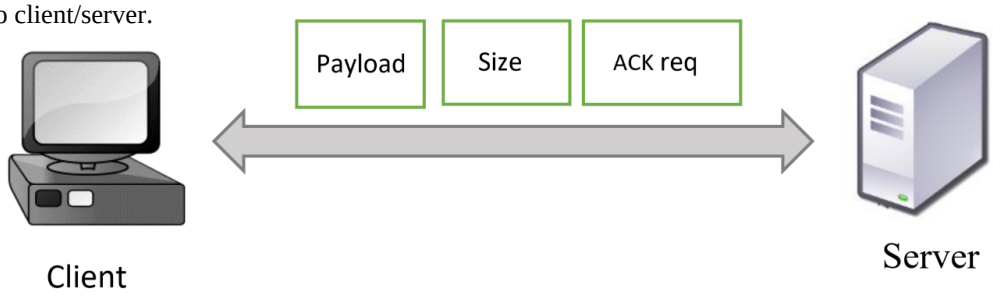


Fig 1: Heartbeat Function

OpenSSL version 1.0.1 to 1.0.1f are affected with this vulnerability, after the release of patch the stable version 1.0.1g is introduced in market. It is highly recommended that we all should use the non-vulnerable version of OpenSSL.

Vulnerability Detection using code clone trends:

Kim et al. [9] proposed VUDDY, which is a scalable approach for detection of vulnerable code clones. This approach can detect vulnerabilities efficiently and accurately in large software. They able to achieve extreme level of scalability by using function-level granularity and a length-filtering techniques that decreases number of signature comparisons. Most interesting feature of this technique is that it can even detect variants of known vulnerabilities. To achieve extreme level of scalability, they used function-level granularity and length filtering techniques to reduce number of signature comparisons.

Li et al. [12] proposed VulPecker, a system to automatically detect whether a software code fragment contains a given vulnerability or not. Firstly, they build a Vulnerability Patch Database (VPD) and a vulnerability Code Instance Database (VCID) from open-source C/C++ products that have some vulnerabilities according to NVD (National Vulnerability Database [27]). Then they used machine learning algorithms for selection of code similarity algorithm that is effective for one specific vulnerability as there is no single code-similarity algorithms that is effective for all kind of vulnerabilities. After the experiment result shows that VulPecker was able to detect 40 vulnerabilities that was not available in NVD, out of these 40 vulnerabilities 22 were “silently” patched by vendors in upcoming release of affected product.

Liu et al. [15] designed a fingerprint model for vulnerable code clone detection and proposed VFDETECT to detect a given vulnerability in a software code fragment based on vulnerability fingerprint. Even after modification from level 1 to level 4, VFDETECT was able to detect codes because they used code block as the granularity unit for clone detection. For

fingerprint generation they leverage MD5 hash function with 8- byte output to generate hash value for all diff files in the patch. The experimental result shows that the time cost is roughly linear with size of code.

Li et al. [11] proposed a novel mechanism CLORIFI to detect code clone vulnerability using code clone verification. To detect code clone vulnerability using code clone verification they combined static and dynamic analysis techniques. Further to reduce number of false positives i.e., to verify a vulnerability, they took the help of concolic testing. However, CLORIFI was able to improve the detection but at the cost of large resource consumption. It had higher false negative rate.

Jang et al. [7] proposed a system named ReDeBug which is capable for quickly finding unpatched code clones in OS-distribution sized code irrespective of programming languages. To detect code clones authors used sliding window of n (n=4 by default) lines. Then applied three different hash function on each window. They detected the clones between files by membership checking in bloom filter. ReDeBug is not able to detect Type-2 clones with modification. It produces many false positive cases.

Li and Ernst [13] implemented a tool CBCD: Cloned Buggy Code Detector that can detect code fragments which are semantically same to buggy code. Authors used directed PDG for code representations, both buggy code as well as source code are represented as PDG, and then CBCD report the presence of bug if bug PDG is subgraph the system PDG. To make subgraph checking cheaper and to reduce complexity, CBCD splits the system PDG as subgraph isomorphism is NP-Complete problem.

Jiang et al. [8] proposed an approach to expose clone related errors by detecting inconsistencies for code clones. In this approach they used a clone detection tool Deckard to detect code clones in the programs. Then they used parse tree to figure out inconsistencies in the context of clones. After classifying inconsistencies based on their potential relations with bugs are reported. Researchers evaluated their approach on Eclipse and Linux kernel. Important finding of their research is that no single existing program analysis technique can discover clone related bugs.

Perl et al. [17] presented an approach VCCFinder to improve code audits. In this approach, first they created vulnerable commit database by mapping CVEs to GitHub commits. Then they used SVM classifier to figure out suspicious commits. Unlike the other approaches VCCFinder can be used on code snippets which make it lightweight analysis approach.

14 Because it is far easier to test code snippets than requiring a full build environment to be setup for each test. VCCFinder produced 99% less false positive at same level of recall as compared to Flawfinder. Researchers released annotated VCC database which can be used as benchmark to compare future approaches with existing approaches.

Zou et al. [26] proposed a semantic based approach to detect vulnerable code clone. They represent the source code fragments in Program Dependency Graph (PDG). Ensuring the integrity of semantic information of PDG was not lost, the PDG was transformed into program feature tree by full path traversal. Presence of subtree corresponding to vulnerable code clone in program feature tree indicate its presence. This approach had shorter execution time as compared to other semantic based approaches to detect vulnerable code clone.

Yamaguchi et al. [25] proposed an approach to accelerate source code auditing by signifying the potentially vulnerable code fragments to an analysist. In this approach researchers used abstract syntax trees to represent code structures and then try to find the structural patterns.

With the help of these structural representations a known vulnerability is decomposed and extrapolated it to code base. Thus, making it possible to figure out potential vulnerabilities having same flaw that can be suggested to analysist. Researchers applied this approach on the source code of four popular open-source projects: Pidgin, LibTIFF, Asterisk and FFmpeg. They are successful in reporting zero-day vulnerabilities.

Table 2: Vulnerability Detection using code clone trend summary

Year and citation	Purpose of study	Match detection technique	Types of clone	Type of vulnerabilities
2010 Pham et al. [18]	an automatic tool SecureSync that can detect reoccurring software vulnerabilities	Hybrid technique	Type-1, Type-2	multiple
2012 Li and Ernst [13]	to report code that is semantically identical to buggy code	PDG subgraph matching	Type-3	multiple
2012 Jang et al. [7]	a scalable approach for vulnerable code clone detection	token based	Type-1, Type-3	multiple
2012 Yamaguchi et al. [25]	A tool to help analyst by suggesting potentially vulnerable functions	AST based	Type-3	multiple
2014 Li et al. [14]	a syntax based fast and scalable approach to detect vulnerable code clones and verifies them using concolic testing	token based	Type-1	overflow buffer

2015 Li et al. [11]	a mechanism to detect software vulnerability using code clone verification	token based	Type-1	buffer overflow
2016 Li et al. [12]	To detect whether a software code fragment contains a specified vulnerability	technique Hybrid	Type-1, 2, Type-3	Multiple
2016 Sajnani et al. [21]	To develop tool that can detect exact and near miss clones from large software project	Semantic Based	Type-3	Multiple
2017 Kim et al. [9]	A scalable approach to detect vulnerable code clones	Token based using Hashing	Type-1, Type 2	Multiple
2017 Liu et al. [15]	a vulnerability fingerprint model based Vulnerable code clones detection system	Hashing	Type-1, Type 2, Type-3	Multiple
2017 Zou et al. [26]	a semantic based approach to detect vulnerable code clone	Semantic based	Type-4	Multiple

Conclusion and future scope: In this paper we put a focus on an important research area of known vulnerability detection using code clone. Numerous applications and gadgets will continue to receive enormous numbers of vulnerable code fragments. We firmly believe that Vulnerability detection using code clone will be an essential strategy for safeguarding a variety of applications from bug propagation. This can be further applied to blockchain technology as well. smart contract are the building blocks for any blockchain based technology, as it is equally important to locate the smart contract with known vulnerability.

References:

[1]. Ross Anderson, "Security in Open versus Closed Systems - The Dance of Boltzmann, Coase and Moore", Open-Source Software: Economics, Law and Policy, Toulouse, France, June 20–21, 2002.

[2]. D. Chatterji, J. C. Carver, N. A. Kraft, and J. Harder, "Effects of cloned code on software maintainability: A replicated developer study," in: Proceedings of the 20th Working Conference on Reverse Engineering, WCRE, 2013, pp. 112–121.

[3]. W. Du and A. P. Mathur. Vulnerability Testing of Software System Using Fault Injection. Technical report, COAST, Purdue University, West Lafayette, IN, US, April 1998.

[4]. J. F. Islam, M. Mondal, and C. K. Roy, "Bug Replication in Code Clones: An Empirical Study," in: Proceedings of the 23rd International Conference on Software Analysis, Evolution, and Reengineering, SANER, 2016, pp. 68–78.

[5]. M. R. Islam and M. F. Zibran, "A Comparative Study on Vulnerabilities in Categories of Clones and Non-cloned Code," in: Proceedings of the 23rd International Conference on Software Analysis, Evolution, and Reengineering, SANER, 2016, pp. 68–78.

[6]. M. R. Islam, M. F. Zibran, and A. Nagpal, "Security Vulnerabilities in Categories of Clones and Non-Cloned Code: An Empirical Study," in: Proceedings of the ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM), 2017, pp. 20–29.

[7]. J. Jang, A. Agrawal, and D. Brumley, "ReDeBug: Finding Unpatched Code Clones in Entire OS Distributions," in: Proceedings of the IEEE Symposium on Security and Privacy, 2012, pp. 48–62.

[8]. L. Jiang, Z. Su, and E. Chiu, "Context-based detection of clone related bugs," in Proceedings of the 6th joint meeting of the European software engineering conference and the ACM SIGSOFT symposium on The foundations of software engineering, ser. ESECFSE '07. New York, NY, USA: ACM, 2007, pp. 55–64.

[9]. S. Kim, S. Woo, H. Lee, and H. Oh, "VUDDY: A Scalable Approach for Vulnerable Code Clone Discovery," in: Proceedings of the IEEE Symposium on Security and Privacy (SP), 2017, pp. 595–614.

[10]. X. Li, X. Chang, J. A. Board, and K. S. Trivedi, "A novel approach for software vulnerability classification," in: Proceedings of the Annual Reliability and Maintainability Symposium (RAMS), 2017, pp. 1–7.

[11]. H. Li, H. Kwon, J. Kwon, and H. Lee, "CLORIFI: software vulnerability discovery using code clone verification," Concurrency and Computation: Practice and Experience, vol. 28, no. 6, 2016, pp. 1900–1917.

[12]. Z. Li, D. Zou, S. Xu, H. Jin, H. Qi, and J. Hu, "VulPecker," in: Proceedings of the 32nd Annual Conference on Computer Security Applications - ACSAC '16, 2016, pp. 201–213.

[13]. J. Li and M. D. Ernst, "CBCD: Cloned buggy code detector," in: Proceedings of the 34th International Conference on Software Engineering (ICSE), 2012, pp. 310–320.

[14]. H. Li, H. Kwon, J. Kwon, and H. Lee, "A Scalable Approach for Vulnerability Discovery Based on Security Patches," in: Proceedings of the 5th International Conference on Applications and Techniques in Information Security, Melbourne, Australia, 2014; 109–122.

[15]. Z. Liu, Q. Wei, and Y. Cao, "VFDETECT: A vulnerable code clone detection system based on vulnerability fingerprint," in: Proceedings of the 3rd IEEE International Conference on Information Technology and Mechatronics Engineering Conference (ITOEC), 2017, pp. 548–553.

[16]. A. Nappa, R. Johnson, L. Bilge, J. Caballero, and T. Dumitras, "The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching," in: Proceedings of the IEEE Symposium on Security and Privacy, 2015, pp. 692–708.

[17]. H. Perl, D. Arp, S. Dechand, F. Yamaguchi, S. Fahl, Y. Acar, K. Rieck, and M. Smith, "VCCFinder," in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15, 2015, pp. 426–437.

[18]. N. H. Pham, T. T. Nguyen, H. A. Nguyen, and T. N. Nguyen, "Detection of recurring software vulnerabilities," in: Proceedings of the IEEE/ACM international conference on automated software engineering - ASE '10, 2010, p. 447-456.

[19]. D. Rattan, R. Bhatia, and M. Singh, "Software clone detection: A systematic review," Information and Software Technology, vol. 55, Issue 7, 2013, pp. 1165–1199.

[20]. C. K. Roy, M. F. Zibran, and R. Koschke, "The vision of software clone management: Past, present, and future (Keynote paper)," in: Proceedings of the IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE), 2014, pp. 18–33.

[21]. H. Sajnani, V. Saini, J. Svajlenko, C. K. Roy, and C. V. Lopes, "SourcererCC: Scaling Code Clone Detection to Big Code," in: Proceedings of the 38th International Conference on Software Engineering - ICSE '16, 2016, pp. 1157–1168.

[22]. W. Scacchi, "Understanding Open-Source Software Evolution," in Software Evolution and Feedback, Chichester, UK: John Wiley & Sons, Ltd, 2006, pp. 181–206.

[23]. K. Scarfone, P. Mell, and M. Souppaya, "Managing Software Patches and Vulnerabilities," in Computer Security Handbook, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2015, p. 40.1 - 40.11.

[24]. M. Shahzad, M. Z. Shafiq, and A. X. Liu, "A large scale exploratory analysis of software vulnerability life cycles," in: Proceedings of the 34th International Conference on Software Engineering (ICSE), 2012, pp. 771–781.

[25]. F. Yamaguchi, M. Lottmann, and K. Rieck, "Generalized vulnerability extrapolation using abstract syntax trees," in: Proceedings of the 28th Annual Computer Security Applications Conference on - ACSAC '12, 2012, p. 359-368.

[26]. D. Zou et al., "SCVD: A new semantics-based approach for cloned vulnerable code detection," in Lecture Notes in Computer Science, 2017, vol. 10327 LNCS, pp. 325–344.

[27]. "NVD - Home." [Online]. Available: https://nvd.nist.gov/. [Accessed: 17-Sept-2021].

Punjabi Text to Speech System for UNICODE and Non-UNICODE based Fonts

Charanjiv Singh Saroa*[1],Kawaljeet Singh*
Punjabi University, Patiala, Punjab, India - 147001
[1]cjsinghpup@gmail.com

ABSTRACT- In this paper, we discuss the need for regional languages for the person's overall development. The main focus of this paper is to create a speech system that can work on non-UNICODE based fonts. We also discuss various studies on text-to-speech converters for regional languages. In this paper, we discuss a method in brief, how we can create a TTS for regional languages, and what are the basic components we need to create an effective Text to speech system like font detector, font converter, spell checker etc. We also touch basics of Braille language, how it is helpful for blind persons.

INDEX TERMS- NLP, Text to Speech, UNICODE, Punjabi, Gurmukhi.

1 Introduction According to estimates, only 2% of India's 70 million disabled people have access to school. India frequently excludes disabled children from attending regular schools. Regional languages are strongly regarded to play a significant part in a person's life. According to the Government of India, India report of the 2011 Census, the count of the Visually Disabled population in India is 18.16% chart shows the classification of the disabled population by Gender wise of India [3]. According to World Health Organization (WHO), 285 million peoples are Visually Blind in the overall world, 39 million peoples are completely blind, and 246 million peoples have low vision [4]. Visually handicapped students can prove to be sensible if they are exposed to text and electronic documents in the regional languages in depth and in a reasonable time. Blind students cannot flip through the pages of a book, skim through the text or use a highlighter. Hence, it is presumed that natural language processing techniques can profitably assist blind students in meeting their academic objectives if supported using regional languages. The basic motivation is to enable access to education for visually impaired children by using ICT, so they may not remain excluded from social participation.

The contributions of this study are:

e. Collects well-reputed journals related to the review objectives.
f. Reviews the prior Text to Speech techniques.
g. Create for corpus of words and sounds.
h. To create a corpus of Punjabi words written in Gurmukhi script with its Braille equivalent and record sounds of Punjabi words and syllables by emphasising upon the Punjabi text and literature being considered prominently by visually impaired persons;
i. To develop a converter to automatically identify popularly used ASCII-based Punjabi fonts and convert them to UNICODE-based Punjabi fonts.

2 EXISTING STUDIES

Text-to-speech (TTS) is a special application that is used to create a spoken sound version of a computer document. TTS can help a visually challenged person to get information from the computer display information. Many TTS products are available [31], including Read Please, Proverb Speech Unit, and Next Up Technology's Text Aloud. Lucent, Elan, and AT&T each have a product called "Text-to-Speech.

Text-to-speech has benefits for all users; some of the specific groups are [30]:

- People with learning disabilities
- People who have literacy difficulties
- People who speak the language but do not read it
- People who multitask
- People with visual impairment
- People who access content on mobile devices
- People with different learning

An overview of some of the relevant research work is reproduced as below:

S. No.	Citation	Major Contribution
1.	Alexandre Trilla , Francesc Alias[13]	This paper discusses the need for sentiment analysis and how sentiment addition to TTS improves the efficiency of speech. The paper also discusses challenges in translating human effect into explicit representations and explains various machine learning principles.
2.	Bhavitha B K, Anisha P Rodrigues and Dr. Niranjan N Chiplunkar[14]	This paper focuses on several machine learning techniques which are used in analyzing sentiments and in opinion mining. This paper presents a detailed survey of various machine-learning techniques and then compares them with their accuracy, advantages and limitations of each technique. Also, discuss two approaches in sentimental analysis. One is by considering symbolic methods, and the other one by machine learning methods.

3.	Sayeda Swaleha Peerzade, Prof. Ramesh Bhat [15]	This paper concentrates on identifying and categorizing the sentiment present in the text, which is the input to the text-to-speech system. It also discusses how the sentence is pre-processed, and classification is done using the classifiers and sentiment tagged sentence is passed as input to the text-to-speech system, and the text-to-speech system selects the voice based on the sentiment and converts text to expressive speech.
4.	Eva Vanmassenhove, Joao P. Cabral and Fasih Haider[16]	In this paper, a method is discussed to predict the emotion from a sentence so that text can convey it through the synthetic voice. It consists of combining a standard emotion-lexicon-based technique with the polarity scores (positive/negative polarity) provided by a less fine-grained sentiment analysis tool in order to get more accurate emotion labels. The sentiWordsTweet tool is used to distinguish sentences with a positive or negative polarity. The paper also discusses the speech clustering method to select the utterances with emotion during the process of building the emotional corpus for the speech synthesizer.
5.	Manjunath K E, K. Sreenivasa Rao and Debadatta Pati [17]	This paper focuses on developing a Phonetic Engine for the Indian languages of Bengali and Oriya. The concept of two separate PEs for decoding the spoken utterances of Bengali and Oriya languages is discussed in this paper. Machine learning approaches such as Hidden Markov Models (HMMs), Feed Forward Neural Networks (FFNNs) and Support Vector Machines (SVMs) are used to build PE.
6.	Lincy Babykutty, Anu George and Leena Mary[18]	This paper explains how Phonetic Engine (PE) is a system that is used to determine the sequence of phones in a spoken utterance. This paper focuses on developing multilingual PE for four Indian languages, namely, Bengali, Hindi, Urdu and Telugu. For developing the PE, a read speech corpus has been used. The system discussed in this paper is based on Hidden Markov Models (HMM). The trained forty HMMs are used to derive a sequence of phonetic units from testing utterances.
7.	Manjunath K E and K. Sreenivasa Rao[19]	The machine learning approaches such as Hidden Markov Models (HMMs), Feed Forward Neural Networks (FFNNs) and Support Vector Machines (SVMs) are discussed with the use of each technique to drive Automatic Phonetic Transcription (APT).
8.	Dr. Surinder Dhanjal and Dr. Satvinder Singh Bhatia[20]	In this paper, a new corpus in the Punjabi language has been designed. This work concentrates only on the Malwai dialect of the Punjabi language. At least 20 special features of the new corpus have been described in this paper. A new term, extended pronunciation, has been used by the authors. The corpus consists of approximately 300 items.
9.	K. R. Aida-Zade, C. Ardil and A.M. Sharifova[21]	This paper contains the main principles of text-to-speech synthesis systems and discusses associated problems which arise when developing speech synthesis systems. This paper also explains the basic building block of TTS. Paper also discusses a brief history of text-to-speech converters. The paper also introduce some of the available TTS like Infovox, Infovox SA-101, DEC talk system, TTs developed by AT&T Bell Laboratories.
10.	Anand Arokia Raj, Tanuja Sarkar, Sathish Chandra Pammi, Santhosh Yuvaraj, Mohit Bansal, Kishore Prahallad and Alan W Black[22]	This paper explains how to build a natural-sounding speech synthesis system. In this paper, the issues of Font-to-Akshara mapping, pronunciation rules for Aksharas, and text normalization in the context of building text-to-speech systems in Indian languages are explained. Unicode and ASCII-based fonts are also discussed and the need for TTS system for ASCII-based fonts is also explained. TTS system that is explained in this paper works for Hindi, Tamil and Telugu.
11.	Priya and Amandeep Kaur Gahier[23]	In this paper, authors discuss basic topics like natural language procession (NLP), Text To Speech Synthesis, the Need of Text to Speech System, Challenges in Text to Speech System and previous work done on these topics.
12.	Parminder Singh and Gurpreet Singh Lehal[24]	In this paper, the authors discuss the development of a Text-To-Speech (TTS) synthesis system for the Punjabi language written in Gurmukhi script. And technique used to develop TTS. The concatenative method has been used to develop this TTS system. The paper also contains information about Syllables and how syllables help to create TTS.The paper also explains why syllables are important for Punjabi TTS. Punjabi is a syllabic

		language, so syllables have been selected as the basic speech unit for this TTS system, which preserves within-unit co-articulation effects. Paper explains development of algorithms for pre-processing, schwa deletion and syllabification of the input Punjabi text, as well as speech database for Punjabi.
13.	Ramanpreet Kaur and Dharamveer Sharma[25]	This paper deals with the improvement of eSpeak. eSpeak provides support for several languages, including Punjabi. This paper discusses some improvements in this formant-based text-to-speech synthesis system for Punjabi text input. After analysis of eSpeak for Punjabi input, some faults are identified and corrected by using eSpeakedit.
14.	Parminder Singh and Gurpreet Singh Lehal[26]	This paper discusses the results of the statistical analysis of Punjabi syllables over a large Punjabi corpus. The paper also explains the role and need of Syllables. TTS discussed in this paper also selects syllables for development. In this paper, Punjabi syllables have been statistically analyzed on the Punjabi corpus having more than 104 million words. The paper also explains how the efficiency of the text-to-speech (ITS) system is improved with the minimum set of syllables.
15.	Sheeba Grover and Dr. Amandeep Verma[27]	In this paper authors presented the design of the hybrid concept of Keyword based Approach and Machine Learning Algorithm for the detection of emotion from Punjabi textual data. Keyword based engine is used to detect whether the emotion is present in the input dataset or not and Machine Learning based classifier detect Ekman's six types of basic emotions (happiness, fear, anger, sadness, disgust and surprise). The proposed design is implemented in python with input Punjabi text in Unicode format. The paper also explains the steps involved in emotion detection.
16.	Perkins [7]	The World Braille Usage book is by UNESCO. This book contains all the languages that can be written using Braille. This book contains more than 100 languages with Braille code and its UNICODE number.
17.	Vandana, Nidhi Bhalla and Rupinderdeep Kaur[28]	The aim of the research paper is to convert the Gurmukhi script to Braille to help blind people for living a good life by learning well. This paper addresses the various aspects of the Braille script. It puts light on the origin and various levels of it, which depends on the type of user, such as simple user, moderate user and expert user. In this paper, the architecture of the Braille system is also explained. The main focus of the paper is on the conversion of Gurmukhi to Braille script.
18.	Vandana, Nidhi Bhalla and Rupinderdeep Kaur[29]	This paper explains the basics of Braille code, Braille sheet, Standards of cell, and Methodology for converting Gurmukhi to Braille. In the implementation character to character conversion is used to convert Gurmukhi script to Braille script.

Some paper related to Braille has also studied Braille as a physical form of reading and writing used by people who are blind or vision impaired. It was developed by Louis Braille in 1829. Braille is based on a six-dot cell with two columns of three, like the six on a dice[6]. By using these six dots, 63 different patterns can be formed. Braille letters do not have printed letters. Each letter may have one dot or a combination of dots. Bharati Braille or Bharatiya Braille or Indian Braille is a unified Braille script for writing the Indian languages. When India gained independence, almost 11 scripts for Braille were used. By 1951, Bharati Braille comes into existence and it is a standard for Indian languages. For the Punjabi language, Punjabi Braille is used to representing the Braille alphabet.[7] [8][9][10][11].

II IMLEMENTATION

Gurmukhi Font converter is created to convert Non- UNICODE Based fonts to UNICODE automatically. Then we create a corpus of Punjabi words written in Gurmukhi script and record the sounds of Punjabi words. In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays, usually electronically stored and processed). In corpus linguistics, they are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. In linguistics, a corpus (plural corpora) or text corpus is a large and structured set of texts (nowadays, usually electronically stored and processed). In corpus linguistics, courpuses are used for statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules within a specific language territory. These corpora also act as a base to create font detector and font converter, and spell checker.

Data is collected from various sources to create a corpus. Till now a total of 1961584 words have been processed to find 69637 unique words. These words are stored with the frequency of their occurrence in the text the following table is generated as a result.

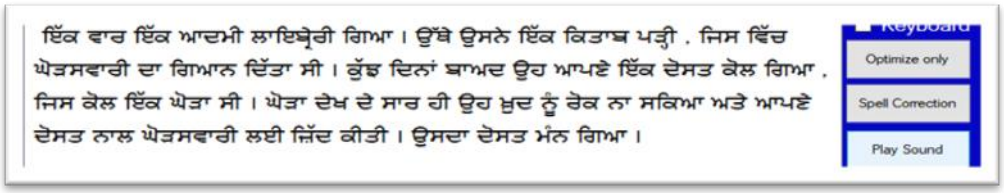
Sr No	Word	Frequency
1	ਹੈ	67366
2	ਵਿੱਚ	50244

3	ਦੇ	47605
4	ਨੂੰ	43829
5	ਦਾ	43163
--	--	--
--	--	--

Similarly, a collection of sounds is created in mp3 and ogg format that is used to create text to speech system.

ੳਚਿਆਈ.mp3 ੳਣਾ.mp3 ੳਧਮ.mp3
ੳਸਾਜਣਾ.mp3 ੳਹਾਂ.mp3 ੳਹੀ.mp3
ੳਗਮਣਾ.mp3 ੳਂਗਰ.mp3 ੳਂਗਰਨਾ.mp3
ੳਂਗਰਾ.mp3 ੳਗਰਾਹਕ.mp3 ੳਗਰਾਹਾ.mp3
ੳਗਰਾਹੀ.mp3 ੳਗਰਾਹੀਆ.mp3 ੳਗਰਾਹੁਣਾ.mp3
ੳਗਰਾਹ.mp3 ੳਂਗਲ ਰੱਖਣਾ.mp3 ੳਂਗਲ.mp3

With the help of a font detector, Font converter, Spellchecker, and sound and text corpus, we implement the Text To Speech system.



Following algorithm is used to convert Gurmukhi script to Speech.

1. Read Input.
2. If there is no input then go to end.
3. Convert NON-UNICODE Based fonts to UNICODE Based Font
4. Check Spells of input text.
5. Break the text into word.
6. Repeat for all the Words of input string
   If Word is Matched with the Database.
      Select the sound from the database and calculate delay time. & Play sound.
   Else
      Split the word into characters and play sound of each character
7. End

IV Conclusion and Future Research Direction

In this paper, we try to implement text to speech converter for the Punjabi Language written in Gurmukhi script. Especially this system is helpful for data written in non-UNICODE-based fonts. We will enhance this system to implement Braille in it. In future work, we first increase the corpus so that conversion from text to Braille becomes easy and fast. We also want to add intonations to the speech system.

REFERENCES

[1] MHRD, India, "Constitutional provisions relating to Eighth Schedule," Link: http://mha.nic.in/hindi/sites/upload_files/mhahindi/files/pdf/Eighth_Schedule.pdf Articles 344(1) and 351 of the Constitution. 2004.

[2] UNESCO, "The Improvement in the Quality of Mother Tongue - Based Literacy and Learning, Bangkok: UNESCO." 2008.

[3] Swaran Lata, Swati Arora, "Exploratory Analysis of Punjabi Tones in relation to orthographic characters : A Case Study," *Technology development for Indian Languages*, pp. 3–7.

[4] Harjeet Singh, Ravinder Khanna and Vishal Goyal, "Comparative Study of Standard Punjabi and Malwai Dialect with regard to Machine Translation," *An International Journal of Engineering Sciences*, vol. 8, pp. 109–118, June 2013.

[5] Kartar Singh Siddharth, Renu Dhir and Rajneesh Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features," *International Journal of Computer Science and Information Technologies*, vol. 2(3), no. 6, pp. 1036-1041, 2011.

[6] Farhan Bodale, Uddhav Bhide and Dilip Gore, "Braille Translation," *International Journal of Research in Advent Technology*,vol. 2, no. 4, p. 372-376, 2014.

[7] Perkins, "World Braille Usage," *National Library Service for the Blind and Physically Handicapped Library of Congress, UNESCO Washington, D.C.*, Third edition, 2013.

[8] Joga Singh, "INTERNATIONAL OPINION ON LANGUAGE ISSUES: Mother Tongue is the Key to Education, Knowledge, Science, and English Learning," pp. 1–22, 2013.

[9] Manzeet Singh and Parteek Bhatia, "Automated Conversion of English and Hindi Text to Braille Representation," *International Journal of Computer Applications*, vol. 4, no. 6, pp. 25–29, 2010.

[10] Nikisha B. Jariwala and Bankim Patel, "Conversion of Gujarati Text into Braille : A Review," *International Journal of Innovations & Advancement in Computer Science*, vol. 4, no. 1, pp. 59–64, 2015.

[11] Anupam Kumar Garg, "Braille-8 - The unified braille Unicode system: Presenting an ideal unified system around 8-dot Braille Unicode for the Braille users world-over," 2016 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), *Bangalore*, 2016, pp. 1-6.

[12] SS.Padmavathi,Manojna K.S.S, Sphoorthy Reddy .S and Meenakshy.D, "Conversion of Braille to Text in English, Hindi and Tamil Languages," *International Journal of Computer Science, Engineering and Applications*, vol. 3, no. 3, pp. 19–32, 2013.

[13] Alexandre Trilla and Francesc Alías, "Sentence-based sentiment analysis for expressive text-to-speech," *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 2, pp. 223–233, 2013.
[14] Bhavitha B K, Anisha P Rodrigues and Dr. Niranjan N Chiplunkar, "Comparative study of machine learning techniques in sentimental analysis," *International Conference on Inventive Communication and Computational Technologies*, pp. 216–221, 2017.
[15] Sayeda Swaleha Peerzade and Prof. Ramesh Bhat, "Categorization of Text into Appropriate Sentiment for Automatic Synthesis of Expressive Speech," *International Journal Of Engineering And Computer Science*, vol. 4, no. 5, pp. 12139–12142, 2015.
[16] Eva Vanmassenhove, Joao P. Cabral and Fasih Haider, "Prediction of Emotions from Text using Sentiment Analysis for Expressive Speech Synthesis," *9th ISCA Speech Synthesis Workshop , 2016, Sunnyvale, CA, USA*, 2016.
[17] Manjunath K E, K. Sreenivasa Rao and Debadatta Pati, "Development of phonetic engine for Indian languages: Bengali and Oriya," *International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*, 2013.
[18] Lincy Babykutty, Anu George and Leena Mary, "Development of Multilingual Phonetic Engine for Four Indian Languages," *International Conference on Next Generation Intelligent Systems (ICNGIS)*, 2016.
[19] Manjunath K E and K. Sreenivasa Rao, "Automatic Phonetic Transcription for read, extempore and conversation speech for an Indian language: Bengali," 2014 Twentieth National Conference on Communications (NCC), *Kanpur*,pp. 1-6,2014
[20] Dr. Surinder Dhanjal and Dr. Satvinder Singh Bhatia, "Development of a standard text and speech corpus for the Punjabi language," *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*, 2013.
[21] K. R. Aida-Zade, C. Ardil and A.M. Sharifova, "The main principles of text-to-speech synthesis system," *International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:7,* vol. 7, no. 1, pp. 395–401, 2013.
[22] Anand Arokia Raj, Tanuja Sarkar, Sathish Chandra Pammi, Santhosh Yuvaraj, Mohit Bansal, Kishore Prahallad and Alan W Black, "Text Processing for Text to Speech Systems in Indian," *Proceedings of 6th ISCA Speech Synthesis Workshop SSW6, Bonn, Germany, 2007*,pp. 188-193 2007.
[23] Priya and Amandeep Kaur Gahier, "Text to Speech Conversion in Punjabi-A Review," *IJCTA*, vol. 9, no. 41, pp. 373–379, 2016.
[24] Parminder Singh and Gurpreet Singh Lehal, "PunjabiText-To-Speech Synthes is System," Preceeding of COLING 2012, Mumbai, pp. 409–416, 2012.
[25] Ramanpreet Kaur and Dharamveer Sharma, "An Improved System for Converting Text into Speech for Punjabi Language using eSpeak," International Research Journal of Engineering and Technology, vol. 3, no. 4, 2016.
[26] Parminder Singh and Gurpreet Singh Lehal, "Statistical syllables selection approach for the preparation of Punjabi speech database," 2010 International Conference for Internet Technology and Secured Transactions, London, 2010, pp. 1-4, 2010.
[27] Sheeba Grover and Dr. Amandeep Verma, "Design for Emotion Detection of Punjabi Text using Hybrid Approach," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, pp. 1-6, 2016.
[28] Vandana, Nidhi Bhalla and Rupinderdeep Kaur "Architecture of Gurmukhi to Braille conversion system," International Journal of Computer Science and Information Technology & Security, vol. 2, no. 2, pp. 467–471, 2012.
[29] Vandana, Nidhi Bhalla and Rupinderdeep Kaur, "Implementation of Gurmukhi to Braille," International Journal of Computer Science and Technology , vol. 3, no 2, pp. 596–599, 2012.
[30] Benefits of Text to Speech, http://www.readspeaker.com/benefits-of-text-to-speech/
[31] TTS products , https://www.understood.org
[32] Text-To-Speech(TTS), http://searchmobilecomputing.techtarget.com
[33] Languages of India, http://www.gutenberg.us/articles/Languages_of_India
[34] Education, http://www.ncpedp.org/Education
[35] Role of ICT, http://wikieducator.org/Vital_role_of_ict_for_disabled_children
[36] Use of Computer Technology to Help Students with Special Needs, https://www.princeton.edu/futureofchildren/publications/docs/10_02_04.pdf

FAKE USER ACCOUNTS DETECTION ON WEB SERVICES

Rajdavinder Singh Boparai[1], Dr. Rekha Bhatia[2]
[1]*Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India.*
[2]*Department of Applied Management, Punjabi University, Patiala, Punjab, India.*
[1]rajiboparai@gmail.com
[2]r.bhatia71@gmail.com

Abstract*— Web services are attracting millions of users throughout the world and their presence on the web has also affected their living style. As dependency of customers is increasing every day to get service from web services such as e-commerce, food, education, entertainment etc. has led to different problems such as presence of fake user or user accounts on the same platform, which may mislead genuine customers to take correct decision. In this study a classification technique is proposed to detect fake user accounts or profiles present on web services. Experiments depicts the procedure involved with gathering data about the records of deceptive clients, featuring highlights, incorporating a dataset for preparing the classifier. For user profiling attributes such as user id, e-mail id, phone number, profile picture, number of reviews, purchase history, information of social profiles and profile details were referred. *The outcome of this research paper is presented in the form of accuracy to detect fake user profiles in comparison with the other classification techniques.

***Keywords:* Fake user profiles, fake user accounts, webservices, deceptive accounts.**

I. INTRODUCTION

Utilisation of the web services is increasing every hour by customers because very attractive offers proposed my companies as compare to physical service model in the form cashback, discounts and service at door step etc. [1]. In order to efficient utilization e-services companies are recommending users to create web accounts and profiles. It has been observed that along with genuine users some fake users also have been noticed on the same platform which is a dangerous thing. Our research is revolving around the major challenge of webservices that is to identify or detect fake account.

Despite the fact that, utilization of webservices are acquiring all-inclusive notoriety yet it brings number of safety and protection challenges like spam, trick, phishing, clickjacking irritating or following an individual or a gathering, slander, data fraud, outsider individual data divulgence and so forth Since client's close to home, proficient, social and political information is jumbled at a solitary spot which similarly draws in digital hoodlums towards these webservices which can be exceptionally hurtful for the two clients just as specialist organizations. These cybercriminals use data fraud assaults, making counterfeit profiles or dispatching mechanized creeping against various well known long range interpersonal communication destinations. Different purposes behind making counterfeit profiles incorporate promoting and crusading, slandering an individual, social designing, fun and amusement, information assortment for research/particular advertising, counterfeit traffic for sites or sites and so forth [2].

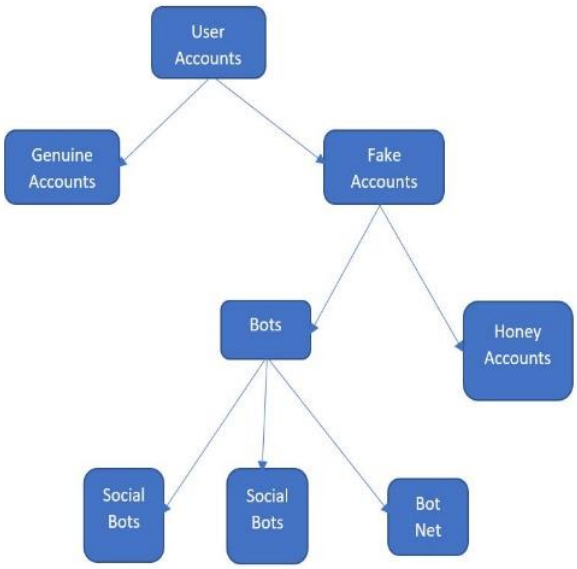


Figure 1: Organization of genuine/fake user accounts on webservices

For the most part the point of these digital lawbreakers is to take the client's close to home, proficient, political, social or monetary data by uncovering the clients with undesirable data on the web to mislead them [3]. There are number of strategies by which the clients' information can be hacked by these foes, and making counterfeit profiles to perform pernicious exercises on webservices is one of the generally utilized techniques. According to the clients' perspective, individual, proficient and surprisingly monetary information is no safer. Figure 1 gives a speedy perspective on different sorts of phony profiles and a few different sorts of profiles found in various web-based informal communities. Genuine

profiles must be classified into compromised and non-compromised ones which are additionally displayed in the figure 1. Profiles [4][5][6] which observe the guidelines and guidelines given by specific webservices administration are genuine. Here rules also, guidelines with regards to web services might mean the proprietor ought not have more than one individual account, it ought not to spread any unlawful, misdirecting, vindictive, or biased substance, and it ought not gather the client's data or access Facebook via computerized implies. An individual dealing with more than one record; for example, a record other than his head account is classified as fake [7][8]. A few choices to improve the security of client accounts like securing the secret key and sending area explicit login alarms and area cautions. Clients can likewise utilize the additional security elements of the organization like how to logout from another gadget, instructions to protect account. To stay away from digital assaults [9][10], one should take appropriate consideration while utilizing on the web social record. Likewise at the hour of record creation, the terms and conditions ought not be abused.

II. RELATED WORK

Awasthi et al. [11] have published the critical review on fake accounts on the different social websites and suggested the different techniques to identify the fake accounts. They have identified the accounts on twitter and facebook. Romanov et al.[12] also published a review of fake account detection and reported the identification of false accounts and its impact on the uses. Apart from this they have also reported the shortcoming in various research published to identify and implement techniques.

Singh et al. [13] carried out research using the machine learning method to identify the fake or false accounts on social media. They have used the techniques based on the number of followers and number of friends as this data is easily available on the social sites without violation of rule of right. As of today, on social media number accounts are available in which have no follower and no friends and such type of accounts are too active in social media and so many accounts are reporting them as fake account.

Kadam and Patidar [14] detected the fake accounts based on the content and attribute estimation in accounts. Fake accounts are mostly used for the abuses and passing the wrong information as they are difficult to detect. They have proposed the methods to identify and analyze the fake accounts.

Rao et al. [15] used the machine learning and natural language processing method to check the fake accounts on the various social networks and suggested to improve the technique to identify such type of accounts. The algorithm which they proposed that is easy and accurate to identify the fake accounts as per their methods.

Rao et al. [16] also used the machine learning techniques to find out the fake accounts on various social media sites. They have reveled that the social sites have great impact on daily life in now a days and fake accounts have also some negative impact on the human. They collected the data from various platforms and analyzed through the algorithm of machining learning techniques and complete process analyzed critically to identify and authentication of non existing or fake accounts. The have used the SVM, RF and Neural Network and reported the strong performance detect the fake accounts. Further they suggested that same dataset and techniques can be used find non real accounts on facebook and twitter sites.

As the machine learning have great impact to identify the fake accounts on the basis of this Rohit [16] also used the machine learning for same process. It was reported that from public domain, daily 1000 fake accounts are created on various social websites. Along with other impact on public these fake accounts also increase the burden of network and data storage space which further ruined in many terms. They have used SVM and CNB to validate and analysis the data which showed around 97% in SVM and 95% accuracy in CNC when it has been implemented on the facebook.

III. PROPOSED METHODOLOGY

Aim of this research to identify fake user accounts using machine learning by understand features/attributes of user accounts, highlighting important features, data cleanliness, model formation followed by performance evaluation of results.

A. *Data Collection*

To gather the work-explicit information from these interpersonal organizations a large portion of the analysts compose their own code utilizing APIs to associate with the designated administrations. Pretty much every person-to-person communication site has its own API for instance for example GRAPH API 10 for facebook which permits its clients to interface with their application and gather client data. Additionally for Twitter there is Twitter API. A few analysts and researchers plan their own information crawlers to extricate information explicit to their exploration from interpersonal organizations. A concentrate in has extricated information from Facebook and Twitter networks to recognize spams in these two interpersonal organizations. In, Authors have composed a content to get associated with currently made honey profiles and removed all the data expected to recognize the malevolent exercises.

B. *Data Cleaning*

Information cleaning is the most common way of fixing or eliminating erroneous, debased, mistakenly designed, copy, or fragmented information inside a dataset. When joining various information sources, there are numerous chances for information to be copied or mis-labeled. In the event that information is wrong, results and calculations are problematic, despite the fact that they might look right. There is nobody outright method for recommending the specific strides in the information cleaning process on the grounds that the cycles will differ from dataset to dataset. Be that as it may, it is vital

to build up a layout for your information cleaning process so you realize you are doing it the correct way without fail. Some steps were performed to extract specific information such as removal of duplicate data values, error fixation, complete missing values and data validation.

C. Feature Selection

Highlight choice is the method involved with lessening the quantity of info factors when fostering a prescient model. It is alluring to diminish the quantity of information factors to both lessen the computational expense of displaying and, at times, to work on the presentation of the model. Factual based element choice strategies include assessing the connection between each info variable and the objective variable utilizing insights and choosing those information factors that have the most grounded relationship with the objective variable. These techniques can be quick and viable, albeit the decision of factual measures relies upon the information kind of both the info and result factors. For user profiling some features were identified which are mentioned in Table I.

TABLE I
Features for user profiling

Attribute/ Feature	Description	Value
user_id	Account id of user	Alpha/ Numeric/Special symbols
email_id	e-Mail of user	e-Mail
phone_number	Phone no	Number
profile_picture	Image	Image
number_of_reviews	Total number of revies posted	Number
purchase_history	Item listing	Links
social_profiles	Profile names, IDs, links	Links
profile_details	Detailed description of profile	Alpha/numeric/special symbols/images/links

These types of clients are typically focused on business goal. This segment arbitrarily chosen a portion of the phony clients from the dataset and physically concentrated on them. Along these lines, an examination of the contrasts among phony and real clients from both substance and conduct perspectives is talked about as follows:

- Most phony records don't contain countless posts; a considerable lot of them have zero or just few posted pictures and recordings.
- Practically all phony clients follow a large number of authentic clients. The fundamental point of phony records is to increment or lessening deals of a particular item, advance or stigmatize somebody and so on

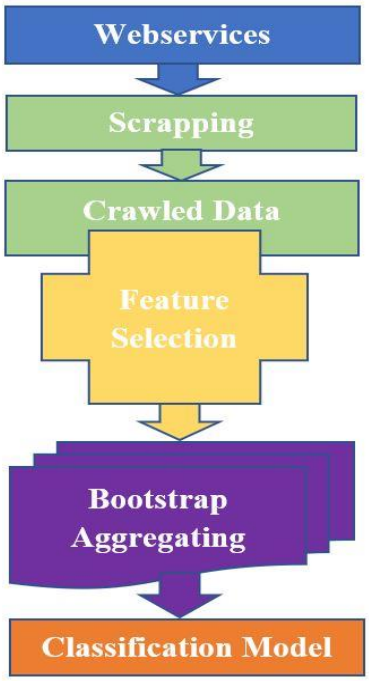


Figure 2: Block diagram of fake account detection model.

- In the wake of investigating, it has been seen that the impressive part of phony clients didn't set an image to their profile, and furthermore their profile doesn't contain a history portrayal more than the name they entered during the time spent making the record.

D. Model formation using classification

To identify fake user accounts a model is presented in this section with the help of machine learning using bootstrap aggregating. As we know bagging, can work on the exactness of your models and empowering you to grow better bits of knowledge. In a noisy dataset, to minimize variance bootstrap aggregation is being used very often. An irregular example of information in a preparation set is chosen with substitution—implying that the singular information focuses can be picked at least a few times.

After a few information tests are produced, these frail models are then prepared autonomously, and relying upon the kind of undertaking—relapse or arrangement. This model works in their sub steps namely bootstrapping, parallel training and aggregation. Overview of presented model to detect fake user accounts present on webservices is presented in figure 2.

IV. PERFORMANCE EVALUATION

Performance of proposed model is presented in this section to detect the fake user accounts. Classification techniques such as super vector machine, naïve bayes and random forest tree were used to evaluate, test and verify performance of the proposed model. Accounts classification is presented as true fake accounts, false fake accounts, true fake accounts for genuine accounts and false fake accounts for genuine accounts. Further to present the outcomes of model mathematically accuracy, recall and precision are used, formulas are represented in equation number 1, 2 and 3.

$$Accuracy = TP + TN\ TP + TN + FP + FN \quad (1)$$

$$Recall = TP / (TP + FN) \quad (2)$$

$$Precision = TP / (TP + FP) \quad (3)$$

Where, TP = TruePositives, TN = TrueNegative, FN = FalseNegatives and FP = FalsePositives.

Further table II and table III; figure 3 and figure 4 is demonstrating comparative performance to proposed model and popular classifiers such as SVM, NB and RFT. Results has shown that proposed model as yield better accuracy as compared to standard algorithms.

TABLE II
Accuracy of proposed model comparative to popular classifiers

Classifiers	Accuracy
SVM	62.3%
NB	93.3%
RFT	94.0%
Proposed Model	96.8%

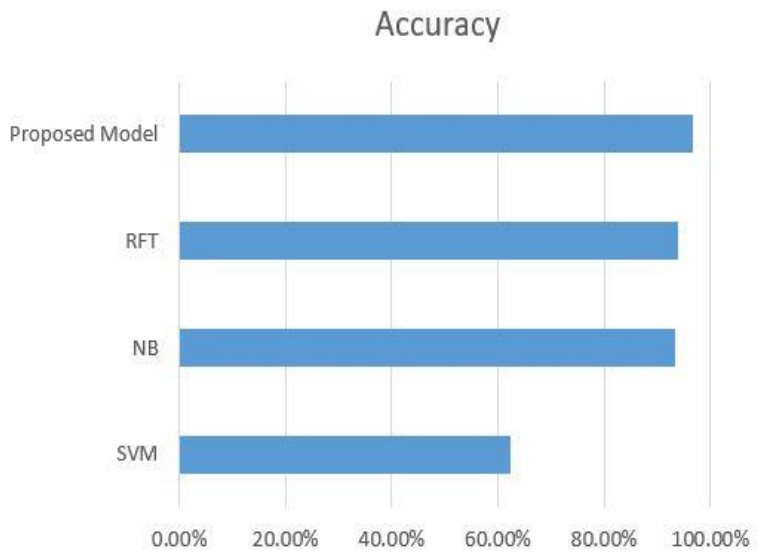


Figure 3: Accuracy of proposed model

TABLE III
Comparison of proposed model with traditional classifiers using precision and recall

Classifiers	Recall		Precision	
	Genuine Accounts	Fake Accounts	Genuine Accounts	Fake Accounts
SVM	76.3%	62.1%	89.1%	72%
NB	95.5%	87.3%	93.9%	95.3%
RFT	95%	93.3%	96.5%	85.6%
Proposed Model	96.1%	98.4%	98.9%	95.3%

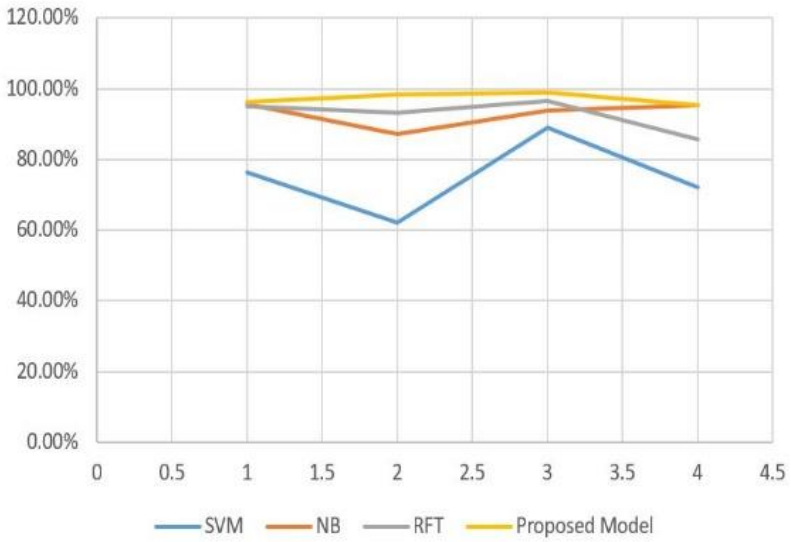


Figure 4: Performance of proposed model using precision and recall

V. CONCLUSION AND FUTURE INVESTIGATIONS

Customer dependency is growing up every day to get them served by web services instead of the traditional service industry. Along with genuine customers, some fake customers have also been noticed on the same platform which is harmful to both i.e., good customers and industries. We have proposed a model based on classification techniques to detect

fake user accounts/profiles present on web services. We have used attributes such as user id, e-mail id, phone number, profile picture, number of reviews, purchase history, information of social profiles, and profile details to find out fake user accounts available on web services. Experiments have yielded good results to justify the performance of the proposed model in the form of accuracy, which is better as compared to popular classifiers such as super vector machine, naïve bayes and random forest tree.

VI. REFERENCES

[1] D. L. Hoffman and T. P. Novak, "Why Do People Use Social Media? Empirical Findings and a New Theoretical Framework for Social Media Goal Pursuit," *SSRN Electron. J.*, 2012.

[2] H. Kwak, C. Lee and H. Park, "What is Twitter, a social network or a news media?," in *Proceedings of the 19th international conference on World Wide Web*, 2010, pp.591–600.

[3] Stein, Tao, E. Chen and K. Mangla, "Facebook immune system," in *Proceedings of the 4th Workshop on Social Network Systems. ACM*, 2011.

[4] Zangerle, Eva, and G. Specht, "Sorry, I was hacked: a classification of compromised twitter accounts," *Proceedings of the 29th Annual ACM Symposium on Applied Computing. ACM*, 2014.

[5] M. Fire, G. Katz, and Y. Elovici, "Strangers intrusion detection-detecting spammers and fake proles in social networks based on topology anomalies," *Proceedings in HUMAN*, vol. 1, no. 1, pp 26-32, 2012.

[6] P. Sowmya and M. Chatterjee, "Detection of Fake and Clone accounts in Twitter using Classification and Distance Measure Algorithms," *2020 International Conference on Communication and Signal Processing (ICCSP)*, 2020, pp. 0067-0070.

[7] A. Balestrucci and R. De Nicola, "Credulous Users and Fake News: a Real Case Study on the Propagation in Twitter," *2020 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS)*, 2020, pp. 1-8,

[8] M. Sevi and İ. Aydin, "Detection of Fake Twitter Accounts with Multiple Classifier and Data Augmentation Technique," *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, 2019, pp. 1-6.

[9] S. Liu, B. Hooi and C. Faloutsos, "A Contrast Metric for Fraud Detection in Rich Graphs," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2235-2248, 1 Dec. 2019.

[10] M. Kolomeets, O. Tushkanova, D. Levshun and A. Chechulin, "Camouflaged bot detection using the friend list," *2021 29th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*, 2021, pp. 253-259.

[11] S. Awasthi, S. R. Jena, and A. Srivastava, "Review of Techniques to Prevent Fake Accounts on Social Media," no. August, 2020.

[12] A. Romanov, A. Semenov, O. Mazhelis, and J. Veijalainen, "Detection of fake profiles in social media: Literature review," *WEBIST 2017 - Proc. 13th Int. Conf. Web Inf. Syst. Technol.*, pp. 363–369, 2017.

[13] N. Singh; T. Sharma; A. Thakral and T. Choudhury, "Detection of Fake Profile in Online Social Networks Using Machine Learning," *International Conference on Advances in Computing and Communication Engineering (ICACCE), 2018.*

[14] N. Kadam and H. Patidar, "Social Media Fake Profile Detection Technique Based on Attribute Estimation and Content Analysis Method," *Int. J. Recent Technol. Eng.*, vol. 8, no. 6, pp. 4534–4539, 2020.

[15] P. Srinivas Rao and J. Gyani, "Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP 1," *Int. J. Appl. Eng. Res.*, vol. 13, no. 6, pp. 4133–4136, 2018..

[16] S. Gutha, "Detecting fake account on social media using machine learning detecting fake account on social media using machine," no. April, 2020.

STATISTICAL KEYFRAME EXTRACTION TECHNIQUE BASED ON DIFFERENCE OF ENERGY AND ENTROPY OF FRAMES

Sumandeep Kaur[1], Dr. Madan Lal[2], Dr. Lakhwinder Kaur[3]
Assistant Professor [1,2], Professor [3]
Department of Computer Science & Engineering, Punjabi University, Patiala
Email Id: sumandhanjal@yahoo.com

Abstract
Key Frame Extraction from video plays an important role in many applications, such as video summarisation, content-based image retrieval and object detection from video. Key frame extraction techniques not only remove duplicate frames from the video but also maintain the useful information of the video. While working on real-time systems, it is important to scrutinize the frames from the video to reduce time and power consumption. It leads to the development of various effective key frame extraction techniques. This paper presents keyframe extraction methods based on the threshold of absolute difference of Energy and Entropy of consecutive video frames. KTH database is used for experimentation. The parameter used for evaluation is the Compression Ratio. Visual and quantitative results show reasonably better performance with a threshold based on the absolute difference of Energy and Entropy than a threshold based on the absolute difference of histograms of consecutive frames.

Keywords: *Key Frame Extraction, Absolute Difference, Histogram, Entropy, Energy.*

I. Introduction

In recent years, the number of videos on the internet and social media has been drastically increasing daily. This tremendous increase in data demands an efficient and effective technique for content-based image retrieval, indexing and data storage. But, the various techniques present cannot serve the purpose [1]. The reason behind this is that the nature of videos is not the same, and traditional methods are not sufficient for indexing, retrieving and storing data. So, there is a great demand to develop an efficient and effective technique that can fulfil the purpose of video data management. Video data is most effective in expressing data as it contains graphics among all the other kinds of data like text, audio, images etc. However, video processing is time-consuming as every video is unstructured and contains a lot of redundant data. So, this duplicate data needs to be removed by maintaining the useful information of the video. This process of removing the redundant frames from video is called key frame extraction. Keyframes are video frames containing important and meaningful information about the video data [2]. These keyframe extraction techniques reduce the size of video data, which is useful in video indexing, retrieval and browsing. It also provides a systematic flow of video data [3]. Moreover, the keyframe extraction technique is used as a pre-processing method in various fields like object detection and tracking.

Many researchers have worked in the field of video summarization, which is used in video surveillance, social media, YouTube and news. Key frame extraction is an important step in video summarization. Key frame extraction techniques can be classified as sequential-based approach and clustering-based approach. The first approach uses various visual features like color, variation, mean etc. and temporal information of different frames. In this, considerable variation in various visual features of different frames is taken to select keyframes. In a later approach, the different clusters of similar frames are formed which is based on some similarity between them. Then, from each cluster, some frames are selected which represent the cluster. The clustering technique is better as it generates less redundant data in comparison to the sequential approach, but at the same time, the clustering technique has drawbacks as it may not preserve the temporal order of frames of the video [4]. Keyframes can be static and dynamic. Frames which are extracted from video holding vital information are called static frames. These are used in static video summarization, whereas in dynamic video summarization, dynamic frames are extracted from videos, maintaining the temporal order of the video.

This paper represents a key frame extraction method by computing a threshold which is based on the absolute difference of consecutive frames based on Energy and Entropy. Results are compared with the existing technique which is using thresholding of absolute difference of histograms of consecutive frames of video. The paper is organized into five sections. Section 1 contains an introduction to the keyframe extraction method. Section 2 presents the existing work done in this field. Section 3 elaborates on the techniques used for key frame extraction. Results and conclusion are presented in section 4 and section 5, respectively.

II. Literature Survey

Wolf et al. performed a keyframe extraction method in which motion metric is calculated by computing optical flow using the Horn and Schnuck technique. Then, local minima are identified from the motion metric to select the keyframes [5]. Markos Mentzelopoulos et al. proposed an algorithm that computes spatial frame segmentation based on entropy difference. It performs very well where the background and object can be clearly distinguished. However, it fails to identify transient changes in an image [6]. Image information entropy and edge matching rate are used to extract keyframes. The information entropy of each frame is calculated to choose candidate frames, and then the Prewitt operator is used to find the edges of the candidate frame. Edges of adjacent frames are matched to eliminate redundant frames [7]. Two visual features, colour and edge profile features, are calculated to find interframe differences. The adaptive threshold technique calculates keyframes based on inter-frame differences [8]. Mutual information [9] difference between neighbouring frames is used to create shots from video. Keyframes are extracted from clusters, which are formed on the basis of the average difference of mutual information in video shots. The absolute difference of histograms [10] is used for

key frame extraction. The threshold is calculated by summing up absolute means and standard deviation of consecutive frames. Frames are selected by comparing the threshold with the absolute difference of histograms of consecutive keyframes. A fast and robust keyframe extraction method [11] which is works on low-level features like color and structure frame difference. Adjacent frames' color difference is used to create an alternative sequence, which is further matched with adjacent frames based on the structure difference of frames, followed by optimization to ensure the effectiveness of the video. Chen et al. [12] used two new gradient features, namely the Histogram of Gradient magnitude (HGM) and Weighted Average Gradient (WAG), which are based on color histograms. A modified Euclidian distance is used for the similarity check. The unsupervised clustering method [13] is used to divide frames into different clusters. From different clusters, frames are extracted based on a threshold calculated by statistical method. The changes in three visual features [14] color histogram, wavelet statistics and edge direction histograms, are measured for each descriptor to find frame difference, which is further used for the dynamic selection of keyframes within each shot. Abbas Rashidi et al. [15] proposed an optimized technique for the selection of keyframes. Often, video files are filled with blurriness, noise and redundant frames as the frame rate of the camera and speed is higher than required. In this algorithm, first of all blurred frames are removed. To remove these frames, threshold is obtained by maintaining minimum level of quality frames required. Then optimum keyframes are selected from the remaining keyframes by using six significant factors: high-quality frame extraction, determining sufficient overlap between adjacent frames, determining baseline length, data degeneracy avoidance, uniform distribution of features in each frame and optimization of the number of extracted key frames.
The related work of key frame extraction reveals that in some cases, a predefined number of keyframes is required, whereas in others, temporal order is to be maintained while calculating keyframes. Standard deviation and mean to determine threshold [10] don't require a predefined number of keyframes and also keep the temporal order of frames. Energy and entropy differences are better than histogram differences of frames as these features measure uncertainties and randomness in data.

III. Key Frame Extraction Techniques

This algorithm is based on the absolute difference of histograms of consecutive keyframes. It works in two phases. In the first phase, the threshold is calculated by adding the mean and standard deviation of the absolute difference of histograms of consecutive keyframes [10]. In the second phase, keyframes are extracted by comparing the absolute difference of the histogram against the threshold. The threshold is calculated as:

$T=\mu_{adh} + \sigma_{adh}$ where μ_{adh} is the mean of the absolute difference of histograms of frames and σ_{adh} is the standard deviation of the absolute difference of histograms of frames.

The same algorithm [10] has been implemented based on the absolute difference of Energy of consecutive frames as described in algorithm 2. It works in two phases. In the first phase, the threshold is calculated by adding the mean and standard deviation of the absolute difference of Energy of consecutive keyframes. In the second phase, keyframes are extracted by comparing the absolute difference of Energy of consecutive frames against the threshold.

The same algorithm [10] has also been implemented based on the absolute difference of Entropy of consecutive keyframes, as mentioned in algorithm 3. It also works in two phases. In the first phase, the threshold is calculated by adding the mean and standard deviation of the absolute difference of Entropy of consecutive frames. In the second phase, keyframes are extracted by comparing the absolute difference of Entropy of consecutive frames against the threshold. All these algorithms are implemented in MATLAB.

Algorithm 1: Keyframe extraction using absolute difference of *Histograms* of consecutive frames [10]

Step 1: Extract all frames from the input video.

Step 2: Calculate the absolute difference between histograms of two consecutive frames.

Step 3: Compute threshold as $T=\mu_{adh} + \sigma_{adh}$ (where adh=absolute difference of Histograms of consecutive frames)

Step 4: Compare difference with T. If difference >T, select it as keyframe; else, go to step 2.

Step 5: Continue till the end of the video.

Algorithm 2: Keyframe extraction using the absolute difference of *Energy* of consecutive frames

Step 1: Extract all frames from the input video.

Step 2: Calculate the absolute difference between the Energies of two consecutive frames.

Step 3: Compute threshold as $T=\mu_{ader}+ \sigma_{ader}$ (where ader=absolute difference of Energy of consecutive frames)

Step 4: Compare difference with T. If difference >T, select it as the keyframe; else, go to step 2.

Step 5: Continue till the end of the video.

Algorithm 3: Keyframe extraction using the absolute difference of *Entropy* of consecutive frames

Step 1: Extract all frames from the input video.

Step 2: Calculate the absolute difference between entropies of two consecutive frames.

Step 3: Compute threshold as $T=\mu_{aden}+ \sigma_{aden}$ (where aden=absolute difference of Entropy of consecutive frames)

Step 4: Compare difference with T. If difference >T, select it as keyframe; else, go to step 2.

Step 5: Continue till the end of the video.

IV. RESULTS AND CONCLUSION

These experiments are conducted on the KTH action database [16]. This video database contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3 and indoors s4 as illustrated below. Currently, the database contains 2391 sequences. All sequences were taken over homogeneous backgrounds with a static camera with a 25fps frame rate. The sequence was down-sampled to the spatial resolution of 160x120 pixels, with an average length of four seconds. Fig.1 shows an illustration of the KTH action database. Some frames of the KTH database under the class running are shown in Fig. 2

Fig.1 KTH action database illustration

To evaluate the performance, the compression ratio is computed. It is the study of the compactness of shot content. The higher value of the compression ratio indicates that the method is good. The compression ratio is computed using the equation.

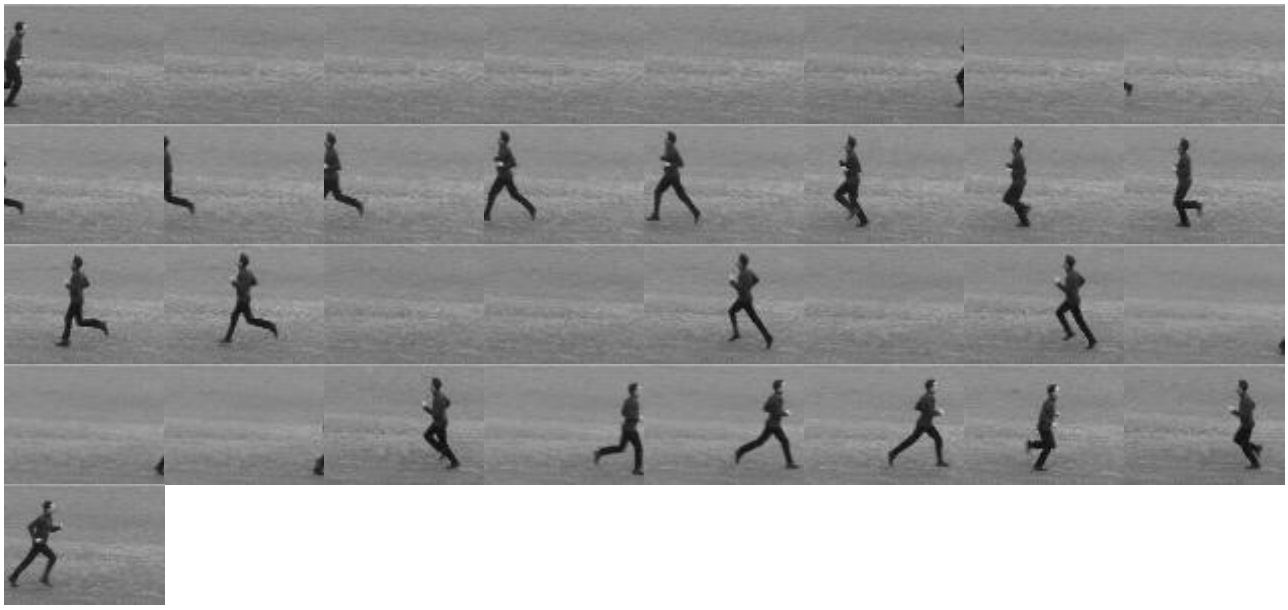


Fig. 2 Sample frames of class running.

CR=Total number of frames in video shot/number of keyframes selected (2)

Table 1 gives the number of keyframes detected and the compression ratio (CR) of selected video data in the KTH action database. It is performed on the frames obtained from video data. Keyframes obtained for video person01_running_d1_uncomp.avi are shown in Fig. 3-5.

Table 1 Results of Compression Ratio for selected videos in the database

Video		Histogram Difference		Energy Difference		Entropy Difference	
	TNF	NKF	CR	NKF	CR	NKF	CR
Person01_running_d1_uncomp.avi	335	59	5.6780	48	6.9792	35	9.5714
Person01_running_d2_uncomp.avi	365	50	7.3000	36	10.1389	43	8.4884
Person01_running_d3_uncomp.avi	350	54	6.4815	56	6.2500	35	10.000
Person02_running_d1_uncomp.avi	314	52	6.0385	50	6.2800	43	7.3023
Person02_running_d2_uncomp.avi	1492	203	7.3498	202	7.3861	187	7.9686

TNF-Total number of Frames, NFK-Number of Ker Frames, CR-Compression Ratio

An automated method for extracting keyframes for video summarization is presented in this. The role of keyframe extraction is to reduce redundant frames that can lead to dimensionality reduction of the feature vector for classification. Directly representing the video sequence by all the frames containing redundant and indiscriminate information would confuse the classifier in action recognition. In this, it is seen that the algorithm is able to compute the keyframes using simple calculations of the histogram of the absolute difference of consecutive frames in video data [10]. Then, this is modified by using the absolute difference of Energy and Entropy. It is shown that results obtained by entropy difference calculation are best. Even results obtained by energy difference are better than histogram difference. The compression ratio values show that the results obtained are reasonably accurate.

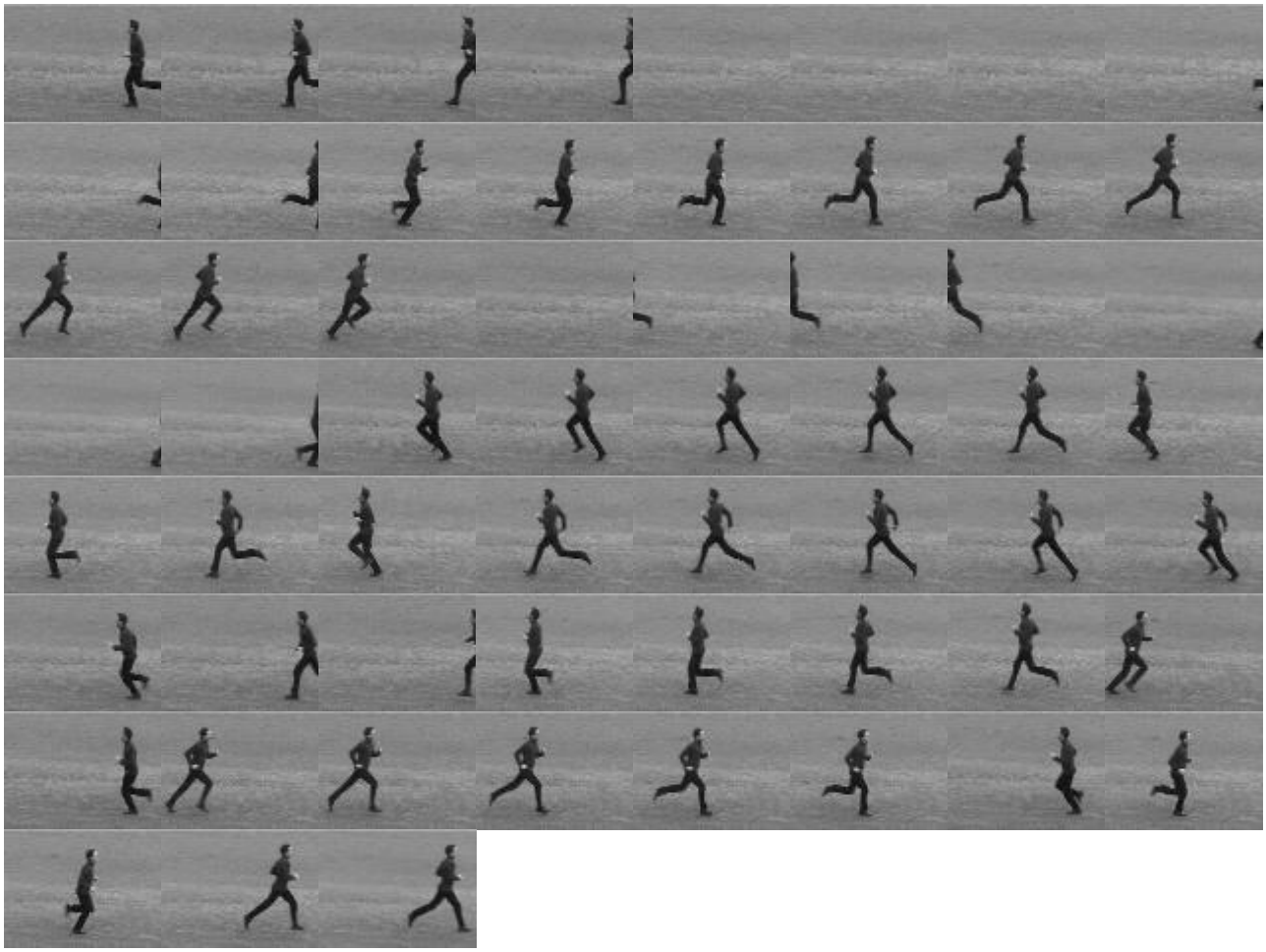


Fig 3. Keyframes (59) of data person01_running_d1_uncomp.avi (histogram difference)

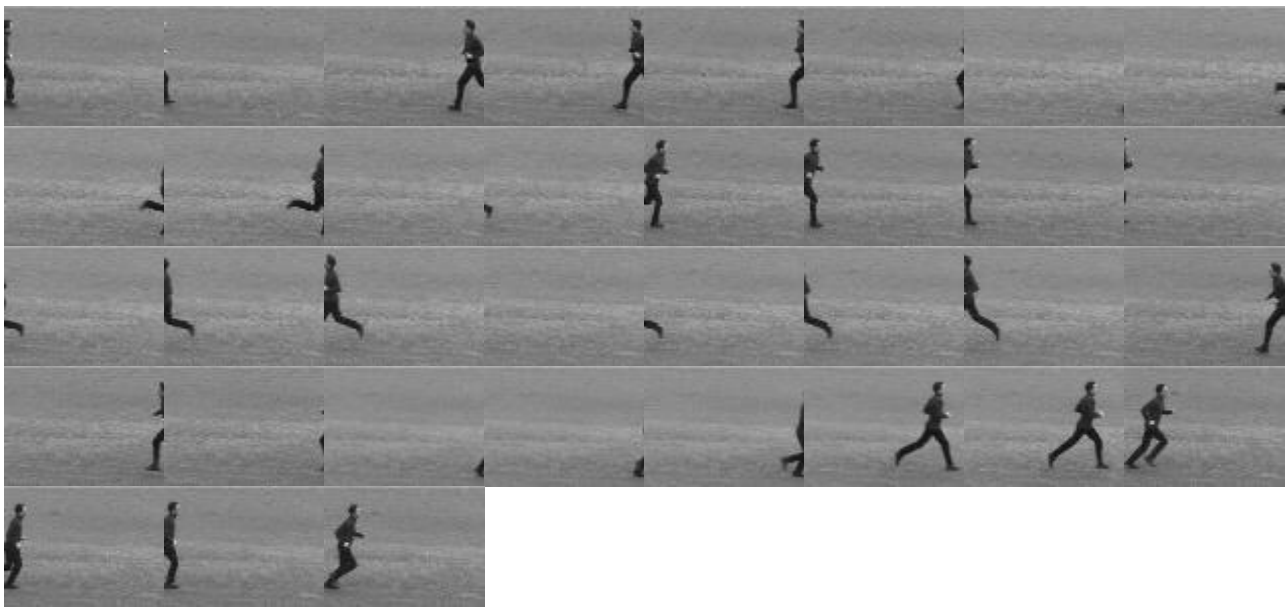


Fig 4. Keyframes (35) of data person01_running_d1_uncomp.avi (entropy difference)

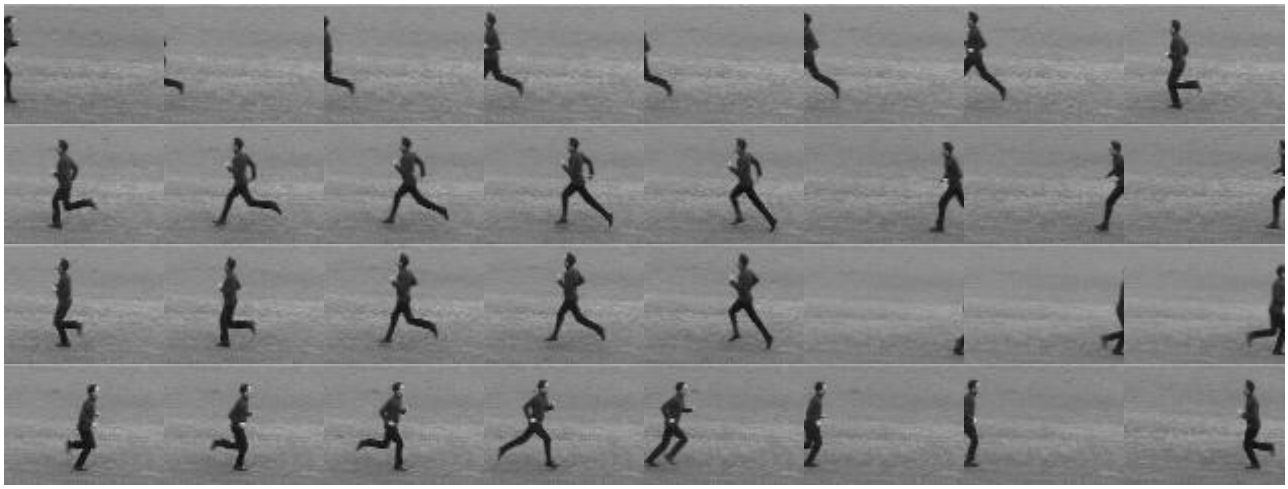


Fig 5. Key-frames (48frames) of data person01

References

[1] A. F. Smeaton, "Techniques used and open challenges to the analysis, indexing and retrieval of digital video," *Inf. Syst.*, vol. 32, no. 4, pp. 545–559, 2007, doi: 10.1016/j.is.2006.09.001.

[2] H. Zhao, W. J. Wang, T. Wang, Z. Bin Chang, and X. Y. Zeng, "Keyframe extraction based on HSV histogram and adaptive clustering," *Math. Probl. Eng.*, vol. 2019, 2019, doi: 10.1155/2019/5217961.

[3] S. Yang and X. Lin, "Key frame extraction using unsupervised clustering based on a statistical model," *Tsinghua Sci. Technol.*, vol. 10, no. 2, pp. 169–173, 2005, doi: 10.1016/S1007-0214(05)70050-X.

[4] S. W. B. Naveed Ejaz, Tayyab Bin Tariq, "Adaptive Key Frame Extraction for Video Summarization using an aggregation mechanism," *J. Vis. Commun. Image*, vol. 23, no. July, pp. 1031–1040, 2012.

[5] W. Wolf, "Key frame selection by motion analysis," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1996.

[6] A. P. Markos Mentzelopoulos, "Key Frame Extraction Algorithm using Entropy Difference," in *MIR'04*, pp. 39–45.

[7] Y. C. Liping Ren, Zhiyi Qu, Weiqin Niu, Chaoxin Niu, "Key Frame Extraction Based on Information Entropy and Edge Matching Rate," in *2nd international Conference on Future Computer and Communication*, 2010, pp. 91–94, doi: 978-1-4244-5824-0.

[8] S. huazhong Haung Min, "An algorithm of key frame extraction based on adaptive threshold detection of multi-features," in *International Conference on Test Measurement*, 2009, pp. 149–152.

[9] L. Huang, "An approach of Key Frame Extraction Based on Mutual Information," in *IEEE*, 2009, pp. 1–4.

[10] C. V. Sheena and N. K. Narayanan, "Keyframe Extraction by Analysis of Histograms of Video Frames Using Statistical Methods," *Procedia Comput. Sci.*, vol. 70, pp. 36–40, 2015, doi: 10.1016/j.procs.2015.10.021.

[11] Y. Shi, H. Yang, M. Gong, X. Liu, and Y. Xia, "A Fast and Robust Key Frame Extraction Method for Video Copyright Protection," *J. Electr. Comput. Eng.*, vol. 2017, no. 2, 2017, doi: 10.1155/2017/1231794.

[12] L. Chen and Y. Wang, "Automatic key frame extraction in continuous videos from construction monitoring by using color, texture, and gradient features," *Autom. Constr.*, vol. 81, no. May 2016, pp. 355–368, 2017, doi: 10.1016/j.autcon.2017.04.004.

[13] C. Gianluigi and S. Raimondo, "An innovative algorithm for key frame extraction in video summarization," *J. Real-Time Image Process.*, vol. 1, no. 1, pp. 69–88, 2006, doi: 10.1007/s11554-006-0001-1.

[14] A. Rashidi, F. Dai, I. Brilakis, and P. Vela, "Optimized selection of key frames for monocular videogrammetric surveying of civil infrastructure," *Adv. Eng. Informatics*, vol. 27, no. 2, pp. 270–282, 2013, doi: 10.1016/j.aei.2013.01.002.

[15] "KTH Databse." https://www.csc.kth.se/cvap/actions/.