Volume 9, Issue 2 (VI)
April – June 2022

ISSN: 2394 – 7780

International Journal of
Advance and Innovative Research

Indian Academicians and Researchers Association
www.iaraedu.com

International Conference

On

Emerging Trends in Digital Technologies-2022 (ICETDT- 2022)

22nd February 2022

Organized by

SVKM's
Usha Pravin Gandhi College of Arts, Science and Commerce

Publication Partner

Indian Academicians and Researchers Association

Advisory Committee

Dr. Anju Kapoor
Principal, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Vile Parle, Maharashtra, India

Dr. Pallavi Jamsandekar
Director, Bharati Vidhyapeeth (Deemed to be University), Sangli.

Dr. Sukhada
Assistant Professor, IIT BHU, Varanasi

Dr. Snehalata Shirude
Assistant Professor, NMU, Jalgaon

Dr. Sangeeta Chakrabarty
Associate Professor, Goa University

Dr. Neel Mani
Associate Professor, Amity Institute of Information Technology, Noida, UP

Ms. Sangeeta Nimkar
Technical Specialist, United Postal Services, New Jersey, USA

Convener

Ms. Smruti Nanavaty
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Co- Convener

Dr. Neelam Naik
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Dr. Manisha Divate
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Organizing Committee

Dr. Swapnali Lotlikar
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Mr. Prashant Chaudhary
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Ms. Sunita Gupta
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Mr. Rajesh Maurya
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

Ms. Neha Vora
Assistant Professor, Usha Pravin Gandhi College of Arts, Science and Commerce

PRESIDENT'S MESSAGE

Higher Education is an imperative milestone for learners in current times. Indian higher education system is maturing towards its presence in Global Higher Education space. This calls for reformative policy initiatives from stakeholders in curricula, pedagogy, use of technology, partnerships, governance and funding. Encompassing this vision, Usha Pravin Gandhi College has a learner centered paradigm of education where the student is placed in a competitive learning environment of the 21st century to foster excellence, equity and quality.

The academic staff at the college constantly commits themselves towards the growth of students to create the desired intellectual, economic and social value.

With firm faith in the saying, "*Vidhyadhanam Sarvadhanam pradhanam"* – knowledge is the only real wealth in this world, I welcome the students of SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce. I am confident that efforts to excel in the field of higher education through innovative practices, immersed and engaged learning and the inculcation of moral and social values in the learners will continue. May you make the best of these openings to shape your careers and future?

Wishing everyone at Usha Pravin Gandhi College of Arts, Science and Commerce all success this new Academic year 2021-22.

Amrish R. Patel
President,
Shree Vile Parle Kelvani Mandal,
Vile Parle, Mumbai

PRINCIPAL'S MESSAGE

I am delighted that SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce is organizing a one-day International Conference on "Emerging Trends in Digital Technologies-2022 (ICETDT-2022) on 22nd February 2022 and on this occasion, we are bringing out a souvenir.

The theme of the conference is very relevant in the present scenario. Society and the professional world continue to evolve and change with the growth of technology. This in turn has had a tremendous impact on the educational sphere, leading to a growing trend in newer forms of the engagement. Educators have to adapt to the decreasing attention spans of the students and use technology in teaching and learning.

To keep the attention of millennials, the content presented to them must have excellent visuals and dialogue that will hold their attention. Teachers have to be innovative presenters having research skills that will inspire their students to take ownership of their learning.

The post pandemic world is opening up to multiple opportunities and to start the year with a research conference is very encouraging. I wish the conference teams with resounding success.

Dr. Anju Kapoor
Principal,
Usha Pravin Gandhi College
of Art, Science and
Commerce, Vile Parle,
Mumbai

International Journal of Advance and Innovative Research

Volume 9, Issue 2 (VI): April - June 2022

SJIFactor

Scientific Journal Impact Factor

CERTIFICATE OF INDEXING (SJIF 2018)

This certificate is awarded to

International Journal of Advance & Innovative Research
(ISSN: 2394-7780)

The Journal has been positively evaluated in the SJIF Journals Master List evaluation process
SJIF 2018 = 7.363

SJIF (A division of InnoSpace)

SJIFactor Project Manager
International Advisory Services
INNOSPACE INTERNATIONAL

SJIFactor Project

International Journal of Advance and Innovative Research

Volume 9, Issue 2 (VI): April - June 2022

CONTENTS

Research Papers

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

MULTI-CLASS SAFETY HELMET DETECTION USING FASTER- RCNNAND SINGLE SHOT DETECTOR ALGORITHMS

[1]Sai Sharan, [2]Preetham Reddy, [3]Fayaz Moqueem and [4]Afraz Moqueem
[1,2]School of Electronics and Communication Engineering, Vellore Institute of Technology, Vellore, India
[3,4]School of Electronics and Communication Engineering, Muffakham Jah College of Engineering and Technology, Hyderabad

ABSTRACT
In recent days, Industrialisation and globalization drastically increasing the construction of massive buildings globally. With the increase in the number of structures, there is also an increased number of workers at risk. Since construction requires workers to do tasks at dangerous places operating heavy and hard materials, it is necessary to wear a safety helmet to avoid serious injuries and deaths. Studies indicate that helmets safety could reduce brain injuries as high as 95% for concrete block impacts. So, this paper put forth two algorithms for detecting safety helmets in real-time. The algorithms used are Single-Shot MultiBox Detector and Faster Regions with Convolutional Neural Networks(Faster-RCNN). This paper also performs a comparative analysis of F-RCNN and SSD in safety helmet detection. Our work has a plethora of applications such as real-time patient monitoring, etc.

Keywords: Faster R-CNN, SSD, Deep learning, Data Augmentation.

I. INTRODUCTION
A large number of hospitalizations of construction workers is related to not wearing a safety helmet at the construction site. Dangerous work conditions and heavy construction materials put works life at high risk. According to the studies conducted by the US Bureau of Labor Statistics, there is an approximately 2% increase in the fatality rates of construction workers every year. Studies in China indicate that about 840 construction workers without helmets died by falling from elevated places. All these statistics emphasize the importance of wearing a safety helmet. This issue can be addressed by deploying surveillance cameras atconstruction sites and constantly monitoring workers.

In recent decades deep learning is developed drastically inobject detection and classification. Even though there are many other pre-existing helmet detections, most of them lack real-time processing and perform poorly. Hence, we proposed two methods for detecting safety helmets in real- time through CCTV cameras. We have deliberately scrutinized SSD and Faster RCNN models and done comparative analysis for the same.

Computer Vision plays an important part in safety helmet detection. Our paper includes helmet recognition, feature extraction, and segmentation. Extraction of frames could ease the recognition in which an image is classified with a label. Segmentation creates a pixel-level understanding and helps to track helmets in the frame. A Faster Regional Convolutional Neural Network is extensively used for helmet detection in real-time. It involves a pre-trained CNN feature extraction network. Then the dataset is augmented and trained on two subnetworks of Faster RCNN. One of them, the Region Proposal Network(RPN), generates object proposals, whereas the other network helps to predict actual object class. Finally, an ROI pooling system followed by a bounding box is used and displays whether the helmet is detected or not.

Single Shot Detector (SSD) uses a multi-box approach to detect objects in a single image. It's an accurate and efficient object classification model. However, it is more complicated than R-CNN architecture, and bounding box proposals are eliminated to get faster output. It is generally based on reducing convolutional filters to predict object classes for low-quality pictures.

This paper follows the following sections. Section I provides an introduction followed by related works in Section II. Section III carried through the overview, and proposed models and methodology are mentioned in section IV Section V includes the result and discussion. The paper comes to an end with a conclusion and references.

II. RELATED WORK
Work at construction sites usually involves high-risk tasks requiring workers to go to dangerous places with heavy materials. However, the construction authorities and government emphasize ensuring the workers' safety by implementing rules that require every worker to wear a safety helmet. This way, there could be a decline in the number of instances of serious injuries and deaths. But, due to little or no awareness among the workers and carelessness of the officials at construction sites, helmets are not efficiently used. So, to counter this dangerous act, there must be a helmet monitoring system to ensure safetyat the construction site. Many previous works

ISSN 2394 - 7780

have implemented helmet detection systems based on various machine learning and deep learning models, which are mentioned in the following literature review.

According to [1], although existing object classification works were based on detecting those without helmets, they relied entirely on image processing techniques, which were tedious and complex. The paper compares various recently developed methods for detecting helmets. Various important concepts such as Deep learning, Machine learning, and Image processing. A clear block diagram of a process of object detection using image processing is presented. Classifiers such as VGG16, VGG19, Inception V3, and MobileNets are compared for the evaluation metric Accuracy. It is observed from the tabular results that MobileNets outperformed other classifiers. The authors also put forth the implementation of number plate detection with the help of Natural language processing. Various stages of NLF for detecting number plates are included in the paper. Finally, it is concluded that the new machine learning methodologies would ease the process of helmet detection.

In [2], Arya et al. developed a video monitoring system based on computer vision for helmet detection of construction workers. The paper proposes a clear overview of machine learning-based helmet detection. The authors considered video input as a dataset and converted it into image frames so that these images can be classified. The extracted frames are then preprocessed to be suitable for further detection. Finally, features are extracted from the input image. In the paper, the authors compared various algorithms such as Faster RCNN, CNN, and YOLO v2 to detect helmets. In the end, It is concluded that almost all the deep learning algorithms considered provide more than 90% accuracy, and each algorithm has its pros and cons in detecting helmets.

According to Runyon et al. [3], although helmet detection effectively reduces accidents at construction sites, it is complex and challenging to implement efficiently. So the authors propose a YOLOV4 deep learning model to detect helmets of construction workers accurately in real-time. To be similar to an actual worksite, the dataset is adjusted for photometric distortions by adjusting brightness, image noise, and other cosmetic parameters. The paper performed the Mosaic data enhancement method in which four images are randomly selected and put together as a single image. To solve the high cost of reasoning calculation, a CSP module is added to YOLOV4 to make CSPDarknet53. The final deep learning model after training with the dataset produced an accuracy of 95.1%.

According to Zijian Want et al. in [4], the existing deep learning models for detecting personal protective equipment lack performance and detection range. The paper put forth an approach for training and evaluating eight detectors based on YOLO architecture. This paper aimed to detect six classes which include four different colored helmets, vests, and person. The dataset consisted of 1300 high-quality CHV images created similar to the real work environment as background. Different versions of YOLO were analyzed based on performance metrics such as accuracy and speed. The comparative results indicated that YOLO v4 could not improve its performance by adding more datasets. YOLO v5x has the best mAP value, and additionally, it is found that YOLO v5s has a faster processing speed for a single image. The paper also mentions the limitations of the model, which include YOLO models inaccurately predict small substances from long distances.

In [5], Haikuan Wang et al. proposes a new detection model based on improved YOLOv3 to aim for higher detection capability on the construction site. This model is named CSTOLOv3. In the beginning, the farknet53, which is considered the backbone network, is improved by applying the cross-stage partial network, which further enhances the training speed and minimizes the calculation cost. Then the spatial pyramid pooling structure is then incorporated in the YOLOv3 model and followed by the fusion top-down and bottom-up feature strategies to enhance multiscale prediction. The dataset is derived from the real construction areas surveillance cameras and consisted of around 10,000 images. Manual annotation is required for training the model with the provided dataset. It is concluded that the novel method can successfully outperform YOLOv3 in terms of speed. Moreover, it can reduce mAP by 28% compared to YOLOv3.

In [6], Guoqing Jin et al. proposes implementing Densenet with YOLOv3 to reduce the technical cost and feature extraction. This collectively is termed as YOLO- Densebackbone Convolution neural network. The results imply that the new model has an improved accuracy of 2.44% greater than the conventional YOLOv3 model.

In [7], Yang Bo et al. devised a method to ensure that workers wear their helmets correctly. The proposed model uses YOLOV3 to detect helmets and heads to categorize and classify the input images. For training purposes, the model used the publicly available COCO dataset. The model was sorely tested using pictures from a real-time situation on a construction site. The model intended to overcome the manual method of safety checks. The detection algorithm used YoloV3 with Darknet53 as a backend engine. The researcher claims that

ISSN 2394 - 7780

their model gave out accuracy of 90 percent. The paper also included comparisons of YoloV3 with other image detection algorithms such as SSD and DSSD.

In [8], Amal Santosh et al. developed a methodology for detecting if motorcyclists were wearing helmets while riding. The model used SSD to determine the bounding boxes and a single-layered CNN architecture to segregate the input into two groups: those who wore helmets and those who did not. SSD was employed to complete both the image segment and classification tasks. The developed model also employed the YOLO technique to recognize license plates of riders who were not wearing helmets. They used images from a video clip collected from a security camera for training purposes. The model's long-term goal is to automate the challan generation process.

In [9], Shoukun Xu et al. suggested a sophisticated deep learning-based approach to address difficulties with safety measures by tracking employees who wear helmets against those who do not. To grasp insights containing tiny features, the model employs an improvised Faster RCNN model. The model also makes use of Online Hard Example mining to improve and stabilize the positive and negative outcomes of the input samples. VOC2007 and VOC2012 datasets were used to train the model. To achieve reliable findings, the IOU was adjusted to 0.95 for categorization reasons. Compared to the generic Faster RCNN, this model promises to provide results that are 7percent more accurate results.

In [10], Zheng Fan et al. came up with a novel approach for helmet detection. The method is integrated with two basic algorithms. These base algorithms take their position based on the size of the required features from an input image. The base algorithms are Faster Rcnn and Cnn. The main motive of these fundamental algorithms is to extract needed insights from a particular image. Integration of the base algorithms is done using ensemble learning. The model is trained using a publicly available data set consisting of over 7000 images and used five-fold cross- validation to improvise the accuracy. The authors claim that this novel approach outperforms Faster Rcnn with a mean average precision of 0.93.

Rui Geng et al.[11] with a novel approach improvised YOLOV3 algorithm for better helmets detection without compromising in its speed. The main motive is to increase the accuracy of detection in an unbalanced dataset. The model incorporates Gaussian Fuzzy algorithm for data augmentation for better target detection. The Gaussian blurring process was employed for better image enhancement and to apply image transformers to the loader. The model uses cross-entropy as a loss function. The researchers claim that this model improved the confidence level of YOLO-V3 target detection by 0.01-0.02 percent of the unbalanced dataset that contained over 600 images.

Jie Li et al. In [12] proposed a model for helmet detection using state-of-art machine learning techniques. The model employed the ViBe background modeling algorithm to segregate the objects in motion and bring them into the foreground. To characterize interior human motion, the Histogram of Oriented Gradient (HOG) feature is retrieved after collecting the motion region of interest. The histogram of the oriented gradient was calculated using HOG, and the original picture was normalized using gamma correlation. The model further used SVM for the classification. Finally, the model used a color feature to detect instances where a safety helmet was worn. The model gives out accuracy of around 80 percent at seven frames per second.

A. LIMITATIONS

Although there are existing works are based on helmet detection, the are not feasable to implement in complex andreal time detection conditions. Some limitations are monomial classification which just involves, weather a person is wearing helmet or not. So, In this paper, multiclass classification of helmets is done to detect various types of helmets and also moniter in realtime conditions.

We have also scrutinized Faster-RCNN and SSD algorithmsto compare each other in detecting helmets.

III. OVERVIEW

A. MOTIVATION

Although there are research works already done on helmetdetection, most of them turn out to be impractical and inefficient when it comes to detecting the helmet in real- time rather than in static conditions. Moreover, existing models consume a huge amount of time for training and involve complex processes. Adding to that, most of the works are based on binary classification, which involves “Helmet detected” or “Not” output, whereas this paper proposes a multiclass detection model in which helmets with different colors can be detected. This multiclass detection is very much helpful in analyzing the type of workers present in the construction site. For instance, if the system does not see a white helmet for 24 hours, it could alert a higher supervisor that the engineer is absent for that day.

ISSN 2394 - 7780

B. DATASET

Since there are no suitable datasets for our paper, they contain one class of helmets, even if available. So for this paper, we have built our Dataset by collecting 961 images from google. We have divided our Dataset into three classes, namely engineers(385), laborers (317), Electricians(259). We have divided the Dataset in the ratio of 60 to 40 percent as a training and testing set. The Dataset built would be released to kaggle soon after works get completed.

C. Pre-processing

The Dataset contains images of different sizes and different lightings. Few of them are in black and white. So we have done preprocessing on the Dataset to make it suitable for training the models. All the images are resized to 125 x 125 pixels to ensure all of the same size and quality. Since the images are not clear, contrast enhancement is done to make them distinguishable using colors on output. After that, the Dataset is augmented to create more images so that accuracy of the model increases.

IV. PROPOSED MODELS

A. Single Shot Multibox Detector

The Single Shot Multibox Detector (SSD) is a deep learning object detection method that adds many layers to the preexisting CCN structure to offer a thorough categorization of the input images from the dataset. SSD is a variant of the VGG16 architecture, which was previously available.SSD is a VGG16 architectural modification that adds layers to identify numerous objects from a single picture input. When compared to VGG16 architecture, SSD has a higher mean average precision. Because of its superiority in object identification from high-quality pictures and flexibility to facilitate transfer learning, the SSD method employs the VGG16 architecture as the basic network.

The input picture is first fed to the VGG16 architecture, which extracts feature maps. The output of the VGG16 is then input into SDD, a six-layer convolutional neural network that makes predictions for each class. With a progressive reduction in the input size to each matching layer, the layers are employed for feature extraction at various sizes. SSD creates several predictions for each item, resulting in many bounding boxes. To eliminate duplicate predictions, we use a non-maximum suppression classifier. SDD examines each object's confidence score and considers the best forecasts to get an accurate result. SSD matches the default bounding boxes generated after the six-layer classification to the ground truth boxes during training. For this purpose, SSD employs IOU (Intersection Over Union), which is computed as given in the equation below.

The initial task of the Faster R-CNN is to employ the Region proposal network(RPN). Region proposal network's main motive is to categorize between the foreground and the background class. To detect the required object orextract the features from an input image of the dataset, RPN

$$\text{Intersection Over Union (IOU)} = \frac{ra\ art1 \cap ta\ ara}{ra\ art1 \cup ta\ ara} \quad (1)$$

In equation 1, the area of part1 represents the area covered by the helmet, and the total area is the complete object area. Only bounding boxes that have a 50% overlap with the ground truth boxes are deemed exact. SSD is quicker than RCNN since it does not require a specialized bounding box proposal. As we continue to convolve a network of this depth, we are going to lose some finer details and characteristics of the input picture, but SSD is piled with several layers to prevent this. The SSD architecture is shown below.

ALGORITHM

1. Create a data collection from the 60 frames per secondcaptured video.
2. Train the algorithm for feature identifications such ashead using the data set.
3. Use VGG16 to extract features from the data set.
4. Finding bounding boxes with feature maps and extraSDD layers.
5. Bounding boxes, which are also utilized to locate IOU.
6. Categorizing bounding boxes with an IOU of more than0.5.
7. A non-maximum suppression classifier is used to removeduplicate predictions.

ISSN 2394 - 7780

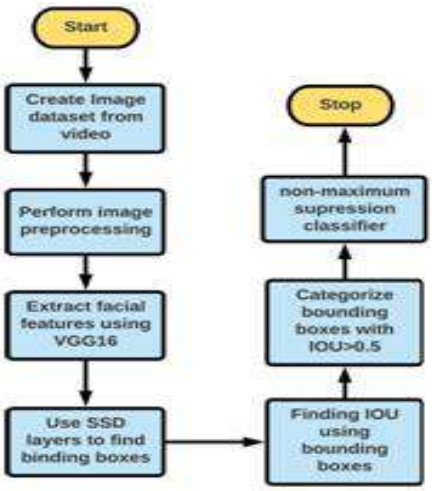


Figure 2. Flowchart of SSD

B. Faster- Regional Convolution Neural Network:

Faster R-CNN is a pre-trained deep learning-based object detection algorithm. Faster R-CNN is a combination of FastRcnn and RPN and uses selective search algorithms for object detection. Faster Rcnn delivers the results in minimaltime and with a greater mean average precision.

Starts generating anchor boxes with predefined dimensions(1:1 and 2:1 and 1:2 length: width ratios) on the input images. Once anchor boxes are developed, Intersection over union (IOU) will be calculated for each box. IOU is the overlap between the anchor boxes and the ground truth boxes. In our setup, the ground truth boxes are the helmets. Only the anchor boxes with IOU>=0.5 will be labeled as the foreground class, and the rest comes under the background class. Finally, the anchor boxes which are classified under the foreground class will be processed for further analysis. In the next stage, the output of the RPN, which has different-sized feature maps, is uninformed usingRegion of Interest (ROI) pooling. We have several options for constructing the final classifier and regressor with the fixed ROI Pooling outputs as inputs. After classifying finally, a Non-max suppression layer is utilized for removing the duplicates, i.e., by considering only the top 100 recommendations.

ALGORITHM

1. Fetch the original dataset.
2. Perform preprocessing on the retrieved dataset.
3. Employe the Region proposal network (RPN).
4. Generate anchor boxes with dimensions(1:1 and 2:1 and1:2 length:width ratios).
5. calculate Intersection over union (IOU) using equation1.
6. Then categorize foreground and background classes.
7. Use the ROI pooling layer for uniformity of the featuremaps.
8. The Fast R-CNN is then used to classify the extractedfeature maps.
9. After classification, a non-maximum suppression layer isemployed to remove duplicates.

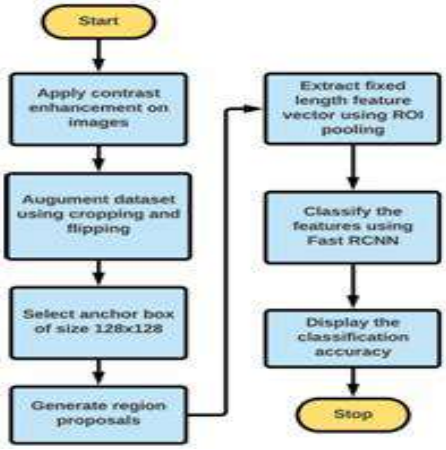


Figure 3. Flowchart of Faster R-CNN

ISSN 2394 - 7780

V. RESULTS AND DISCUSSION

A. Experimental results and discussion.

Evaluation Parameters:

The Faster Rcnn and SSD algorithms are compared on parameters which includes mean Average precision (mAp) ,Recall and Intersection over union (IOU).

The simplest basic performance metric is accuracy, which is just the ratio of accurately predicted observations tothe total number of observations considered.

$$\text{Accuracy} = \frac{\text{True positive}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \quad (2)$$

The ratio of precisely predicted positive findings to the totalexpected positive findings is known as precision.

are tabulated as shown in Table-1. The table-1 depicts that Faster Rcnn outperforms SSD in every parameter. The SSD and Faster RCNN have an average mAp of 67.34 percent and 78.92 percent, respectively. In general, SSD is 27 percent less accurate than HDD. Faster RCNN surpasses SSD by 12 percent in terms of mAp. However, considering computation time SSD with regression architecture and a single neural network performs tasks faster than Faster RCNN. Overall, the Faster Rcnn with multiple data preprocessing input outperforms SSD for helmet identification on our dataset.

CONCLUSION

Construction workers nowadayas are at increased prone toriskes and loss of lives due to heavy materials. Most of the accidents takes place because of carelessness of the workers to not wear a safety helmet at work place. Our proposed models helps to moniter the safety helmets on the workers by using algorithms such as Faster-RCNN and Single Shot Detectors. After peroforming a deliberate comparation of both the algorithms, it is concluded that Faster RCNN outperformed SSD algorithm in varipus performancemetrics such as mAp, IOU, Recall and Precision.

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \quad (3)$$

The ratio of precisely predicted positive findings to thepredicted results is known as recall.

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False Negative}} \quad (4)$$

True Positive, False Positive, False Negative, and True Negative in the above equations represent the total number of predictions that correctly identified the helmets, the total number of predictions that correctly identified other objects as helmets, and the total number of predictions that correctly identified helmets as other objects.

Our methodology works well for the Construction worker dataset. With larger data, our model aims to provide many accurate results. Our aim for future research is to assess the success of our methodology on more complex datasets. We plan to do research with infrared and depth data for real- world scenarios. Since it is not feasible to track drivers with a regular grey scale camera during a dim light, by processing temporal details in images, we can increase the efficiency of our model.

$$\text{Intersection Over Union (IOU)} = \frac{ra\ vra}{ra\ nn} \quad (5)$$

"A review on various methodologies used for vehicle classification, helmet detection and number plate recognition," Evolutionary Intelligence, vol. 14, no. 2, pp. 979–987, Sep. 2020, doi: 10.1007/s12065-020-00493-7.

The only bounding boxes which has over 50 percentage overlap with the ground truth boxes ((I.e IOU>0.5)) are considered to deliver precise results.

Table of Comparision:

TABLE -I

Parameter	Faster - RCNN	SSD
mAp	78.92%	67.34%

ISSN 2394 - 7780

IOU	72.14%	67.20%
Recall	86.40%	81.33%
Precision	94.24%	88.24%

B. INTERPOLATION

Faster RCNN and SSD algorithms are compared based on mAp, recall, IOU, and precision. The experimental results

REFERENCES

[1] S. Sanjana, S. Sanjana, V. R. Shriya, G. Vaishnavi, and K. Ashwini, A. K M and A. K K, "A Review on Deep Learning Based Helmet Detection," SSRN Electronic Journal, 2020, doi: 10.2139/ssrn.3791099.

[2] R. Cao, H. Li, B. Yang, A. Feng, J. Yang, and J. Mu, "Helmet wear detection based on neural network algorithm," Journal of Physics: Conference Series, vol. 1650, p. 032190, Oct. 2020, doi: 10.1088/1742-6596/1650/3/032190.

[3] Z. Wang, Y. Wu, L. Yang, A. Thirunavukarasu, C. Evison, and Y. Zhao, "Fast Personal Protective Equipment Detection for Real Construction Sites Using Deep Learning Approaches," Sensors, vol. 21, no. 10, p. 3478, May 2021, doi: 10.3390/s21103478.

[4] H. Wang, Z. Hu, Y. Guo, Z. Yang, F. Zhou, and P. Xu, "A Real-Time Safety Helmet Wearing Detection Approach Based on CSYOLOv3," Applied Sciences, vol. 10, no. 19, p. 6732, Sep. 2020, doi: 10.3390/app10196732.

[5] F. Wu, G. Jin, M. Gao, Z. HE, and Y. Yang, "Helmet Detection Based On Improved YOLO V3 Deep Model," 2019 IEEE 16th International Conference on Networking, Sensing and Control (ICNSC), May 2019, doi: 10.1109/icnsc.2019.8743246.

[6] Y. Bo et al., "Helmet Detection under the Power Construction Scene Based on Image Analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Oct. 2019, doi: 10.1109/iccsnt47585.2019.8962495.

[7] Santhosh, "Real-Time Helmet Detection of Motorcyclists without Helmet using Convolutional Neural Network," International Journal for Research in Applied Science and Engineering Technology, vol. 8, no. 7, pp. 1112–1116, Jul. 2020, doi: 10.22214/ijraset.2020.30426.

[8] Y. Gu, S. Xu, Y. Wang, and L. Shi, "An Advanced Deep Learning Approach for Safety Helmet Wearing Detection," 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), Jul. 2019, doi: 10.1109/ithings/greencom/cpscom/smartdata.2019.00128.

[9] Z. Fan, C. Peng, L. Dai, F. Cao, J. Qi, and W. Hua, "A deep learning- based ensemble method for helmet-wearing detection," PeerJ Computer Science, vol. 6, p. e311, Dec. 2020, doi: 10.7717/peerj-cs.311.

[10] X. Wang, D. Niu, P. Luo, C. Zhu, L. Ding, and K. Huang, "A Safety Helmet and Protective Clothing Detection Method based on Improved- Yolo V 3," 2020 Chinese Automation Congress (CAC), Nov. 2020, doi: 10.1109/cac51589.2020.9327187.

[11] J. Li et al., "Safety helmet wearing detection based on image processing and machine learning," 2017 Ninth International Conference on Advanced Computational Intelligence (ICACI), Feb. 2017, doi: 10.1109/icaci.2017.7974509.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

ANALYZING TRENDS IN CRYPTOCURRENCY USING MACHINE LEARNING TECHNIQUES

Bindy Wilson and Ansa Jovel Kunnathettu
Department of Information Technology and Computer Science, Keraleeya Samajam Dombivli's Model College Dombivli, India

ABSTRACT
Cryptocurrency has so changed the financial world that it is difficult to ignore the relevance it is gaining. As technology keeps evolving, a stronger cryptocurrency financial system has found its place, raising its popularity. Manual analysis of large volumes of text data is not accurate. A system for measuring the spread and current trend of cryptocurrency by using text mining and data analysis is presented in this paper. News feed data from different domains were collected using NewsAPI. Latent Dirichlet Allocation (LDA) has been used to perform topic modelling, to extract trending topics related to cryptocurrency. Topic clustering is used to identify the topics that are related to each other. The general sentiment over the topic as reflected in the news items was evaluated in Sentiment Analysis, using spaCy, a Natural Language Processing (NLP) library in Python. Performance is evaluated using precision, recall, F1 score and support.

Keywords: Cryptocurrency, Text mining, Topic modelling, Sentiment Analysis, Latent Dirichlet Allocation, Natural Language Processing

I. INTRODUCTION
Cryptocurrency is a digital currency system where the payment token is a string of bits[1].It is a form of digital asset based on a decentralized, distributed network. Blockchains, which can be termed as a Distributed Ledger Technology, are an essential component of many cryptocurrencies. It is devoid of a central controlling authority; instead relies on the above distributed network for verification of transactions, as well as for updating and storage of the record of transaction histories. Cryptocurrencies are the major mode of transaction in different applications and networks such as online social networks, online games, and peer to peer networks[2]. The major formats include Bitcoin, Ethereum and Dogecoin. Lately, Non-Fungible Tokens or NFTs also use the same blockchain technology and represent a type of digital asset that tokenizes unique items like art, music and a wide range of other items. Since these terms have been appearing in various news articles recently, a sentiment analysis can help to reveal the sentiment given out by these articles.

Topic modelling comes under unsupervised machine learning which is used for natural language processing. It can be used to aid text mining where a set of documents can be represented with the help of a list of topics that explain the underlying information in them. Each such topic can have a collection of words which are grouped in such a manner that each group of words represents a topic in the set of documents. There are several topic modelling algorithms, of which the one discussed in this paper is a probabilistic model called Latent Dirichlet Allocation (LDA). The topics hence extracted have been subjected to sentiment analysis.

Clustering is an unsupervised approach to identify groups with similar characteristics in a collection of observations. Among different clustering approaches, hierarchical clustering technique has been used in the paper to group related topics.

II. LITERATURE REVIEW
Topic modelling is a very popular tool for exploratory analysis. LDA is a topic modelling method which is simple to implement and to understand. Toqir A. Rana et al [3] explored different topic modelling techniques for sentiment analysis to compare the accuracy of different systems. They have found that LDA-based models are the most widely used and can extract those topics which appeared in the document frequently and may lack the ability to handle implicit aspects. Hye-Jin Kwon et al [4] used topic modelling and sentiment analysis on the posts of Skytrax (airlinequality.com) to figure out important words in online reviews. As a result, 'seat', 'service', and 'meal' were found to be significant issues in the flight. The result also revealed that delay was the main issue, which mainly affects customer dissatisfaction while sentiment analysis revealed 'staff service' can make customers satisfied.

Maibam Debina Devi and Navanath Saharia [5] made use of Latent Dirichlet Allocation (LDA) model to extract the topics from 150 lyrics samples of Manipuri songs written using Roman script. They used the unsupervised machine learning models to obtain the different sentiment class underlying the lyrics, in the form of topics. Rania Albalawi et al [6] have investigated topic modelling methods as applied to short textual social data, to detect important topics. They used two textual datasets: the 20-newsgroup data and short conversation data from

ISSN 2394 - 7780

the Facebook social network site and found that LDA methods delivered more meaningful extracted topics and generated the most valuable outputs.

Khusbu Thakur et al [7] investigated the applications of text mining among researchers and found that the LDA and R package is the most extensively used tool and technique. Xing Fang et al [8] did experiments on sentiment analysis or opinion mining for sentence-level categorization and review-level categorization. They collected the online product reviews from Amazon.com and used this data for the analysis.

Sangeeta Rani et al [9] used topic modelling and sentiment analysis to identify trending topics related to Clean India Mission. The data they used were extracted from Twitter and the results were used to analyze the interest of the Indian population towards the mission. Bisma Shah et al [10] used social media content to perform a comparative study on different methods of sentiment analysis, extracted polarity of the dataset and categorized it as positive, negative or neutral.

Table 2.1 Summary of Literature Survey

Author	Data selection for analysis	Topic modelling algorithm used
Toqir A. Rana et al [3]	Customer Reviews	Latent Dirichlet Allocation
Hye-Jin Kwon et al [4]	Posts of Skytrax (airlinequality.com)	Latent Dirichlet Allocation
Maibam Debina Devi and Navanath Saharia [5]	150 lyrics samples of Manipuri songs written using Roman script	Latent Dirichlet Allocation, Heuristic Dirichlet Process
Rania Albalawi et al [6]	Textual dataset of conversation from Facebook and newsgroup data.	Latent Semantic Analysis, Latent Dirichlet Allocation, Non-negative matrix factorization, Random projection, Principal Component Analysis.
Xing Fang et al [8]	Online product reviews from Amazon.com	Sentiment Analysis
Sangeeta Rani et al [9]	Twitter data on Clean India Mission	Latent Dirichlet Allocation
Bisma Shah et al [10]	Extracting polarity of social media dataset	Naive Bayes Classifier, Maximum Entropy Classifier, Interdependent Latent Dirichlet Allocation

Table 2.1 gives a summary of the literature review done by several authors in Topic Modelling. Majority of the authors have used the algorithm of Latent Dirichlet Allocation (LDA) for Topic Modelling. Hence in our work LDA is the selected algorithm for Topic Modelling.

III. TOPIC MODELLING AND LDA METHODOLOGY

Topic Modelling is an unsupervised Machine Learning technique in which the model identifies the topics by detecting patterns like word clusters. LDA is an unsupervised, probabilistic topic modelling method which extracts topics from a collection of papers [11]. It is used to extract the most likely terms of topics and most likely topics related to the documents [12]. LDA assumes that the document consists of words which help to determine topics and it maps documents to a group of topics by assigning each word in the document to different topics. Fig. 1 represents the graphical model of LDA [13].

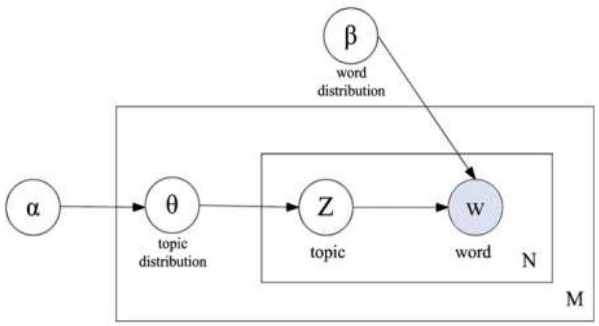


Fig. 1. Graphical model representation of LDA [13]

ISSN 2394 - 7780

In a corpus X containing Y words which are unique and M documents, where each document contains a sequence of strings s{s1,s2,s3……sn}. In a topic number P, the generative process followed by LDA for a document s is as follows:

- Sample P-vector θs from Dirichlet distribution p(θ|α), θs is the topic proportion mixture of documents.
- For i=1..n, sample word wi in the s form of document multinomial distribution p(wn| θs, β) where α is a P-vector parameter, and p(θ| α) is given in (1), β is a P*V matrix of word probabilities and βij= p(wj=1|zi=1), i=0,1,…P; j=0,1,2..V.

(1)

$$P(\theta|x) = \frac{\Gamma\left(\Sigma_{i=1}^{k}\alpha_i\right)}{\Pi_{i=1}^{k}\Gamma-(\alpha_i)}\theta_1^{\alpha_1-1}\ldots\theta_k^{\alpha_k-1}$$

LDA assumes independence between topics that are randomly drawn from a Dirichlet distribution. The process flow diagram of the LDA system applied in the current study is given in Fig. 2. It uses an unsupervised learning approach.

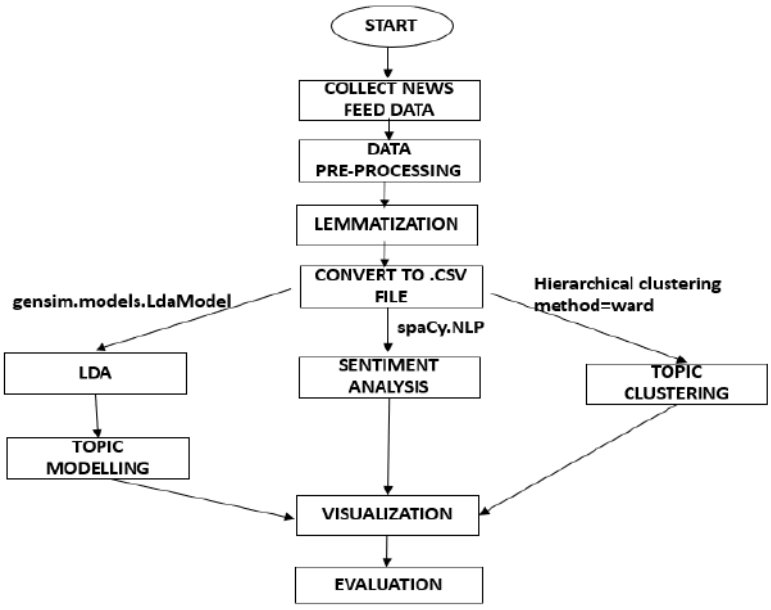


Fig. 2. Process flow diagram for Topic Modelling.

- Data collection: The system begins with collection of the data for analysis. The newsfeed data is collected using News API.
- Data pre-processing: Raw data is subject to errors, duplicates, numbers, punctuation, special characters and stop words. As a result, data cleaning is needed to make the data reliable for accurate analysis. The resultant data undergoes lemmatization and tokenization.
- Lemmatization: It refers to using vocabulary by removing inflectional endings and returning the base form of words known as lemma.
- LDA and Topic Modelling: Topic Modelling using LDA is applied on the matrix to extract trending topics and terms on newsfeed data related to cryptocurrency.
- Sentiment Analysis: It is needed for opinion mining as it identifies the emotional tone in a text. An approach used in Natural Language Processing.
- Topic Clustering: To identify how the topics are correlated to each other clustering is used.
- Visualization: To effectively communicate the results obtained from analysis. The extracted topics and their probabilities are plotted along with the clustering output.
- Evaluation: Evaluation is performed using precision, recall and f-score. Precision gives the percentage of truly positive values out of all the positive predicted values. Recall gives the percentage of predicted positive values out of all the total positives. F-score gives the harmonic mean of precision and recall.

ISSN 2394 - 7780

IV. IMPLEMENTATION

The data required for analysis is collected from news feed using News API from the domains coindesk.com, economictimes.com, indiatimes.com, techcrunch.com based on keywords cryptocurrency, bitcoin, dogecoin and ethereum. A total of 160 articles were retrieved for analysis from a period of three months. As raw data is subject to errors it undergoes preprocessing and cleaning for making the data accurate for analysis. Lemmatization and tokenization are performed on the resultant articles. The newsfeed data is then converted to a CSV file for further analysis.

LDA is applied to the .csv file data using gensim.models.LdaModel. The LDA model constructed on the corpus obtains the result as shown in Fig. 3. It shows the most relevant terms for each trending topic with respect to their probabilities.

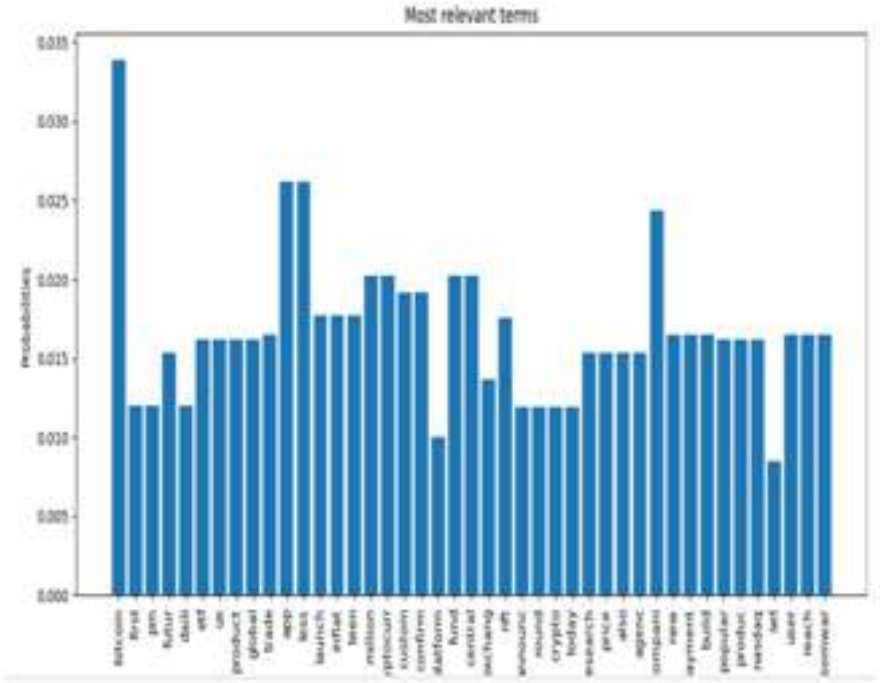


Fig. 3. Result of most relevant terms versus its probabilities.

Topic Modelling is used to extract trending topics on Newsfeed data related to cryptocurrency. 100 iterations were performed to get optimal results for retrieving 20 most trending topics and related top terms. Part of the output is shown in Fig. 4.

	content
0	The new ETF is set to start trading on Friday,...
1	Central banks control the circulation and supp...
2	It's no secret that NFTs have been a hot space...
3	As investors race to capitalize on surging int...
4	Assuming Wall Street embraces these ETFs, the ...

Fig. 4. Result of most trending topics.

Clustering is performed to identify how strongly the topics are related to each other. Hierarchical clustering is applied using the ward.D method in hclust() function. Resultant two clusters are the related topics as represented in the form of dendrogram shown in Fig. 5.

ISSN 2394 - 7780

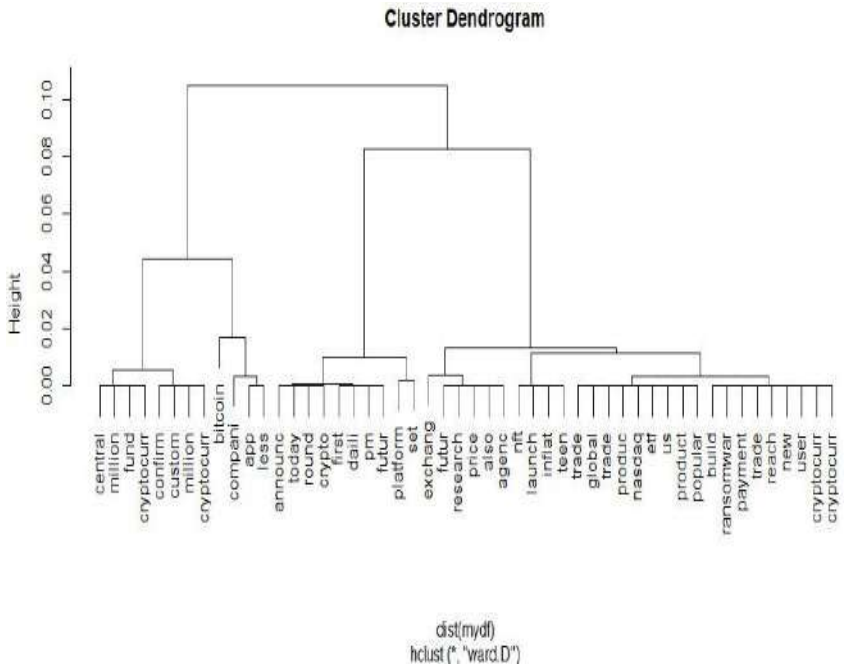


Fig. 5. Cluster Dendrogram representing clusters using ward.D method.

For identifying general sentiments over the topics, an NLP library of Python called spacy, is used.

V. EXPERIMENTAL RESULTS AND EVALUATION

The results reflect the positive outcome of the keywords. More frequently coined terms of cryptocurrency in news feed data are visually shown in the form of word cloud as shown in Fig. 6. Word Cloud is a visualization technique which uses the frequency of words in the corpus, to display words with sizes corresponding to their frequency.

Fig. 6. Most frequent terms in the form of Word Cloud related to cryptocurrency.

Probabilistic topic models such as LDA, are popular text analysis tools which provide predictive and latent topic representation of corpus. Chunksize controls how many documents are processed. Chunksize is set to 2000. Number of passes controls how we train the model on the entire corpus. LDA model output is represented in Fig. 7.

```
LdaModel(num_terms=525, num_topics=10, decay=0.5, chunksize=2000)
[(0, '0.034*"bitcoin" + 0.012*"first" + 0.012*"daili"'), (4, '0.020*"million" + 0.020*"fund" + 0.020*"
cryptocurr"'), (7, '0.024*"compani" + 0.016*"new" + 0.016*"cryptocurr"'), (3, '0.019*"million" + 0.019
*"cryptocurr" + 0.019*"custom"'), (6, '0.015*"futur" + 0.015*"research" + 0.015*"price"'), (2, '0.026*
"app" + 0.026*"less" + 0.015*"launch"')]
```

Fig. 7. LDA model output showing 6 documents with 3 terms each.

ISSN 2394 - 7780

LDA helps to create a Document Term probability matrix in which each cell represents the frequency count of word W_j in the document D_i of the corpus. The following probabilities are considered – P(topic t / document d) which denotes the proportion of words in a document currently assigned to topic t, and P(word w / topic t) which denotes the proportion of documents with word w assigned to topic t. With each iteration, the topic-document distribution tends to become steady. Output based on 100 iterations is represented in Fig. 8. The figure shows 10 documents, each with 10 words which have the most probabilities.

```
[(0, '0.034*"bitcoin" + 0.012*"first" + 0.012*"pm" + 0.012*"futur" + 0.012*"daili" + 0.012*"ralli" + 0
.012*"fund" + 0.012*"etf" + 0.012*"sure" + 0.012*"contract"'), (1, '0.016*"etf" + 0.016*"us" + 0.016*"
product" + 0.016*"global" + 0.016*"trade" + 0.016*"bitcoin" + 0.016*"build" + 0.008*"could" + 0.008*"s
et" + 0.008*"right"'), (2, '0.026*"app" + 0.026*"less" + 0.018*"launch" + 0.018*"inflat" + 0.018*"teen
" + 0.009*"space" + 0.009*"mani" + 0.009*"ceo" + 0.009*"said" + 0.009*"use"'), (3, '0.019*"million" +
0.019*"cryptocurr" + 0.019*"custom" + 0.019*"confirm" + 0.010*"platform" + 0.010*"two" + 0.010*"gener"
+ 0.010*"capit" + 0.010*"week" + 0.010*"onlin"'), (4, '0.020*"million" + 0.020*"cryptocurr" + 0.020*"f
und" + 0.020*"central" + 0.014*"exchang" + 0.014*"rais" + 0.014*"ventur" + 0.014*"invest" + 0.014*"ban
k" + 0.014*"control"'), (5, '0.018*"nft" + 0.012*"announc" + 0.012*"round" + 0.012*"crypto" + 0.012*"t
oday" + 0.012*"would" + 0.012*"plan" + 0.012*"compani" + 0.012*"question" + 0.012*"see"'), (6, '0.015*
"futur" + 0.015*"research" + 0.015*"price" + 0.015*"also" + 0.015*"agenc" + 0.015*"secur" + 0.015*"onc
hain" + 0.015*"data" + 0.015*"bitcoin" + 0.008*"analyst"'), (7, '0.024*"compani" + 0.016*"new" + 0.016
*"cryptocurr" + 0.016*"payment" + 0.016*"build" + 0.016*"stripe" + 0.016*"help" + 0.016*"broker" + 0.0
09*"seri" + 0.009*"led"'), (8, '0.016*"popular" + 0.016*"trade" + 0.016*"nasdaq" + 0.016*"produc" + 0.
008*"set" + 0.008*"space" + 0.008*"ethereum" + 0.008*"b" + 0.008*"behind" + 0.008*"breakdown"'), (9, '
0.016*"cryptocurr" + 0.016*"user" + 0.016*"trade" + 0.016*"reach" + 0.016*"ransomwar" + 0.016*"oper" +
0.016*"share" + 0.009*"close" + 0.009*"monday" + 0.009*"pressur"')]
```

Fig. 8.TopicWord Probability Matrix after 100 iterations.

Evaluation of sentiment analysis is done using statistical measures of precision, recall and f-score to assess its accuracy in different iterations. Precision, Recall and F-score calculations of five iterations are represented in Fig. 9.

```
Precision:0.4999999999358974    Recall:0.9999999997435898       F-Score:0.6666666665527066
Precision:0.4999999999358974    Recall:0.9999999997435898       F-Score:0.6666666665527066
Precision:0.4999999999358974    Recall:0.9999999997435898       F-Score:0.6666666665527066
Precision:0.4999999999358974    Recall:0.9999999997435898       F-Score:0.6666666665527066
Precision:0.4999999999358974    Recall:0.9999999997435898       F-Score:0.6666666665527066
Review text: at the same time bitcoin was surging above  around  pm utc monday ether the second larges
t cryptocurrency by market capitalization also set an alltime high hitting  accor
Predicted sentiment: Positive   Score: 0.9999545812606812
```

Fig. 9. Evaluating the predicted sentiment with score.

Recall gives the metrics that measure the proportion of relevant keywords among recommended keywords. Precision measures the retrieved recommended keywords to the actual keywords. F-score represents the effectiveness of retrieval combining precision and recall. Predicted sentiment is positive with a score of 99.9%.

CONCLUSION

Being one of the trending topics in the financial sector recently, news websites frequently publish articles related to Cryptocurrency and other related topics. Topic modelling is used as a text mining technique for identifying the topics from a collection of news articles related to cryptocurrency and then applying hierarchical clustering for finding a connection between these topics. Visualization tools are used to present the topics and their probabilities, and the association among them. LDA model was used as the topic modelling technique to extract the most trending topics from the collected news articles after initial pre-processing. Sentiment analysis performed by the Natural Language Processing library spaCy revealed that the sentiment reflected recently was mostly positive.

REFERENCES

[1] Jonathan Chiu and Thorsten Koeppl, The economics of cryptocurrencies – bitcoin and beyond (2017).

[2] Shailak Jani, The growth of cryptocurrency in India: its challenges & potential impacts on legislation (2018).

[3] Toqir A. Rana, Yu-N Cheah & Sukumar Letchmunan, Topic modelling in sentiment analysis: a systematic review (2016).

[4] Hye-Jin Kwon, Hyun-Jeong Ban, Jae-Kyoon Jun, and Hak-Seon Kim, Topic modelling and sentiment analysis of online review for airlines (2021).

ISSN 2394 - 7780

[5] MaibamDebina Devi and NavanathSaharia, Exploiting topic modelling to classify sentiment from lyrics (2020) pp 411-423.

[6] Rania Albalawi, Tet HinYeap, and MoradBenyoucef, Using topic modelling methods for short-text data: a comparative analysis (2020).

[7] Khusbu Thakur and Vinit Kumar, Application of text mining techniques on scholarly research articles: methods and tools (2021).

[8] Xing Fang and Justin Zhan, Sentiment analysis using product review data (2015).

[9] Sangeeta Rani, Nasib Singh Gill, and PreetiGulia, Sentiment analysis and topic modelling on twitter for clean India mission (2021).

[10] Bisma Shah, and Farheen Siddiqui, Sentiment analysis approaches for social media monitoring (2018).

[11] Asmussen, C.B., Møller, C. Smart literature review: a practical topic modelling approach to exploratory literature review. *J Big Data* 6, 93 (2019). https://doi.org/10.1186/s40537-019-0255-7.

[12] Onan, A.; Korukoğlu, S.; Bulut, H. (2016): LDA-based Topic Modelling in Text Sentiment Classification: An Empirical Analysis. International Journal of Computational Linguistics and Applications, 7(1), pp. 101–119.

[13] D. Blei, A. Ng, M. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

ASCERTAINING THE ROLE OF GENDER IN INFLUENCING PERCEPTION OF STUDENTS TOWARDS VIRTUAL CONFERENCES DURING COVID-19: A STUDY IN KOLKATA METROPOLIS

Samuel S Mitra[1], Peter Arockiam A.[2], Joseph K[3], Dr. Milton Costa[4] and Payal Sharma[5]
[1]Staff and Research Scholar in Commerce, St. Xavier's College (Autonomous), Kolkata, West Bengal, India
[2]Vice Principal of Commerce (Evening) & BMS and Financial Administrator, St. Xavier's College (Autonomous), Kolkata, West Bengal, India
[3]Vice Principal of Commerce (Morning) and Assistant Director of Library, St. Xavier's College (Autonomous), Kolkata, West Bengal, India
[4]Assistant Professor in Commerce and Management, St. Xavier's University, Kolkata, West Bengal, India
[5]Assistant Professor in Commerce (Morning) and Research Scholar, St. Xavier's College (Autonomous), Kolkata, West Bengal, India

ABSTRACT
The Covid-19 lockdowns imposed certain restrictions in outdoor activities and triggered catastrophic impacts on all sectors of the economy. In this light, the sacrosanct sector of education is no exception. To combat the menacing pandemic, every institutions across the globe had to be shut down due to the lockdowns, as a result getting a move on was felt quite difficult. However, the ambit of technology has emerged as an impeccable talisman to rescue the pandemonium and keep the hallowed milieu of education afloat. The facilitation of teaching and learning facilities along with a plethora of other academic activities conducted through digital technologies online has been a force to reckon. In this context, capturing the perceptions of individuals towards such online academic endeavours seems pivotal. The current research study is attempted at examining and analyzing the attitudes and behaviour of the students towards virtual conferences by the application of Technology Acceptance Model (TAM). For this purpose, a survey has been conducted on 334 college students in the metropolitan setting of Kolkata. The findings reveals that such students have a strong penchant towards virtual conferences as there exists a positive and significant relationship amongst different dimensions of TAM constructs with the attitude and behaviour of students amidst COVID-19 pandemic. The study also revealed that gender had no impact on perceptions of students towards online conference.

Keywords: COVID-19; Virtual Conferences; Attitudes and Behaviour; Technology Acceptance Model (TAM); Students

INTRODUCTION TO THE STUDY
In the backdrop of this fast paced world, where rapid technological advancements are exhibiting myriad uncanny pyrotechnics it has been witnessed that there has been a bewildering intensification in the use of digital technologies, largely triggered by the dramatic rise in the usage of internet, smartphones and relentless percolation of technologies in the ambit of E-commerce. The Indian consumers are fast relinquishing the offline activities for the innovative comfy online services. In this context, the burgeoning of virtual conferences and their acceptance by students has been of the striking revelations during the COVID-19 pandemic. Virtual Conferences also termed as E-Conferences are held online by various digital platforms like Zoom, Google Meet, Webex and Microsoft Teams. The usage of such online tech-savvy platforms has boosted the current life of students as well as academicians to keep education upbeat. Contrary to this, wherein offline conferences had certain travails like physical travelling, expenses, spoil of energy, wastage of time and much more. What is more important is that, online conferences is also serving the need of the situation when social distancing is to be strictly maintained and harsh COVID-19 protocols need to be followed while still staying at home for the most of the time. A well-observed fact in the ambit of academia during these challenging times has been the adoption and usage of digital technologies for conducting online academic endeavours like teaching-learning, online FDP, online MDP, online conferences, online webinars and many more such eye-twitching activities. The surge in the acceptance and usage of such online activities among the Indians and all across the world bears a strong testimony to this fact. The present research study purports to investigate the intrinsic motivations, perceptions and adoption mechanisms of college students towards online conferences in the context of the COVID-19 pandemic. The current research study has been undertaken in city of Kolkata by surveying as many as 334 respondents who are college students.

One of the arduous challenges for any academic researcher lies in augmenting the current level of consciousness of multiple factors which trigger the mobile wallets to be accepted and adopted especially in times of Covid-19 when talking in the context of Technology Acceptance Model (TAM), where the crux objective is to probe into

ISSN 2394 - 7780

the underlying perceptions, motivations, attitudes and behavioural intentions of Indian consumers towards these agile mobile applications. TAM is an information system model describing a number of decisions which influence how consumers would accept and use a new technology when presented with it. In the current research study, we have rejigged the model of TAM to include concepts of 'Subjective Norm' and 'Exigencies (Covid-19).' Albeit, few researchers in the past have blended the various components of 'Attitude-Intention-Behaviour', in this particular research study we intend to emphasize on conventional TAM with certain minor modifications.

LITERATURE REVIEW

TAM is an extension of the "Theory of Reasoned Action" (TRA). TAM is an amelioration over TRA as it was built on certain independent variables like "Perceived Usefulness" (PU) and "Perceived Ease of Use" (PEOU) as well as dependent variables like "Attitude towards Usage" (ATU). Fred Davis had coined the term PU as the degree or extent to which a person believes that using a particular system would lead to enhancement of his or her job performance. Davis (1989) defined PEOU as the degree or extent to which a person believes using a particular system would be free from effort. Further adding to the theory, Davis (1993) said that the usage of actual information system was determined by a concept called Behavioural Intention (BI), which was determined jointly by the users' attitude towards the use of the system and perceived usefulness. He defined it as "the subjective probability that an individual will perform a specified behaviour." Attitude towards Usage (ATU) is a crux dependent variable in the TAM and in the words of Ajzen & Fishbein (2000), "it is the evaluative effect of positive or negative feeling of individuals in the usage of a particular system." With the passage of time, the concept of TAM began bolstering from the dynamics of retrospective information technology to integrate novel concepts like e-commerce and m-commerce. Lin et al. (2008), in their study focused on the application of TAM, in that they endeavoured to corroborate the influence of crux elements like mobile trust, perceived usefulness, perceived ease of use and service fee on wireless mobile data service categorizing them as independent variables having an inexorable impact on the customer adoption of SMS technology. It is noteworthy that the traditional model of TAM has also proven to be quite flexible to include independent constructs most notably 'Subjective Norm', as first introduced by Taylor & Todd (1995), who defined it as "the influence gained from social circle on whether or not to use a particular." TAM is still being relentlessly studied and expanded. "The two major upgrades under the umbrella of TAM are TAM 2 (Venakatesh & Davis 2000 and Venkatesh 2000) and Unified Theory of Acceptance and Use of Technology (UTUAT, Venkatesh et al. 2003)." As per Venkatesh & Bala, 2008, "TAM 3 has also been proposed in the context of e-commerce which would include the effects of trust and perceived risk on system use." Though very few researches has been conducted in the past taking into consideration the concept of Subjective Norm in the context of TAM, we find it extremely grueling to come across the domain of Exigencies used as a construct in TAM. This in fact is the essence of our present research study.

RESEARCH OBJECTIVES

1. To introduce a novel Technology Acceptance Model for capturing student perception towards e-conferences
2. To probe into the perception of the Indian students towards virtual conferences.
3. To explore if gender has any effect on perceptions of students towards online conferences.

Development of Research Model and Hypothetical Statements

The below model is a re-modified TAM. The constructs namely 'Subjective Norm' and 'Exigencies (Covid-19)' has been incorporated to cater to the influence of peer groups and urgent unforeseen needs respectively. Therefore, our research model comprises of 6 constructs, which has been developed and presented below.

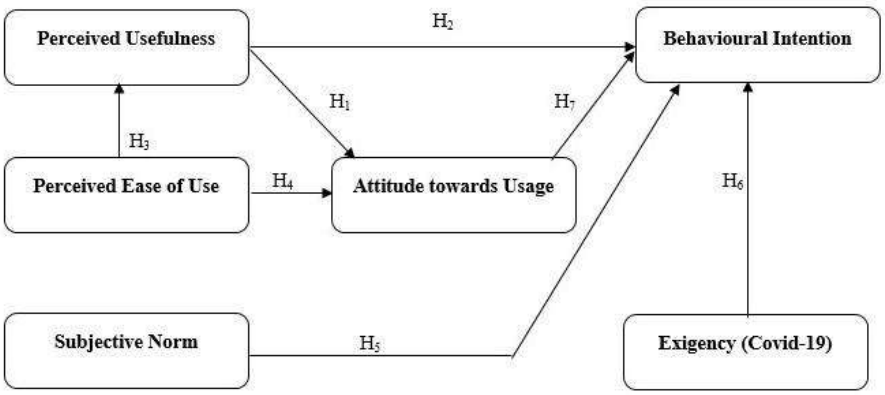


Figure 1: Research Model

ISSN 2394 - 7780

The current research study purports to develop a research framework for the user acceptance and intention of mobile wallets, pillared on a re-modified Technology Acceptance Model (TAM). For this purpose, the following hypotheses have been developed and substantiated through the above research model represented above.

H_1: "Perceived Usefulness has a positive influence on Attitude towards Usage"

H_2: "Perceived Usefulness has a positive influence on Behavioural Intention"

H_3: "Perceived Ease of Use has a positive influence on Perceived Usefulness"

H_4: "Perceived Ease of Use has a positive influence on Attitude towards Usage"

H_5: "Subjective Norm has a positive influence on Behavioural Intention"

H_6: "Exigencies (Covid-19) has a positive influence on Behavioural Intention"

H_7: "Attitude towards Usage has a positive influence on Behavioural Intention"

DATA AND METHODOLOGY

A rigorous research was conducted including both primary and secondary data. Secondary data was used to create a robust foundation of conceptual framework of the present research study. For this purpose, several research papers has been acquired from various authentic and reliable e-resource sites like EBSCO, BASE and Google Scholar. For the purpose of primary data collection, a survey has been conducted on a total sample size of 350 respondents, in the age group of below 18 to 55 years and engaged in education and various occupations. For the purpose of data collection a close-ended questionnaire was developed. Most of the questionnaire were mailed while others were randomly doled out to the respondents. The respondents in the present study form the residents of Kolkata. The questions in the questionnaire were mostly self-developed albeit few questions have been adopted from previous researches. The questionnaire contained 22 questions under 6 segments, namely, "Perceived Usefulness" (PU), "Perceived Ease of Use" (PEOU), "Subjective Norm" (SN), "Exigencies" (Covid-19), "Attitude towards Usage" (ATU) and "Behavioural Intention" (BI). A "Five-point Likert Scale" has been used to measure the concepts. There were few responses which were erroneous and some were not returned, hence, those responses had to be rejected. After the rejection of such responses, the final valid responses stood at 334. The data obtained has been processed by using SPSS version 23.0.

ANALYSIS AND PRESENTATION OF DATA

- **Demographic Profiling**

Table 1: Representation of Demographic Statistics

Demographic Construct	Classification	Population Statistics	Percentage
Gender	Male	178	0.53
	Female	156	0.47
	TOTAL	**334**	**1.00**

We observe that the ratio of male and female is moderately balanced in the proportion 178:156 where the total respondents are 334.

- **RELIABILITY ANALYSIS**

A reliability analysis has been conducted to check the internal validity and consistency of the items used for each factors. For conducting, reliability statistics, IBM SPSS version 23 has been used. As per Nunnally (1978), "questionnaire for various factors are judged to well reliable measurement instrument, with Cronbach's Alpha scores being all above 0.6." The reliability statistics prove that the "Cronbach's Alpha" score were above the standard value of 0.6, thus, validating that all the 22 items fit perfectly in our questionnaire and support our proposed research model. Cronbach's Alpha is actually a measure of internal consistency of items, implying the related closeness of a set of items as a group. In the present study, it serves as a basis for measuring the scale reliability of the items used in the questionnaire.

Table 2: Reliability Statistics for all variables (n=22)

Cronbach's Alpha	Cronbach's Alpha based on Standardized Items	N of items
0.812	0.812	22

ISSN 2394 - 7780

- **Correlation Analysis**

After conducting the reliability analysis, it is vital to find out the relationship between the 6 factors as well as to examine the hypotheses of our proposed research model. To serve this purpose, we have conducted a correlation test by using SPSS version 23. The below table shows that the correlation between PEOU, PU, ATU, SN, EX (Covid-19) and BI are positive and significant, thereby, confirming, our original hypotheses made in the literature related to TAM. The correlation statistics has been presented below.

Table 3: Representation of Correlation Matrix

Factor		PEOU	PU	ATU	SN	EX (Covid)	BI
PEOU	Pearson Correlation		0.738**	0.763**	0.586**	0.751**	0.680**
	Sig. (2-tailed)	1	.000	.000	.000	.000	.000
	N		334	334	334	334	334
PU	Pearson Correlation	0.738**		0.768**	0.645**	0.705**	0.728**
	Sig. (2-tailed)	.000	1	.000	.000	.000	.000
	N	334		334	334	334	334
ATU	Pearson Correlation	0.763**	0.768**		0.606**	0.610**	0.712**
	Sig. (2-tailed)	.000	.000	1	.000	.000	.000
	N	334	334		334	334	334
SN	Pearson Correlation	0.586**	0.645**	0.606**		0.542**	0.587**
	Sig. (2-tailed)	.000	.000	.000	1	.000	.000
	N	334	334	334		334	334
EX (Covid)	Pearson Correlation	0.751**	0.705**	0.610**	0.542**		0.786**
	Sig. (2-tailed)	.000	.000	.000	.000	1	.000
	N	334	334	334	334		334
BI	Pearson Correlation	0.680**	0.728**	0.712**	0.587**	0.786**	
	Sig. (2-tailed)	.000	.000	.000	.000	.000	1
	N	334	334	334	334	334	

- **REGRESSION ANALYSIS**

To further bolster our research findings, we have also conducted a regression statistics to test the different proposed hypothesis. First, we examine the relationship between H_1 and H_4.

Table 4: Regression Statistics

Table: Predictors: PU & PEOU → Dependent Variable: ATU

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of Estimate
1	.876[a]	.717	.713	.60436

a. Predictors: (Constant), PEOU, PU

Coefficients[a]

Model	Unstandardized Coefficients		Standard Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	.378	.193		1.534	.143
PEOU	.363	.050	.381	6.826	.000
PU	.587	.057	.553	10.342	.000

a. Dependent Variable: ATU

As we can see from the above table, the value of R square indicates that the two predictors (PU, PEOU) explains 71.7% variations in ATU. It explains the rationality of this model, albeit there might be other oblivious factors having an impact on the respondents' ATU. The standardized coefficients (β) shows that PU (β=0.553)

ISSN 2394 - 7780

have a larger impact than PEOU (β=0.381). Also, the Sig. indicates that both of the predictors have a significant and positive impact on ATU scores being less than 0.001 level. "Unstandardized Coefficients" are those which are obtained after conducting the regression test on variables which are measured in their original scales and it reveals the impact of each of the individual variable on the dependent variable, while "Standardized Coefficients" are obtained by the regression test on rescaled variables and it reflect the comparison the impact of various predictors on the final outcome.

Table 5: Regression Statistics

Table: Predictors: PU, SN, EX (Covid-19) & ATU ⟶ Dependent Variable: BI

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of Estimate
1	.895[a]	.724	.707	.57964

a. Predictors: (Constant), PU, SN, EX (Covid-19), ATU

Coefficients[a]

Model	Unstandardized Coefficients		Standard Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	.319	.186		1.770	.093
PU	.599	.072	.558	8.768	.000
SN	.285	.061	.287	4.376	.000
EX (Covid-19)	.306	.062	.315	5.545	.000
ATU	.383	.067	.394	6.425	.000

a. Dependent Variable: BI

From the above table it is confirmed that all the four predictors namely PU, SN, EX (Covid-19) and ATU had a significant and positive influence on BI, with (β=0.558), (β=0.287), (β=0.315) and (β=0.394) respectively for each predictor. Each of the four predictors have Sig=0.

Finally, we conduct a regression analysis to examine H_3.

Table 6: Regression Statistics

Table: Predictors: PEOU ⟶ Dependent Variable: PU

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of Estimate
1	.809[a]	.687	.690	.56438

a. Predictors: (Constant), PEOU

Coefficients[a]

Model	Unstandardized Coefficients		Standard Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	.378	.193		1.534	.143
PEOU	.363	.050	.381	6.826	.000

a. Dependent Variable: PU

Finally, one more determination of a regression model was done to test our fourth hypothesis, i.e. influence of PEOU on PU. As evidenced from the above table, the value of R Square is 0.687 which represents that PEOU explains 68.7% variations in PU. We also notice that Standard Coefficient value is (β=0.381), PEOU had a significant and positive impact on PU. Hence, our proposed research model along with the hypotheses are rightly proven correct as evidenced by the robust examination and analysis.

ISSN 2394 - 7780

- **Chi-Square Test**

It is an important objective of our current research study to explore the relationship between demographic variable of gender among college students and perception towards virtual conference. Chi-Square Test would help us to see whether the observed frequencies in the data results are supporting the relationship or not.

Table-7. Gender and Perception towards Virtual Conferences

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	2.331[a]	4	.782
Likelihood Ratio	2.377	4	.769
Linear-by-Linear Association	1.634	1	.843
N of Valid Cases	334		

DELIBERATIONS OF RESEARCH FINDINGS

This study was a pioneering effort in the application of virtual conferences into a TAM model which was re-modified especially by including the novel constructs of "Subjective Norm" and "Exigencies (Covid-19)". According to our research findings, "Perceived Usefulness" (PU) had a significant impact on "Attitude towards Usage" (ATU). It was also observed that PU was significantly related to "Behavioural Intention" (BI). The reason behind this could be that students are willing to adopt a beneficial technology that could make their life more convenient. "Perceived Ease of Use" (PEOU) was significantly related to ATU. Furthermore, the domain of "Subjective Norm" (SN) which deals with the influence of social circle had a significant impact on the "Behavioural Intention" (BI) of the students. Social interactions in today's world plays a pivotal role in shaping the perception and attitude of people, in this case, the perception and attitude of respondents towards online conferences. We also notice that "Exigencies (Covid-19)" also influence the attitude of students as evidenced by our research findings. In fact, the spectacular adoption and usage of online education endeavours in the era of Covid-19 is a fact well-documented. This shows, that the perception of students will be quite different in case of urgent need or emergency situation or any other kind of exigencies. The findings of the present research study also proved that PEOU had a strong influence on PU, suggesting that providing adequate user training is vital for fine-tuning the consumers' perception about the usefulness of a technology which is quite new. Ultimately, we comprehend that the attitude of students has been prodigious in shaping up their behaviour as both psychological and physiological faculties are a nifty driving force in the development of perceived likelihood of students.

CONCLUSION TO THE STUDY

The findings of the present study further highlights the importance of the study as the various hypothesis put forth has been proved by robust empirical results which have been obtained. The face of technology has witnessed a seismic shift, particularly in the recent years because of a plethora of instauration. The digital technological platforms possess uncanny qualities supporting myriad technological applications which has proven to extremely useful for the many people. Amidst this pandemic, many tech-savvy digital academic platforms have served the purpose of students by absorbing their despondency. We have been able to bring out the effectiveness of virtual conferences when talking in the context of perception of students towards it especially in the COVID-19 era. The present research study has highlighted certain crux components under the TAM constructs, which will be quite useful to guide future researches. College Students in Kolkata highly value virtual conferences. A reason for this lies in the confidence and the eagerness among the erudite students who are always ready to embrace new-fangled technology.

FUTURE SCOPE

A similar study can be conducted finding the perception of teachers towards not only virtual conferences but also towards online mode of teaching in the context of COVID-19 pandemic. Furthermore, it would also be worthwhile to take into account the influence of various learning apps on both students and teachers. To bolster the worth of the study, a larger sample should be selected and surveyed as well as the study could also be conducted across a wider geographical boundary to gain more insights about perception of students and teachers towards various aspects of online academic activities.

REFERENCES

1. Ali, W. (2020). Online and remote learning in higher education institutes: A necessity in light of COVID-19 pandemic. Higher Education Studies, 10(4), 16-25.

ISSN 2394 - 7780

2. Allo, M. D. G. (2020). Is the online learning good in the midst of Covid-19 pandemic? The case of EFL learners. Journal Sinestesia, 10(1), 1-10.

3. Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behaviour. Englewood Cliffs, NJ: Prentice Hall.

4. Ajzen, I., & Fishbein, M. (2000). Attitudes and the attitude-behavior relation: Reasoned and automatic processes. European Review of Social Psychology, 11(1), 1-33.

5. Bhaumik, R., & Priyadarshini, A. (2020). E-readiness of senior secondary school learners to online learning transition amid COVID-19 lockdown. Asian Journal of Distance Education, 15(1), 244-256.

6. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340.

7. Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. Management Science, 35(8), 982-1003.

8. Davis, F. D. (1993). User acceptance of information technology: System characteristics, user perceptions and behavioural impacts. International Journal of Man-Machine Studies, 38(3), 475-487.

9. Jena, P. K. (2020). Impact of COVID-19 on education in India. Purakala, 31(46), 142-149.

10. Luam, P., & Lin, H. H. (2004). Towards an understanding of the behavioral intention to use mobile banking. Computer in Human Behaviour, 21(6), 340-348.

11. Mohalik, R., & Sahoo, S. (2020). E-readiness and perception of student teachers' towards online learning in the midst of COVID-19 pandemic. Retrieved from https://ssrn.com/abstract=3666914

12. Nagar, S. (2020). Assessing students' perception towards e-learning and effectiveness of online sessions amid COVID-19 lockdown phase in India: An analysis. Tathapi, 19, 272-291.

13. Ngampornchai, A., & Adams, J. (2016). Students' acceptance and readiness for elearning in Northeastern Thailand. International Journal of Educational Technology in Higher Education, 13(34), 1-13.

14. Nunnally, J. C. (1978). Psychometric theory (2nd Ed.). New York: McGraw-Hill.

15. Roy, S. (2017). Scrutinizing the factors influencing customer adoption of App-based cab services: An application of the technology acceptance model. IUP Journal of Marketing Management, 16(4), 54-69.

16. Smith, C., Hoderi, M., & Mcdermott, W. (2019). COVID-19 impact on education: Quality assurance and recognition of distance higher education and TVET. Academy of Educational Leadership Journal, 23(2).

17. Taylor, S., & Todd, P. A. (1995). Understanding information technology usage: A test of competing models. Information Systems Research, 6(2), 144-176.

18. Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. Management Science, 46(2), 186-204.

19. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425- 478.

20. Venkatesh, V., & Bala, H. (2003).Technology acceptance model 3 and a research agenda on interventions. Decision Sciences, 39(2).

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

SECURITY AND PRIVACY CONCERNS IN WEARABLE DEVICES

[1]Dr. Swapnali Lotlikar and [2]Ms. Alicia Cotta
[1]Assistant Professor and [2]MScIT, Part I Student, Usha Pravin Gandhi College of Arts, Science and Commerce Mumbai

ABSTRACT
The wearables industry has developed substantially in the last decade, notably in the fitness and e-health tracker sectors. These trackers provide new capabilities that need a huge amount of personal data to be collected. Fitness trackers have been the victim of targeted assaults such as eavesdropping, unauthorized account access, and fraudulent software upgrades as a consequence. Since the arrival of the internet, security has been a risk and a highly debatable topic. In terms of data sharing on the internet, privacy has also been a critical problem. Wearable technology poses comparable concerns, and in some respects, they are amplified by the sensitivity of the data collected by these devices. There are many different sorts of wearable devices, each with its own set of issues and drawbacks, but security and privacy are considerations that apply to all forms of wearable devices, whether big or small. The objective of this paper is to get a better understanding of the various ways in which today's wearable technology may violate our personal privacy and security, as well as suggest ways for organizations to tackle concerns connected to the privacy and security of their wearable devices.

Keywords: fitness tracker, security, privacy

INTRODUCTION
As a consequence of technological developments, the use of wearable technology has gained in popularity recently. As more companies strive to make current CPUs smaller and more effective, and develop new sensor devices that can measure even more body information with better precision, more companies are seeing this as an opportunity to bring technology into everyday goods like garments and other items. [1] Wearable technology's advantages and possibilities are evolving constantly. Wearable technology has advanced in the previous decade to collect more sensitive data such as heart rate, blood pressure, oxygen levels, and much more. Establishing customer trust, security and privacy are also a crucial part. Wearable devices have emerged as a powerful tool for personal fitness, allowing proactive individuals to establish personal health and wellness objectives, track their progress, and push themselves to achieve greater goals. While wearables have a variety of applications for personal health, they also raise some serious ethical issues. Questions have been raised, in particular, concerning the role of organizations that capture and handle the data collected by wearable devices. These ethical concerns about wearables touch on a number of critical issues such as individual privacy rights and data permission. The goal of this paper is to get a deeper understanding of the various ways in which today's wearable technology might violate our personal security and privacy, as well as the steps taken by the firms that manufacture these to prevent such a violation.

BACKGROUND
Wearable technology is a form of technology that may be worn on certain areas of the body, as the names indicates. These are minicomputers that have been developed in the shapes of wrist worn bands, chest worn straps, and several other items. They are intended to be easy to wear for long periods of time, and are typically made of materials that aren't associated with electrical equipment, such as fabric and rubber or silicone as a substitute of metal and plastic. [1] They can either assist the user in simplifying routine chores or allow the user to accomplish them in a far more efficient manner while they are worn. They are intended to enhance and improve a user's life by delivering valuable information specific to them, such as the ability to detect the quality of their sleep and heart rate, or just receiving useful alerts like reminders and phone calls. Some devices are called "complementary," in that they require other sort of primary device to operate and, in general, communicate between these devices via a wireless connection, with certain devices preferring to link to the primary device via a physical connector. Typically, these devices include a number of sensors that work together to complete a task. These sensors include accelerometer, pedometer, and heart rate sensor, which are most commonly used.

In a number of digital situations, researchers investigated consumers understanding of data gathering as well as their attitudes about privacy and security. The majority of research focus on web/online environments or mobile technology. Some studies looked at security and privacy in a broader sense, examining a variety of existing and future technologies while others concentrated on digital apps and social media environments. IoT devices, wearables, and fitness trackers have received relatively less attention. Wearable devices have been conquering

ISSN 2394 - 7780

the gaming sector; we can see that VR (virtual reality) headsets have begun to acquire more attention and appeal in recent years, and this is due to the fact that VR headsets have come to be utilized in more creative ways to entertain people. For instance, today's games may be applied to immerse the user and make them feel like they are a part of the game. Another usage for VR headsets is in health care, where they are used to treat individuals suffering from PTSD (post-traumatic stress disorder). [1] Fitness trackers are one of the most prevalent types of wearables presently being used by customers today, and many of them feature a variety of sensors aimed at measuring various body-related statistics. Companies also go so far as to incorporate sensors that can detect rate of breathing and body temperature, in the hopes of providing users in maintaining healthier lives.

METHODOLOGY

Null Hypothesis: The gender of the individual does not affect the security and privacy risks of population of individuals owing wearables or

Alternative Hypothesis: The gender of the individual does affect the security and privacy risks of population proportion of individuals owing wearables.

The analysis of wearable devices security and privacy, was done with the help of numerous secondary resources and some primary research that was performed through a survey with around 72 participants. The survey was conducted to get a fair idea about the knowledge and attitude of the users and non-users towards the fitness tracker data collected. Analysis of the collected data will give a look into what the fitness tracker users think of the data collected by their fitness trackers.

An online survey for fitness tracker users was conducted to find out what they think about the following: rust in the understanding of data collecting and usage methods, as well as their understanding of the plausibility and possibility of privacy breach and risks; attitudes about security and privacy breaches and sharing preferences of the data

Respondents were to think about data their fitness tracker generated and to respond to questions asked in [2]:

A. Data Sensitivity: How concerned would you be if your fitness tracker data were compromised, through a breach of security at the company's end? (Very Concerned, Somewhat Concerned, Not at all Concerned; Mean=0.70, SD=0.03)

B. Personal Data Value: Compared to other kinds of personal data (like financial information) how valuable is your fitness tracker data to you? (Very Valuable, Somewhat Valuable, Not at all Valuable; Mean=0.58, SD=0.03)

C. Data Value to advertisers: Compared to other kinds of personal data (like financial information) how valuable do you think is your fitness data is to third-party advertisers? (Very Valuable, Somewhat Valuable, Not at all Valuable; Mean=0.65, SD=0.03)

Other questions asked were related to the sharing activities of the user:

1. Do you share your fitness tracker statistics online? (Yes, No; Mean=0.22, SD=0.04)
2. Do you participate in group fitness activities? (Yes, No: Mean=0.38, SD=0.05)

These questions were asked to gain information about the sharing tendencies of the participants. The participants were also asked what their biggest concerns regarding the use of wearable devices was from a list of concerns including privacy, security, health and extensive usage. The results revealed that a majority of participants were concerned about the security and privacy factors.

KNOWLEDGE

Regardless of the fact that fitness tracker security has improved as a result of these research, in [3,4] it has been demonstrated that new security measures are still required. The authors in [3] evaluated the privacy and security of data transfers in eight common fitness trackers. They discovered security flaws in the majority of the devices they looked at, and just one manufacturer used BLE privacy. In [4] the authors conducted a security study of the representative sample of currently available fitness trackers. They concentrated on malicious user settings that attempt to insert misleading data into cloud-based applications, resulting in inaccurate data analytics. The fitness tracker does not have data integrity checks, and the acquired data is saved in simple text on the smart phone, according to their findings.

On activity based on social networks, there has been a recent trend of sharing information and competing with peers or mutual connections [5]. Because effective and granular access control over data must be adequately given, this would pose even more security and privacy concerns.

ISSN 2394 - 7780

In [6], the authors provide a table which lists the qualities that may be controlled, which are divided into basic and additional categories. The data provided by the user while creating an account and registering their device, as well as information obtained by the tracker and technical details about the device when it synchronizes with the mobile application, are referred to as "basic attributes". "Additional attributes" allow users to share more information in order to take use of additional services (e.g., linking a Google account, messaging, and monitoring meals) or to personalize their profile (e.g.: profile pictures).

Data type	Basic Attributes	Additional Attributes
Personal data (PD)	Gender Age Date of birth Height Weight	Food log Water log Location Profile photos Alarms
Contact info (CI)	email	Contacts Basic info from other accounts (e.g., social networks)
Activity data (AD)	Steps Floors Heart rate Sleep quality	
Device info (DI)	Battery level Sync time App info	

Fig. 1. Data managed by fitness trackers

PRIVACY AND SECURITY CONCERNS

Many users choose wireless connections because they might be less troublesome and offer them with a degree of control that is impossible to attain with traditional wired connections. While several customers prefer wireless connections for wearable devices, there are many dangers and downsides that might possibly lead to privacy and security violations. [1] Many systems depend on their wireless connections to accomplish their many varied duties. Most businesses prioritize low latency (time taken to send the data) while ensuring that their connections are safe using the most advanced encryption methods.

Bluetooth Low Energy (or BLE) is an instance of a connectivity that has the potential to compromise a user's privacy and security. Manufacturers, in their eagerness to get their products to market, frequently overlook the security flaws they may be leaving up to all types of attackers. Some of the devices broadcasted a unique ID, defeating the purpose of masking, whereas the other would only alter the last few bytes of their MAC addresses, making them clearly identifiable and traceable. [7] Because of the inadequate implementation of Bluetooth devices, it is significantly easier for many people to identify devices within a specific range and to carry out various cyber-attacks, such as man-in-the-middle attacks, in which users may easily listen in on other users.

Most individuals are only concerned about their privacy if it concerns to wearable technologies. This is because they do not want their data to be available online, and many individuals are cautious about what they post for others to see.

RESULTS

A total number of 72 participants responded to the survey. The data collected from these different individuals from different age groups. These respondents were asked various questions related to the data usage of fitness trackers. From the total number of 72 participants, 36 females (50%) and 36 males (50%). The age groups ranged from 10 to 39 years and most of the participants belonged to the age group of 20 to 19 years (86.1%).

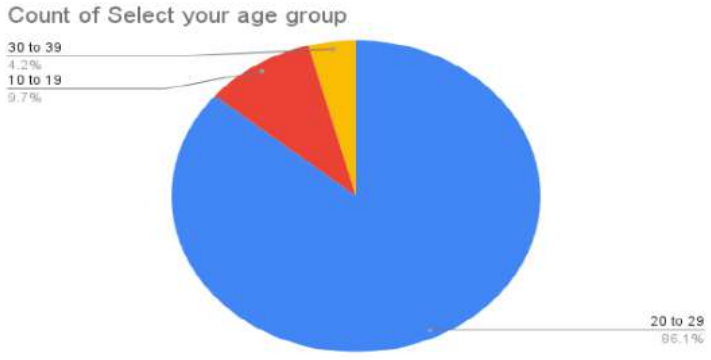


Fig. 2. Count of age groups of participants

ISSN 2394 - 7780

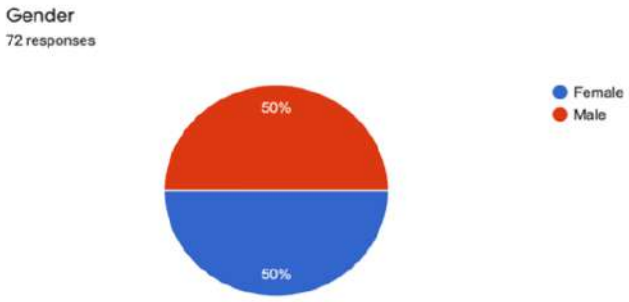


Fig. 3. Count of gender of respondents.

Furthermore, within the sample 42 individuals (61.1%) owned a wearable activity tracker (18 women and 24 men), whereas 26 individuals (38.8%) did not use such a device (16 women and 10 men).

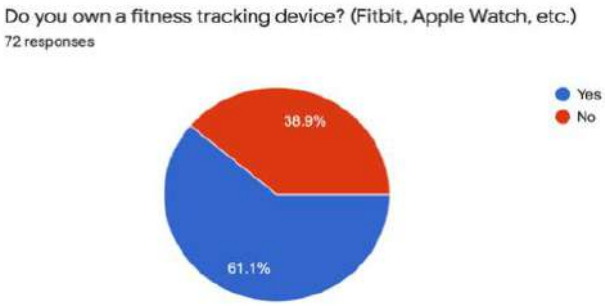


Fig. 4. Count of individuals owning a fitness tracker device

A major part (77.8%) of the total participants said that they do not share their fitness data on the internet.

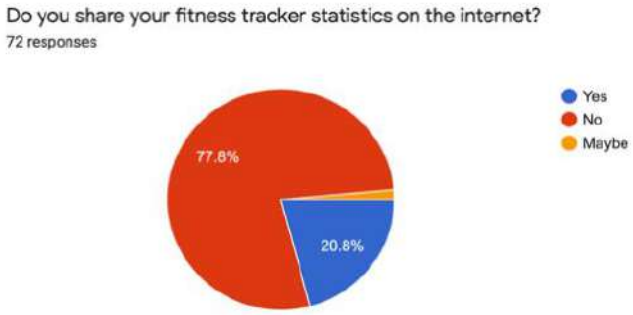


Fig. 5. Count of willingness of individuals to share their fitness data on the internet

When asked whether the participants were aware of the data policy of the fitness tracker company, 56 of 72 (77.8%) answered no showing that large percentage of participants had no knowledge about the company data policies.

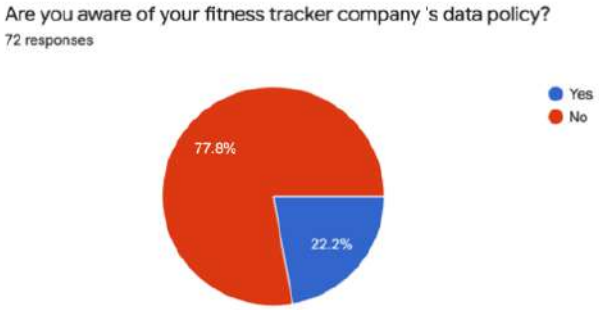


Fig. 6. Result of question related to awareness about the fitness tracker company data policy

ISSN 2394 - 7780

As can be seen in the figure below, the majority of individuals who participated in the survey were concerned about security and privacy when using their fitness trackers.

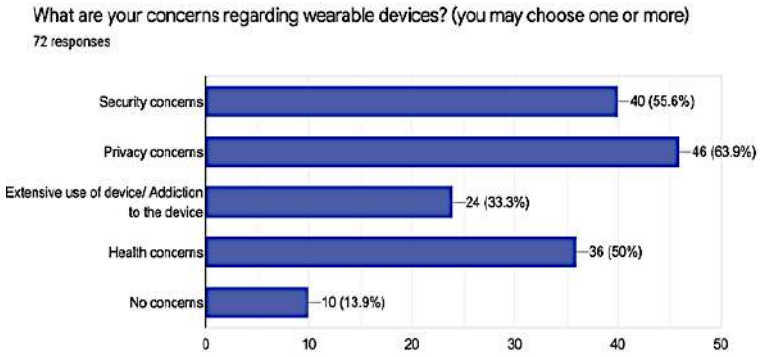


Fig. 7. Result of question related to concerns regarding wearable devices.

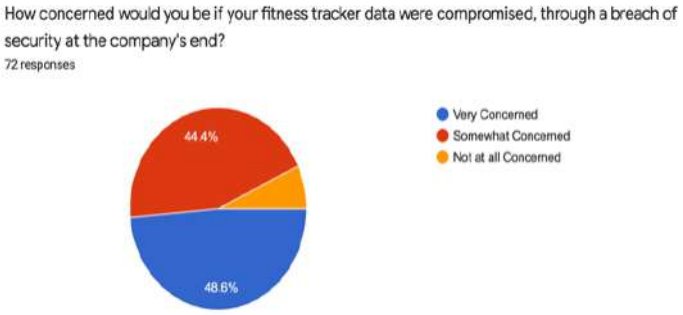


Fig. 8. Result of concern related to the breach of fitness data at the company's end

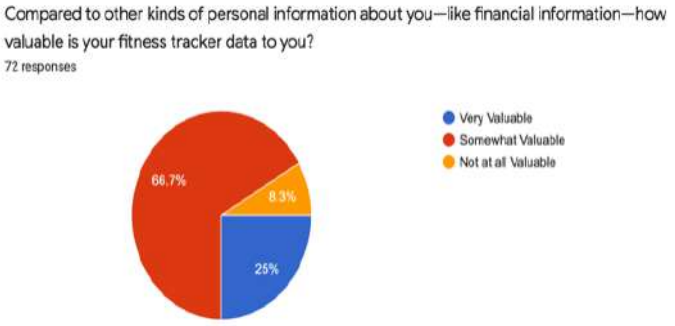


Fig. 9. Count of value of personal fitness information

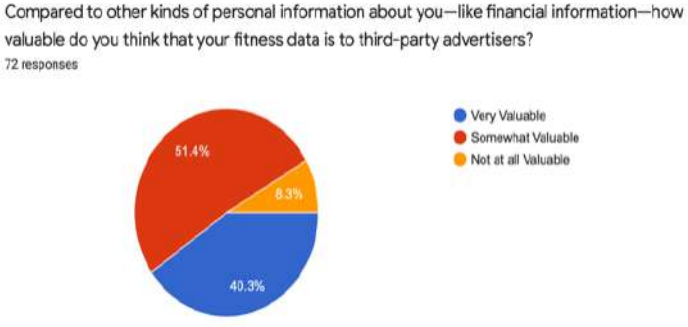


Fig. 10. Result of value of information to third-party advertisers

The above figures 8, 9 and 10 show that a sizable number of individuals are concerned when it comes to privacy of their data whether fitness or other.

ISSN 2394 - 7780

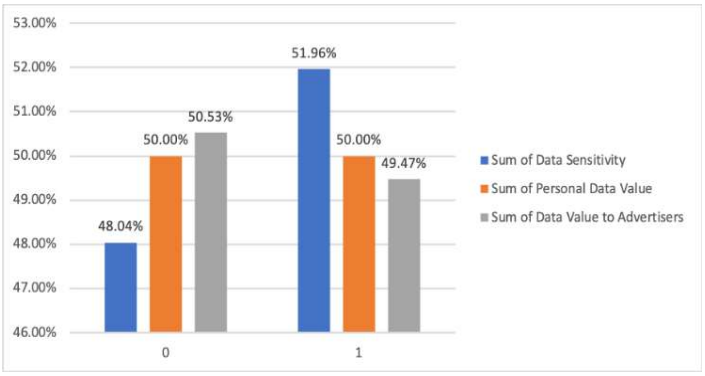


Fig. 11. Interpretation of Data Sensitivity, Personal Data Value and Data Value to Advertisers with respect to gender

The results in figure 11, shows that gender of an individual does not affect the security and privacy risks of wearable devices. ($F(3,68)=0.44$, $p<0.8$, $R^2=0.02$)

Using MLR, because our p-value for the variables above is more than the significance level of 0.05, we can reject the alternative hypothesis and accept the null hypothesis.

INTERPRETATION OF RESULTS

A. Users are unaware of the data being collected: As seen in figure 6, participants are quite uncertain about data their fitness trackers gather, and are unknown to the data usage. Given that the majority of users do not read their tracker's privacy policy or terms and conditions, a bigger question arises: do users truly understand the scope of their data collection? It's impossible to speculate further without knowing more about what they do know.

B. Users take little measures to defend themselves from dangers: Users take few efforts to protect themselves against risks to their privacy and security. This supports the findings of Zimmer, et al. [8], which revealed that many participants had not reviewed their privacy settings since the gadget was first set up. Those who changed their settings likely limited their shared data even more.

Several fitness trackers record personal health data like as daily steps and heart rate, but one needs to bear in mind that all of this information is frequently recorded via applications on ones smart phone, which have access to information such as calendar, contacts list, and ones whereabouts. It's not difficult to figure out where and with whom one spends time using this information. Therefore, one should take some precautions when it comes to the privacy and security of ones data.

1. Thoroughly examine the privacy policies of the fitness tracker: Consider the default privacy policy with caution. In the device's "Settings," go for "Privacy Controls." Look for details on how the information will be accessed by friends or the general public, as well as if it will be sold.
2. Disable "Location Tracking.": Physical location, such as residential address and daily commute, may reveal a lot about a person. Consider enabling location services for only a few applications.
3. When recommended, update the device: Important security patches are frequently included in software updates, which should not ignored.
4. Stay away from insecure networks: As the fitness tracker captures more and more personal information, one should avoid using public Wi-Fi networks to protect that information.
5. Consider the fitness tracker like it is a valued item: It really wouldn't take much for ones identity to be compromised if it were lost or stolen.

SOLUTION

A. Creating wearable technology that can function as a stand-alone device: Wearable technology, such as smartwatches, may be made to work without the use of additional devices, according to this solution. For example, the smartwatches described earlier require connection with a phone through Bluetooth or other protocols and the internet of things. This approach would address the issue of data privacy in wearable devices, such as health records. Because many wearables are connected to other devices, such as mobile phones and computers, data is frequently kept there (due to the lack of storage space in the smart watch). Making the item a stand-alone device would protect the customers data and ensure that their information would not be shared or used without their consent. Because the devices do not need to be connected to a mobile device, this solution can address the user's security concerns.

ISSN 2394 - 7780

B. User transparency: Developers must provide consumers with more and better information about their wearable devices. As wearable devices become smarter, users must understand what information the devices collect about them. Users must also understand how to use devices effectively and in a way that does not compromise or threaten their personal security, because if the security is compromised, it would result in a privacy issue, as data will be released and private information will be made public.

C. Incorporating different policies to secure the user's privacy: Adding numerous policies can aid in the user's privacy solution; nevertheless, the user must review the device's documentation and instructions in order to understand what information is shared; moreover, the user must understand where his data is kept and who has access to it. For example, with GPS tracking, the user must be made aware that their location is always being tracked; organizations should make it a policy to tell users of this information so that their privacy and safety are not compromised.

D. Terms of Service and User Guidelines: Developers will need to include rules for the usage of their goods in order to prevent additional privacy concerns. This will benefit both the user and the developers of wearable devices, because these recommendations will help safeguard both the customer and the developers of wearable devices. Furthermore, following these guidelines will aid in the prevention of any potential legal issues that may arise with the clients in the future.

LIMITATIONS AND CONCLUSION

The sample size may not be representative of the entire fitness tracker users community The participants, who ranged in age from 10 to 39, do not represent older users' views. The survey captures subjective self-reported answers. While a questionnaire allows for the identification of trends in terms of knowledge, adding user interviews might yield a more detailed analysis.

The survey findings have given us a fair idea of what fitness tracker users know regarding security and privacy procedures, as well as their opinions regarding their device's data. Users reported a lack of trust in their understanding of what their fitness tracker gathers and how this information is used.

Overall, the findings have led to reach the conclusion that fitness tracker users need to be more informed of the policies around data collecting, ownership, storing and distribution.

Privacy is more often than not an afterthought as technology advances at a tremendous rate. It is unlikely that one would realize the full importance of privacy unless it has been invaded. One may take certain safety steps if they want to keep their privacy and use fitness trackers as discreetly as possible. It all starts even before purchasing a fitness tracker.

REFERENCES

[1] Alrababah, Z. (2020). Privacy and Security of Wearable Devices.

[2] Vitak, J., Liao, Y., Kumar, P., Zimmer, M., & Kritikos, K. (2018, March). Privacy attitudes and data valuation among fitness tracker users. In International Conference on Information (pp. 229-239). Springer, Cham.

[3] Hilts, A., Parsons, C., & Knockel, J. (2016). Every step you fake: A comparative analysis of fitness tracker privacy and security. Open Effect Report, 76(24), 31-33.

[4] Fereidooni, H., Frassetto, T., Miettinen, M., Sadeghi, A. R., & Conti, M. (2017, July). Fitness trackers: fit for health but unfit for security and privacy. In 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE) (pp. 19-24). IEEE.

[5] Pham, A., Huguenin, K., Bilogrevic, I., Dacosta, I., & Hubaux, J. P. (2015). Securerun: Cheat-proof and private summaries for location-based activities. IEEE Transactions on Mobile Computing, 15(8), 2109-2123.

[6] Mendoza, F. A., Alonso, L., López, A. M., & Patricia Arias Cabarcos, D. D. S. (2018). Assessment of fitness tracker security: a case of study. In Multidisciplinary Digital Publishing Institute Proceedings (Vol. 2, No. 19, p. 1235).

[7] Degeler, A. (2015). Bluetooth low energy: Security issues and how to overcome them.

[8] Zimmer, M., Kumar, P., Vitak, J., Liao, Y., & Chamberlain Kritikos, K. (2020). ‘There’s nothing really they can do with this information’: unpacking how users manage privacy boundaries for personal fitness information. Information, Communication & Society, 23(7), 1020-1037.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

BREAST CANCER CLASSIFICATION WITH HISTOPATHOLOGICAL IMAGES USING MACHINE LEARNING

[1]Dr. Swapnali Lotlikar and [2]Ms. Movina Dhuka
[1]Assistant Professor and [2]MscIT Students, Usha Pravin Gandhi College of Arts, Science and Commerce

ABSTRACT
Breast cancer has impacted approx. 2.3 million people worldwide in 2020, with 685 000 fatalities. Breast cancer had been diagnosed in 7.8 million women in the previous 5 years as of the end of 2020, making it the most common cancer in the world. The death rate can be greatly reduced if the disease is detected and treated at an early stage. Traditional manual diagnosis, on the other hand, necessitates a high workload, and pathologists' long hours are vulnerable to diagnostic errors. Automatic histopathological image identification is critical for speeding up and enhancing the accuracy of diagnoses. Here the proposal is to use machine learning models for prediction of malignant and benign tumours where Convolutional Neural Network (CNN) is used to detect cancerous tumor using histopathological images

Keywords: breast cancer, convolutional neural network, histopathological images

I. INTRODUCTION
Cancer is a term that encompasses a wide range of illnesses. It can appear in practically any part of the body. Carcinomas, sarcomas, and other cancers are examples of distinct types of cancer. Breast cancer is a carcinoma that starts in the skin or the tissue that covers the surface of internal organs and glands. When healthy cells in the breast alter and expand out of control, a tumour forms. A tumour might be malignant or benign. A malignant tumour is one that can grow and spread throughout the body. The term "benign tumour" refers to a growth that does not spread. It affects women of all ages after puberty in every country on the planet, with rates rising as they get older. According to WHO, from the year 1930s to the 1970s, breast cancer mortality remained relatively constant. Survival rates began to rise in the 1980s in countries where early detection programs were combined with various treatment options to eradicate invasive illness [7]. Breast cancer survival rates range from more than 90% in high-income nations to 66 percent in India and 40 percent in South Africa five years after diagnosis. Early identification and treatment has been shown to be effective in high-income nations and should be implemented in countries with limited resources that have access to some of the standard instruments [7]. The World Health Organization's Global Breast Cancer Initiative (GBCI) aims to cut global breast cancer mortality by 2.5 percent each year between 2020 and 2040, avoiding 2.5 million deaths. Health promotion for early detection, quick diagnosis, and comprehensive breast cancer management are the three pillars that will help to achieve these goals. Machine Learning has proven to be a good practice in automating the detection of breast cancer.

In this paper we will be using CNN (Convolutional Neural Network) model along with Transfer learning to achieve the goal of detecting accurate and faster results on breast cancer diagnosis.

II. LITERATUR REVIEW

A. RELATED WORK
The most recent research (Zhou, X., Li, Y., Gururajan, R., Bargshady, G., Tao, X., Venkataraman, R., … Kondalsamy-Chennakesavan, S) uses a 19-layer deep CNN based algorithm. The suggested prediction model's performance was assessed using standard data classification measures such as accuracy, area under the Receiver Operating Characteristic (ROC) Curve (AUC), Classification Mean Absolute Error (MAE), and Mean Squared Error (MSE) [1]. Additionally, they compared their suggested model against a state-of-the-art deep learning model, GoogLeNet, as well as a traditional machine learning classifier, Support Vector Machine (SVM) and CNN as a model. The SVM algorithm is a widely used classification algorithm. It's been used in a variety of cancer prediction systems. Google researchers proposed GoogLeNet, commonly known as Inception v1, in 2014 [8]. The model is made up of "Inception cells," which are basic units. A sequence of convolutions are run at various scales, and the results are then pooled in Inception modules [1].

A similar study done by Singh, Shiksha; Kumar, Rajesh where histopathology-based features have been included in this study for breast cancer detection and categorization. They evaluated K-Nearest Neighbor (KNN), Random Forest, and roughly six varieties of Support Vector Machine (SVM) classification algorithms as part of their research. The experimental results reveal that using a cubic SVM classifier, the suggested strategy for breast cancer detection and classification has a maximum accuracy of 92.3 percent. Classifier

ISSN 2394 - 7780

goodness parameters such as accuracy, precision, recall, f-score, specificity, confusion matrix, and ROC curve are used to verify the results[10]

Another research by Md Zahangir Alom & Chris Yakopcic & Mst. Shamima Nasrin & Tarek M. Taha1 & Vijayan K. Asari in which they used the Inception Recurrent Residual Convolutional Neural Network (IRRCNN) model to present binary and multi-class breast cancer recognition algorithms in this paper.

The trials were carried out using the IRRCNN model on two different benchmark datasets, BreakHis and the 2015 Breast Cancer Classification Challenge, and the results were evaluated using several performance measures. Image-level, patient-level, image-based, and patch-based analyses were used to assess the suggested method's performance [11]

A completely new perspective on the same problem was proposed in a paper written by Benhammou, Yassir; Achchab, Boujemâa; Herrera, Francisco; Tabik, Siham. They presented a taxonomy that divides BreaKHis-based CAD (Computer Aided Systems) systems into four reformulations in this paper (MSB, MIB, MSM, and MIM). Using this taxonomy, they conducted a complete survey of all BreaKHis dataset-using CAD systems. They highlighted their significant contributions, as well as the preprocessing methods they employed, the models they adopted, and the learning strategies they used, as well as the outcomes they achieved at various levels of evaluation. They analyzed these different reformulations to identify the optimal one from both a clinical and practical standpoint, and found that MIM is the best from both perspectives. They then used deep learning CNN to evaluate this MIM technique, which had never been done before in the literature.[9]

II. DATA AND METHODS

A. DATATSET

The dataset being used is from BreakHis (Breast Cancer Histopathological Image Classification) which is made up of 7,909 microscopic images of breast tumour tissue taken from 82 people using various magnifying factors (40X, 100X, 200X, and 400X). There are now 2,480 benign and 5,429 malignant samples in the database (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, PNG format). The P&D Laboratory - Pathological Anatomy and Cytopathology, Parana, Brazil (http://www.prevencaoediagnose.com.br) collaborated on the creation of this database. Breast tissue biopsy slides stained with hematoxylin and eosin are used to create samples (HE). Pathologists from the P&D Lab prepared the tissue for histological analysis and labeled it. Immunohistochemistry analysis of breast tumour specimens (IHC). Core Needle Biopsy (CNB) and Surgical Open Biopsy (SOB) are two types of biopsies. [6].

Table 1 | Distribution of images on the basis of the subclasses and magnification property

Magnification	Benign	Malignant	Total
40X	652	1370	1995
100X	644	1437	2081
200X	623	1390	2013
400X	588	1232	1820
Total	2480	5429	7909

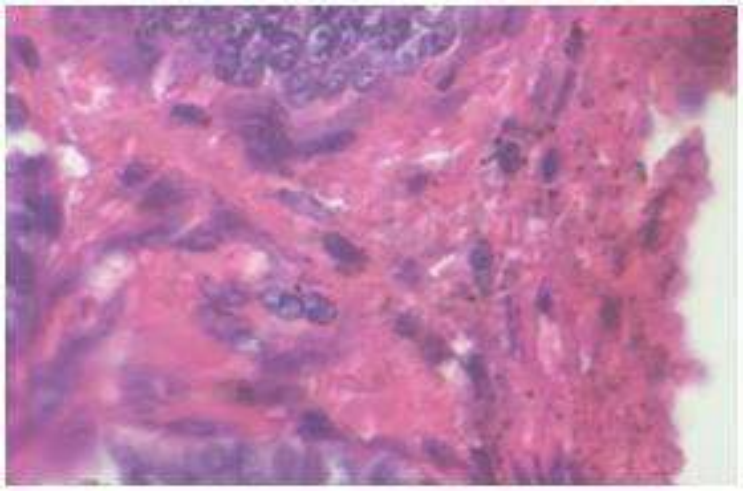


Fig1. Sample image of benign tumour from BreaKHis dataset (400X magnification)[6]

ISSN 2394 - 7780

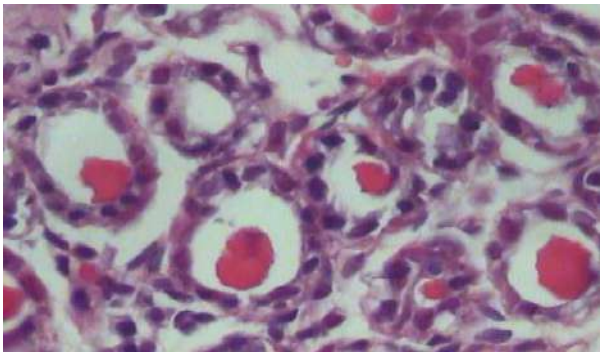


Fig2. Sample image of malignant tumour from BreaKHis dataset (400X magnification)[6]

Breast tumours, both benign and malignant, can be classified into distinct categories depending on how their cells appear under a microscope. Breast tumours come in many distinct forms and subtypes, each with its own prognosis and treatment options. There are currently four histologically distinct types of benign breast tumours in the dataset: adenosis (A), fibroadenoma (F), phyllodes tumour (PT), and tubular adenona (TA); and four malignant tumours (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC).

B. PROPOSED METHOD

Number of studies are been carried out in order to build an efficient model with maximum accuracy. Machine leaning models have proved to be one of the way to help accurate detection of cancer. The use of multiple machine learning and deep learning models on histopathological dataset has proven to be efficient to an extent. This paper will be using CNN model with transfer learning on the histopathological dataset of BreaKHis to predict whether the tumour is malignant or benign.

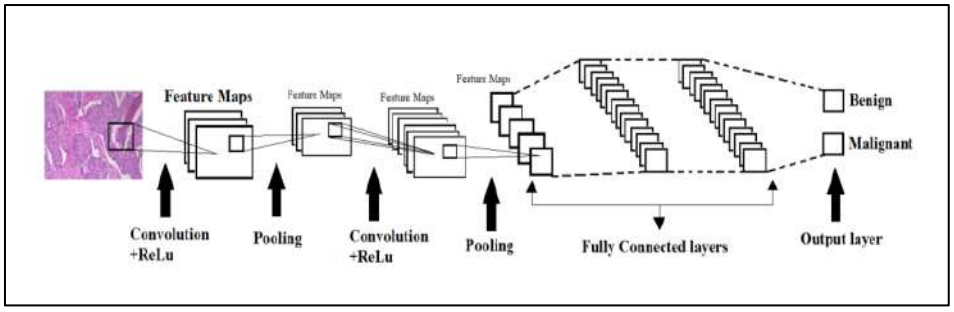


Fig 3.CNN Architecture (Benign or Malignant)[12]

Convolutional Neural Network is a class of Deep Learning used to classify computer vision, images, natural language processing etc. The mathematical function of convolution, which is a special sort of linear operation in which two functions are multiplied to produce a third function that expresses how the shape of one function is modified by the other, is denoted by the term "Convolution" in CNN. In simple words, two matrices are multiplied to provide an output that is used to extract information from a picture.

Cnn Architecture

1. **Input:** The input layer will be taking the pixel values which includes height , width , etc.
2. **Convolution:** In this layer the features maps are created. To create the feature map, the value of each position of the input data was convolved with the kernel in the convolutional layer [5]. For higher level of features more filters are used in CNN whereas for low level feature detect less number of filters are used. A linear output is produced by each of the neurons. When a neuron's output is fed to another neuron, it eventually creates a linear output. Nonlinear activation functions, such as ReLU (Rectified Linear Unit), are employed to solve this problem. Rectified Linear Unit (ReLU) is the most widely used nonlinear operator, as it filters out all negative data [5].
3. **Pooling:** This layer's purpose is to offer spatial variance, which basically implies that the system will be able to recognize an item even if its look changes.
4. **Fully Connected:** We flatten the output of the previous convolution layer and connect every node of the current layer to the other nodes of the next layer in a fully connected layer. As in conventional Neural Networks, neurons in a fully connected layer have full connections to all activations in the previous layer and work in a similar fashion.

ISSN 2394 - 7780

The proposed system consists of the following sub-steps:

- Data Collection – The data being collected is from the BreaKHis dataset which consists of 7909 microscopic images of breast tumour tissues. BreaKHis is one of the many datasets available as open source online for research purpose.
- Data Pre-Processing – This step includes rearranging and processing image in order to structure he dataset along with the 2 predication classes' i.e malignant and benign.
- Data Augmentation– Data augmentation is a useful technique for increasing the size of the training set. Adding to the training examples allows the network to view a wider range of data points while still remaining representative.
- Callbacks– Prior to building a model it is important to set callbacks (Keras) so as to monitor and control the training of the data with respect to the model created
- Building Model– Before building the model we will choose the pre-trained model to implement transfer learning. With all the steps combined the model will trained on the 80% of the dataset and the rest 20% will be used to test.
- Predict Outcomes–Once the model built the prediction of the model will be done by giving different input images.
- Generate Performance Metrics– Performance metrics will be generated after multiple data is being given to the model. Confusion Matrix, ROC curve. To get a better view at misclassifications Precision ,Recall and F1 score can be used.

III. REFERENCES

[1] Zhou, X., Li, Y., Gururajan, R., Bargshady, G., Tao, X., Venkataraman, R., … Kondalsamy-Chennakesavan, S. (2020). A New Deep Convolutional Neural Network Model for Automated Breast Cancer Detection.

[2] Xie, J., Liu, R., Luttrell, J., & Zhang, C. (2019). Deep Learning Based Analysis of Histopathological Images of Breast Cancer. Frontiers in Genetics

[3] Al-Haija, Qasem Abu; Adebanjo, Adeola (2020). - Breast Cancer Diagnosis in Histopathological Images Using ResNet-50 Convolutional Neural Network

[4] Zhu, Chuang; Song, Fangzhou; Wang, Ying; Dong, Huihui; Guo, Yao; Liu, Jun (2019). Breast cancer histopathology image classification through assembling multiple compact CNNs

[5] Nahid, Abdullah-Al; Mehrabi, Mohamad Ali; Kong, Yinan (2018). Histopathological Breast Cancer Image Classification by Deep Neural Network Techniques Guided by Local Clustering.

[6] Spanhol, F., Oliveira, L. S., Petitjean, C. and Heutte, L. Breast Cancer Histopathological Database (BreaKHis)
[Online] Available at: https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

[7] WHO Breast Cancer (2021) [Online] Available at : https://www.who.int/news-room/fact-sheets/detail/breast-cancer

[8] Szegedy, Christian; Wei Liu, ; Yangqing Jia, ; Sermanet, Pierre; Reed, Scott; Anguelov, Dragomir; Erhan, Dumitru; Vanhoucke, Vincent; Rabinovich, Andrew (2015). - Going deeper with convolutions.

[9] Benhammou, Yassir; Achchab, Boujemâa; Herrera, Francisco; Tabik, Siham (2019). BreakHis based Breast Cancer Automatic Diagnosis using Deep Learning: Taxonomy, Survey and Insights. Neurocomputing

[10] Singh, Shiksha; Kumar, Rajesh (2020). - Histopathological Image Analysis for Breast Cancer Detection Using Cubic SVM

[11] Md Zahangir Alom1 & Chris Yakopcic1 & Mst. Shamima Nasrin1 & Tarek M. Taha1 & Vijayan K. Asari (2019). - Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network

[12] Abhinav Sagar (2019) - Convolutional Neural Network for Breast Cancer Classification [Online] Available at: https://towardsdatascience.com/convolutional-neural-network-for-breast-cancer-classification-52f1213dcc9

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

IMPACT OF COVID-19 ON THE MINDSET OF TEENAGERS

[1]Smruti Nanavaty and [2]Ms. Sheetal Y. Kushalkar
[1]MSc-IT Co-coordinator and [2]MSc-IT Student/ Professor, Usha Pravin Gandhi College of Arts, Science & Commerce, Mumbai, Maharashtra, India

ABSTRACT
SARS CoV2 is a deadly emerging pandemic. Studies show that different age groups have been differently affected during the pandemic. It has caused a lot of psychological distress to the children and specially the teenagers across the world. Abundance of chronic stress is experienced by the teenagers because of the confinement to place with little or no access for moving out or any involvement in any physical activity due to Covid-19 restrictions. One more major reason of acute stress among the teenagers is the disturbance in their daily routines which has affected them psychologically. Various concerns have been raised on the impact of mental health of the people because of lockdown.

This paper aims at reviewing at various articles related to mental-health aspects of children and adolescents impacted by COVID-19 pandemic and study the various factors affecting the teenagers. The study further focuses on finding the significantly related factors to mental health issues in children and adolescents during the COVID-19 pandemic. The data here is collected using an online survey. The questionnaire was built with an objective of analyzing the social as well the demographic information. The Research Paper further aims to analyze gender wise mental health issues on teenager's due to covid-19 pandemic.

Keywords: Covid-19, Teenagers, Mental health,Psychological, Lockdown, Gender, Adolescences

INTRODUCTION
The covid-19 pandemic, also called as corona virus pandemic, it is an ongoing global pandemic, which is caused by severe acute respiratory syndrome coronavirus2 which is also termed as SARS-CoV-2. The government has also declared it as a national disaster. The global pandemic (COVID-19) which has been considered as the biggest health crisis which has hit the country so bad since independence.

Covid-19 has impacted bad not only on the lives of people but also the world including children's and teenagers in an unprecedented manner. As we know that essential modus of prevention from the biggest humanitarian infection is been isolation and social distancing as this is the strategies been used to protect our-self form the risk of infection. Teenagers experiences a higher rate of peer interaction and social world interaction as compared to their family, and even from the complex peer relationships as compared to their younger counterparts such as babies and teenager's. Teenagers are also affected by the impact of the pandemic on their Academic structure, studies, unemployment, emotional stress, and fear of infection and the need for adults to receive adequate care and support as well. Due to the restrictions imposed as a measure to control the spread of virus, it has also led to the closure of schools, increased usage of internet among the youths which has badly affected their mental health. In India the lockdown was imposed on 24th March 2020 with strict rules and regulations. We even saw some relaxations in the month of September but, schools were still closed as a matter of safety concern and now they started replacing the physical mode education system with that of the online one.

The aim of this research paper is displaying the impact of the COVID-19 pandemic on teenager and young people's mental health and psychological well-being.

LITERATURE REVIEW
Deepak Nathiya et al[1] represents the different ways of psychological interventions which specifically targets on the youths living in the rural areas and also have made a special report stating irrespective government schemes of educational status only on women's data. In which authors have collected 684 responses from red zone and carried out the performance. Shweta Singh et al[2] proposed a review based research in which they reviewed various articles based on mental-health aspect of adolescent and children who had an impact due to pandemic. Authors have carried out a review-based research which states the advisories affects on mental health of adolescents during the covid-19 pandemic. Sree Latha B Venkat et al in [3] showed that the main objective was to find out how the prevalence of behaviour abnormalities have assessed in school-aged children and adolescents and also the impact of lockdown on children behaviour. Form the given research authors have stated that there was a huge psychological impact of lockdown on adolescents and children's, also the longer screen time increased due to pandemic and the parental conflicts. Mohan Kumar M et al[4] represents a main aim to provide a direct cash transfer and food supplements system as an need which will provide rations and can

ISSN 2394 - 7780

sustain the nutrition of the teenagers. Main aim of the paper is to carry out the flow of uninterrupted amenities which may cause an effect on adolescent in various sectors such as personal hygiene, mental health, education and other such sectors.

Proposed System

The studies included here are classified on some important variable which are discussed below based on the reports found. The data collected is qualitatively analyzed as follow:

Anxiety

Anxiety is normal and often called as a form of healthy emotion. Anxiety disorder is a form of mental health diagnoses that lead us to excessive nervousness, fear, apprehension, and worry. With colleges and school closures and moreover the can-celled events have created a huge impact on teenagers such as they are confined to a place andhave restrictions for moving out and this have adversely affected the mental health and growth of teenagers.

From the survey been collected it is noticed that the anxiety level of females is more that male as shown in diagramgiven below:

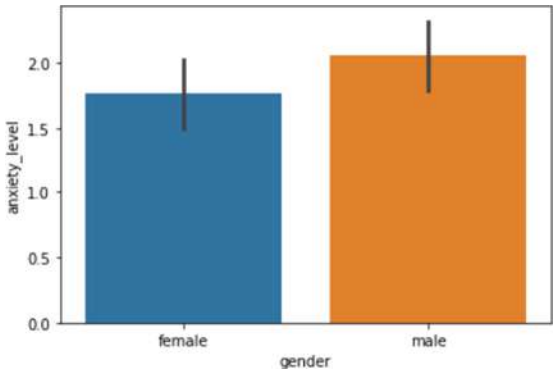


Figure 1: Effects of anxiety-level.

Communication

Communication is one of the most important mode of imparting and exchanging of messages by speaking, writing, or using some other medium. Now a days, due to covid-19 many people have relied on media for communication and has opted it as a effective communication strategy. The sudden changing nature of the pandemic have made communication process easy. Social media plays an important role in communicating during crises and that has been the case during pandemic. It is also seen that effective way of communication during pandemic include contents, methods, people and partners.

The distribution for effective communication by teenagers in COVID scenarios are summarized in image below:

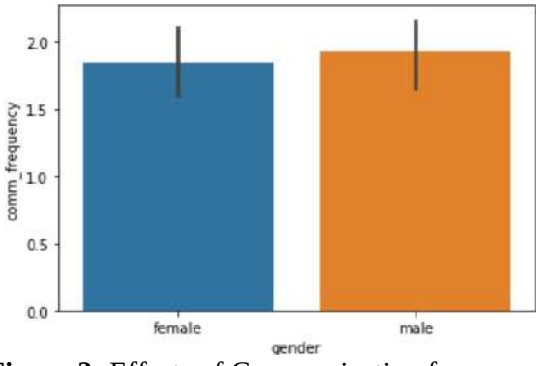


Figure 2: Effects of Communication frequency.

Mental Health

Mental health refers to cognitive, behavioral, and emotional well-being. Mental health can affect daily living, relationship, and physical health. Mental distress during the pandemic have created a negative impact on the mind-set of teenagers, and as a result of the pandemic many repercussions are faced by them such as closures of Schools,colleges, offices, shifting to work from home culture, loss ofincome and rise in the unemployment rate has deteriorated the mental of the people.

ISSN 2394 - 7780

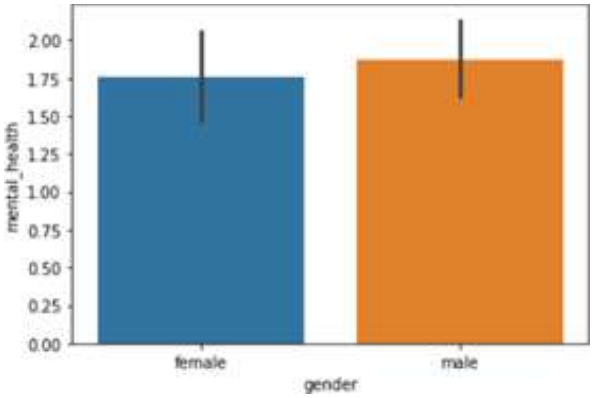


Figure 3: Effects of Mental_health

Education

We all are aware about the covid 19 crisis and how it has impacted the whole world. It has also enthralled every human being to properly follow all the safety norms in the rise of Covid 19 such as washing hands, maintaining social distancing, sanitization, etc. It has adversely affected the Education sector critical determinant of a country's economic future. The pandemic has led to the school and college closures and all the other activities such as exams or events are either being postponed, or they are held online now. Restrictions and confinements have destroyed the routine and schedule of every student. Although, due to the outbreak of covid 19 pandemic the evolution in the Digitalization have also emerged as a positive platform where now people can manage things remotely like now the schools and colleges have shifted their Physical teaching model to online.

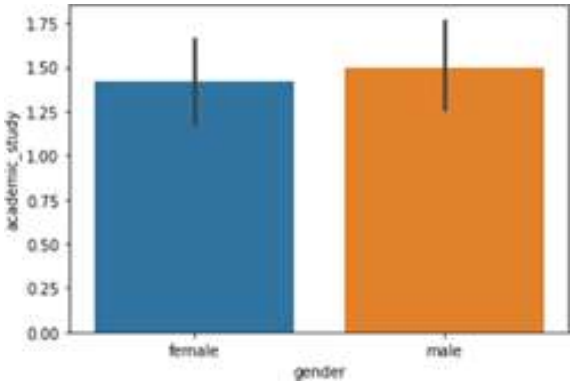


Figure 4: Effects of pandemic on students' academic study.

Sleep Pattern

As we know teenagers and adolescents had an adverse effectdue to covid-19 on various factors. In which Sleep pattern is an important factor which has disturbed a lot. As due to pandemic the sleep pattern has been more prolonged which may lead to several health issues. Teenagers had faced a lot of anxiety and depression especially due to Covid-19 rumination and worries, which has interfered sleep pattern a lot. As it is equally important to understand the risk of psychopathology and the sleep problem.

As shown below in the graphical representation it is measured that the sleep patter of teenagers has adversely affected due to covid-19.

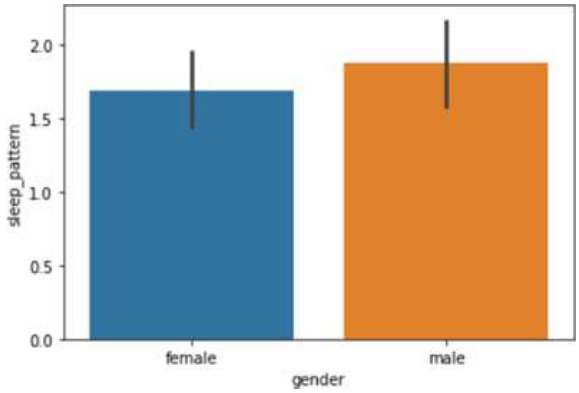


Figure 5: Effects of pandemic on sleep pattern.

ISSN 2394 - 7780

Social Isolation

Social isolation is often described as loneliness which can also be define as a state of incompleteness or near-complete lack of contact. It also deals with been in a isolated mode from the community and society which may give rise tomany mental disorders such as insomnia, depression,adjustment and many more.

Loneliness is common among the older age group, leading to increased rates of depression and suicide. It has been well documented that long periods of institutional isolation or disease quarantine have negative effects on mental well- being. (Wilson et al., 2007)

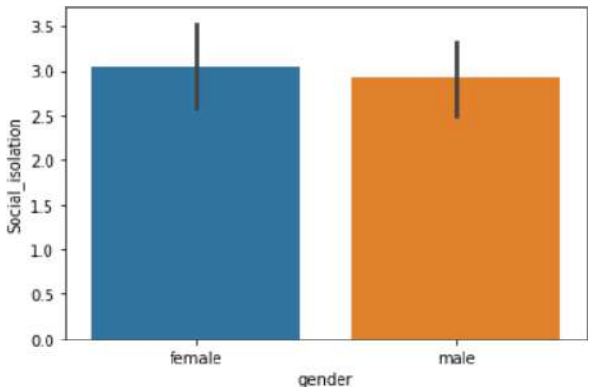


Figure 6: Effects of Social-isolation.

As we know, that the use of social media is increased due to the pandemic, As social media not only has the benefits but also the consequences. Social media does gives us various vital equipment for teens to get the proper access to all the related information and resources related to covid-19. It also helps teenagers to do various activities such as creative expression, identity exploration and social connection. Adolescence is a crucial face of development stage in which all the activities are equally important. Also due to pandemic the use of social media has been increased a lot and which may cause limited growth in many areas such as education, physical activities etc. Given below are some keyconsiderations points for parents to be taken for their children for using phone during the Covid-19 pandemic.

Table No1: Key considerations for parents on teens social media use during the COVID-19 pandemic.

Domain	Key Consideration
Peers and Socialization	•What platforms are youth using to interact with their peers? • How much time do young people spend passively scrolling rather than interacting directly? • What features are available in the social media tools they use?
Physical Health	• How much time are teenspending on social media? • At what times of day are teens spending time on social media?
Self-Esteem, Body Image and Mood	• What social media activities do teens find make them feel good, and which do not? • Seek professional help with serious teenager mental health problems.
Resources and Information	• Where are teens reading or learning information about COVID-19, and are these sources accurate? • Are teens accessing helpful resources for managing emotions related to COVID-19? • Direct teen to trusted information source and resources for managing stress.
Overall	• How can parents best support teens in the COVID-19 pandemic, particularly in regard to social media use? • Be a role model with how, and when to use social media. • Be forgiving of yourself an your teen during this time. • Take advantage of social media's together with teens. • Check out tip for parents www.commonsensemedia.org

ISSN 2394 - 7780

UNITS

DATA AND METHODS

Data Collection

Data of the respondents were collected with the help of the social media. After obtaining the data from the given age criteria, which was filled and completed by the teenagers on June 12. Google forms were used to collect pre-tested structured questionnaire containing basic information of the teenagers and the question was related of how covid 19 has impacted on their daily life's activities. Teenagers between 19 and 27 years (n=75) living with their parents under lockdown was the main age criteria.

Data Analysis

The data collected was analyzed using different data analysis tools with the help of python, numerical outcome was predicted and also using different types of graphical methods such as barplot, histogram and boxplot.

RESULTS AND DISCUSSION

The Teenagers were electronicallyapproached to fill the assessment. This data here was collected from different people, asked them different category of questions related to how covid-19 has impacted their mental health.

After collecting the data, we have performed some analysis to know how does covid-19 has affected their mindset and also carried out some operations on that data. There were total 80 participants who successfully completed the Assessment, which include 38 (50.7%) males and 37(47.3%) females. The age group which we have specified for our research usually ranged from 19 to 30 year. And from the whole data which we collected 90% of our data belongs to the age group of 20-25 years. Further in the analysis both male and females who answeredto the question asked in the survey has been analyzed and the outcome is been carried out that Females are more plagued as compared to male in the symptoms of random thoughts and over thinking.

As shown in the below table and graphical representation:

Table 2: Responses collected from the data towards PsychologicalImpact of Covid-19 w.r.t. gender.

Sr No	Variables	Male	Female	Mean	Var	SD
1	Anxiety_level	78%	67%	1.907	0.778	0.6324
2	Study_experience	82%	89%	2.4078	0.6908	0.8355
3	Risk_of_Contagion	100%	100%	2.8684	2.835	1.683
4	Mental_health	71%	67%	1.8157	0.7122	1.8157
5	Stress	92%	99%	2.552	0.9171	0.9577

As shown below is a visual representation of the above data in the form of graphical view and the distribution is shown through using this variable with respect to genders:

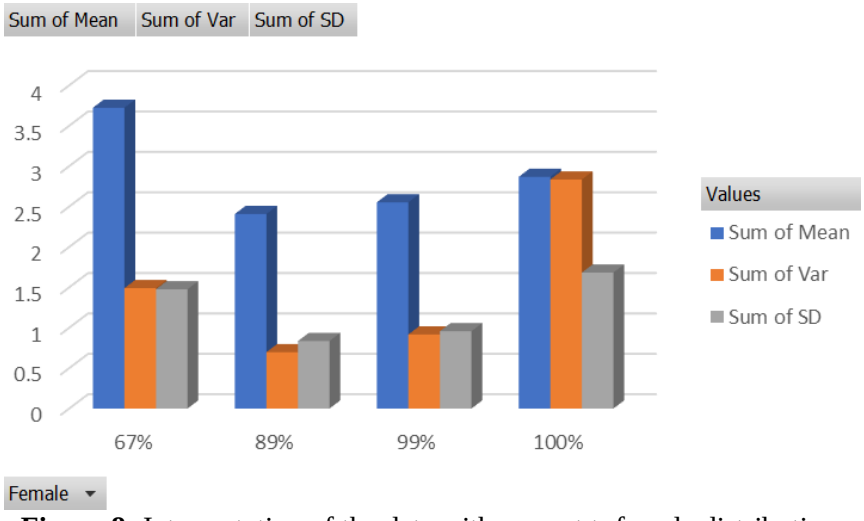


Figure 8: Interpretation of the data with respect tofemale distribution.

ISSN 2394 - 7780

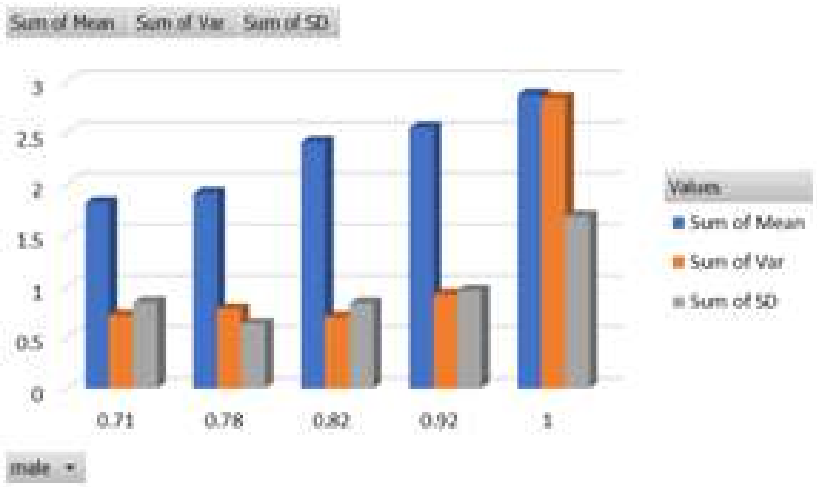


Figure 9: Interpretation of the data with respect to male distribution.

CONCLUSION

In Summary, we have demonstrated that how much covid-19 has impact on teenagers' mental health we have also demonstrated the assessment of reaction to perceptions approximately the effects of covid-19 at the participants. A growth has been measured in the symptoms of stress, anxiety_level, mental_health, study_experience, risk_of_contagion has been measured through the data which is been collected and showed above, also we can conclude by looking at the outcome that Females are more plagued as compared to male in the symptoms of random thoughts and over thinking.

REFERENCES

[1] Suman, Supriya and Singh, Pratima and Nathiya, Deepak and Raj, Preeti and Tomar, Balvir. (2020). Mental health problems and impact on the minds of young people during the COVID-19 outbreak: cross-sectional survey (RED-COVID).

[2] Venkat SL, Ananthakrishna K R. Psychosocial impact of COVID-19-related confinement in school-aged children and adolescents in Karaikal - A longitudinal study. India J Soc Psychiatry 2021; 37: 77-8.

[3] Cortese, S., Asherson, P., Sonuga-Barke, E., Banaschewski, T., Brandeis, D., Buitelaar, J., Coghill, D., Daley, D., Danckaerts, M., Dittmann, RW, Doepfner, M., Ferrin, M., Hollis, C., Holtmann, M., Konofal, E., Lecendreux, M., Santosh, P., Rothenberger, A., Soutullo, C.,. .. Simonoff, E., 2020. ADHD management during the COVID-19 pandemic: advice from the European ADHD guideline group. Lancet Child Adolesc. Health. https://doi.org/10.1016/S2352-4642(20)30110-3.

[4] National Academies of Sciences, Engineering and Medicine. Quick Expert Consultations on the COVID-19 Pandemic: March 14, 2020 - April 8, 2020. Washington DC:National Academies Press; 2020.

[5] Qiu J, Shen B, Zhao M, Wang Z, Xie B, Xu Y. A national investigation into Chinese psychological distress in the COVID-19 epidemic: implications and policy recommendations. Gen Psychiatr 2020; 33: e100213.

[6] Nishiura H, Kobayashi T, Miyama T, Suzuki A, Jung SM, Hayashi K, et al. Estimation of the asymptomatic ratio of new coronavirus (COVID-19) infections. Int J Infect Dis 2020; 94: 154-5.

[7] Zhao X, Zhang B, Li P, Ma C, Gu J, Hou P, et al. Incidence, clinical features, and prognostic factor of patients with COVID-19: a systematic review and meta-analysis. MedRxiv 2020.

[8] Coronavirus Disease 2019 (COVID-19) WHO Situation Report 94. Available from: https://www.who.int/docs/defaultsource/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.

[9]Anderson, M. and Jiang, J. (2018). Teens, social media and technology. Becker, S.P., Sidol, CA, Van Dyk, TR, Epstein, J.N. and Beebe, D.W. (2017). Intra-individual variability of wake / sleep patterns in relation to child and adolescent functioning: a systematic review. Opinion on leephttps://doi.org/10.1016/j.smrv.2016.07.004

[10]Coronavirus Disease 2019 (COVID-19) WHO SituationReport 94. Available at: https://www.who.int/docs/defaultsource/coronaviruse/situation-reports/20200423-sitrep-94-covid-19.

[11] Anderson, M. and Jiang, J. (2018). Teens, social media and technology. Becker, S.P., Sidol, CA, Van Dyk, TR, Epstein, J.N. and Beebe, D.W. (2017). Intra-individual variability of wake / sleep patterns in relation to

ISSN 2394 - 7780

child and adolescent functioning: a systematic review. Opinion on sleep medicine, 34.94 121https://doi.org/10.1016/j.smrv.2016.07.004

[12] Brechwald, W. A. and Prinstein, M. J. (2011). Beyond Homophilia: A Decade of Advances in Understanding Peer Influence Processes. Journal of Research on Adolescence, 21, 166-179 https://doi.org/10.1111/j.1532-7795.2010.00721.

[13] Calhoun, AJ, & Gold, J.A. (2020). "I seem to know them": the positive effect of celebrity exposure to mental illness. Academic Psychiatry, 44, 237-241 https://doi.org/10.1007/s40596-020-01200-5

[14] Centers for Disease Control and Prevention. (2020). Coronavirus disease 2019 (COVID-19): social distancing,quarantine and isolation. https://www.cdc.gov/coronavirus/2019-ncov/prevent- getting-sick/social-distancing.html.

[15] Choukas-Bradley, S., & Thoma, B.C. (in press). Mental health in LGBT youth. In W. I. Wong & D. VanderLaan (eds), Gender and Sexuality Development: Contemporary Theory and Research. New York: Springer.

[16] Cohen, R., Newton-John, T., & Slater, A. (2017). The relationship between appearance-oriented activities on Facebook and Instagram and body image issues in young women. BodyImage, 23,183,187https: //doi.org/10.1016/j.bodyim.2017.10.002.

[17] Arafat, S.Y., Kar, S.K., Marthoenis, M., Sharma, P., Apu, E.H., Kabir, R., 2020. Psychological fundamentals of panic buying during the pandemic (COVID-19). Research in psychiatry.PMC,113061.https://doi.org/10.1016/j.psychres.2020. 113061.

[18] Bhat, R., Singh, V.K., Naik, N., Kamath, C., R, Mulimani, P, Kulkarni, N., 2020. COVID 2019 outbreak: the disappointment of Indian teachers. Asian J. Psychiatry,102047.https://doi.org/10.1016/j.ajp.2020.102047

[19] Dalton, L., Rapa, E., Stein, A., 2020. Protect children's psychological health through effective communication onCOVID-19. Lancet Child Adolesc.

[20] Health4 (5), 346-347. https://doi.org/10.1016/S2352-4642(20)30097-3. Protecting and Mobilizing Youth in COVID-19 Responses. Economic Analysis & Policy Division, Dept of Economic & Social Affairs UnitedNations; 2020.

[21] Rieger MO. To wear or not to wear? Factors influencing wearing face masks in Germany during the COVID-19 pandemic. Soc Health Behav 2020;3:50-4.

Questioner's used to carried out this research are:

1.Is there a change in the frequency of communication with your friends and family in pre and post pandemic situations.
2. Has covid impacted your anxiety levels?
3. Has pandemic affected your social life?
4. Did you experience any interruption or disruption in your sleep patterns due to outbreak concerns ?
5. Do you think your daily routine and working hours had a huge Impact due to covid-19 ?
6. Do you think your life was different before the pandemic?
7. How often have you started using your phone as compared to pre-pandemic situations?
8. Do you think covid has affected your mental health?
9. How do you perceive your academic studying experience during this period of Covid-19 pandemic?
10. Has covid affected you even after following the safety norms properly?
11. Did this pandemic impact your academic life?
12. Do you think that due to covid there is an evolution in the online and digital platforms?
13. Do you think it is easy for you to adopt the new online lifestyle due to Covid 19?14. Do you think that there is a huge increase in the usage of your mobile phones due to the pandemic?
15. Have you experienced any stress related to the pandemic?
16. Do you intend on learning or acquiring any skills during lockdown?
17. How do you perceive the condition of social isolation imposed during this period of the COVID-19 pandemic?
18. Do you think that pre-pandemic academic structure(physical mode) was better than the post-pandemic(online mode)?
19. How do you perceive the risk of contagion(covid) during this period of the pandemic ?

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

FINANCIAL DISTRESS PREDICTION USING MACHINE LEARNING

Manisha Divate and Siddhesh Gupta
Usha Pravin Gandhi College of Arts, Commerce and Science, Vile parle, Mumbai, Maharashtra, India

ABSTRACT
Financial distress occurs when a company or individual is unable to generate adequate revenue or money to fulfil or payback its financial commitments. This research looks at how Machine Learning can be used to identify personal financial distress. Financial frauds are a rising problem in the financial services industry with far-reaching implications, while numerous techniques have been developed, Machine Learning has been used to automate the processing of large volumes of complicated data in finance systems. In the identification of distress, Artificial Intelligence has played a significant role in the financial industry. Predicting different frauds or patterns is a big data challenge that is made more difficult by two factors: first, the profiles of normal and fraudulent behaviour vary regularly, and second, cybercrime data sets are highly skewed. This research explores and compares the performance of Different Machine Learning Models on publicly available dataset. Dataset of 15,000 individuals is sourced from public repository by Lending.com. The Algorithms are implemented on the raw and pre-processed data and the outcome of these Algorithms/Models is evaluated based on accuracy, sensitivity, specificity and precision.

INTRODUCTION
This research tries to predict if an individual can face financial distress over the period of next 2 years. This information can be very important to financial institutions which will cater its services to such individuals. If this research can predict the financial distress of an Individual, then that data can be used by financial institutions to limit or deny services to any individual who can face financial distress in near future. Financial fraud is a rising problem in the government, business organisations, and the financial industry, with far-reaching consequences. The heavy reliance on internet technologies in today's environment has accelerated financial transactions. As online transactions have become a more common way of payments, emerging computational approaches for dealing with financial services difficulties have gotten a lot of attention. Many credit scoring systems and tools are available to help organisations such as credit card industry, retail sector, e-commerce services, insurance, and other industries to avoid fraud. It is difficult to be absolutely confident of an application's or transaction's real intention and legality. The most effective method is to use mathematical algorithms to search for probable fraud evidence in the existing data. The procedure of identifying those individuals that are suspected is converted into two classes of real class and distress class, various algorithms and models are developed and deployed to solve such tasks as deep neural networks, frequent item set mining, machine learning models, migrating bird’s optimization algorithm, logistic regression, Support Vector Machines, decision tree and random forest.

These problems are quite prevalent in the financial world, yet they are also hard to resolve. First, it's difficult to match a pattern for data set because of the fact that there is just a little amount of data. Second, several data collection items with separate truncations may likewise fit within acceptable conduct patterns. There are also several limitations to the problem. First of all, data sets are not easily accessible to the Public, and the outcomes of study are often obscured and monetized, making the results unavailable. Datasets with actual published studies are not mentioned in previous research. Furthermore, it is more difficult to develop techniques by limiting the interchange of ideas and methodologies in these studies as a result of the security issue. Finally, the data sets are always changing, which makes it possible to distinguish the profiles of ordinary and malicious behaviours that the legal transaction in the past has been or is still a fraud. This study examines the four approaches of machine learning, decision tree, vector support, logistic regression and random forests, followed by a joint comparison to assess which model was best performed.

LITERATURE SURVEY
In [1] this paper represents a case study involving the prediction of fraud, which shows that before modelling, data standardisation is used and with the results obtained from the use of unattended learning networks and deep neural fraud detection networks that clustering characteristics can minimise neural input. The use of standardised data with already trained data can also provide intriguing outcomes. In this study, unsupervised learning was applied and new techniques for fraud prediction were designed and the findings accurately improved.

In [2] A new comparative measure was created in this study, which effectively aggregates evaluation metrics. A cost-sensitive method based on Bayes is presented with the proposed cost measurement. When comparing this

ISSN 2394 - 7780

approach with other state-of-the-art algorithms, up to 25% gains are achieved. The data collected for this research was based on transactional real-life data of a major Global firm, and personal data was kept confidential. The accuracy of an algorithm is about 60%. This effort was aimed at developing an algorithm and reducing costs. The result was a 26% increase, with Bayes' least risk approach.

In [3] To identify fraudulent transactions, several current approaches based on pattern recognition, deep neural networks, machine learning, artificial intelligence and others have been developed and are continuously being developed. All these techniques require a thorough and clear knowledge, which will undoubtedly lead to an effective system. This paper comprises an examination of several techniques and an evaluation of each methodology on the basis of certain performance standards. The survey in this paper aimed to assess each technique's efficiency and sensitivity. The relevance of this study is carrying out a study to assess multiple algorithms so that the best way to solve the problem is determined.

In [4] In this study, a comparison of artificial intelligence models is done, as well as a comprehensive explanation of the created fraud detection system, such as the Naive Bayes Classifier and the model on Bayesian Networks, the deep neural network model. Finally, judgments regarding the outcomes of the models' evaluative testing are reached. Using the Bayesian Network, it was found that the number of lawful truncations was higher or equal to 0.68, indicating that their accuracy was 68 percent. The purpose of this work is to compare artificial intelligence models, along with a general parametrization of the produced system, and to indicate the specificity of each model, as well as recommendations for improving the model.

In [5] Nutan and Suman supported the theory of what is fraud, types of fraud such as telecommunications, fake bankruptcy, and how to detect it in their review on fraud detection. They also explained numerous algorithms and methods for detecting fraud, including the Glass's Algorithm, Bayesian networks, Hidden Markov model, Decision Tree, and others. They offer detailed explanations of how the algorithms operate as well as mathematical explanations. The goal of this study is to identify fraud in a dataset collected from ULB website by utilising Logistic regression, Decision trees, and other models to evaluate their accuracy, sensitivity, specificity, and precision, and compare them to the best feasible model to address the fraud detection problem.

BACKGROUND

The ability of a system to learn and improve without explicit programming is machine learning. It includes the development of computer systems that can use data for their own learning. That a classifier algorithm may be described as an algorithm for classification, especially when implemented, as well as a mathematical function that is implemented in categories by an algorithm and maps input data. It is a supervised learning example, which provides a training set of correctly accepted observations.

Logistic Regression: Logistic regression is a supervised classification technique predicting the likelihood of a binary variable depending on the independent variable in the data set. The probabilities of a result with two values, zero or one, yes or no, false or true, are predicted using logistic regression. Linear regression is like a straight-line regression, but logistic regression generates a sigmoid curve. Based on a predictor or an outcome variable, the logistic regression produces sigmoid curves which represent zero to one value based on logarithmic functions. Regression is a model with a category dependent variable which analyses the link between several independent variables. The logistic regression models, including binary, multiple and binomial logistic models, are many variants. The Binary Logistic Regression Model calculates the probability of a binary response depending on one or more variables.

SVM (Support Vector Machine): SVM is a method of machine learning regression and classification. It is a supervised form of learning that captures information. Modelling SVM involves two steps: training a data set to build a model, and then predicting information from a test data set by using that model. The SVM model depicts the training data points as points in the n-dynamically spatial range, then maps them in a way that separates the points of various classes from the broadest range possible. In the SVM technique, each data item is treated as a point for n-dimensional space, where n is the number of characteristics and the value of each feature is the value of a specific co-ordinate. The classification is then performed by locating the hyperplanes which separate the two groups clearly.

Decision Tree: Decision tree is an algorithm that provides a tree-like graph or model of decisions and their likely consequences for probabilistic choosing. This method uses conditional assertions of control. It is an algorithm for an objective function, which represents an alpha function in the decision tree. These algorithms are famous for inductive learning and have been used to various applications efficiently. It assigns a label to a new block, indicating whether the class label is valid or false, then test the transaction value against the decision tree, and then trace the journey from the root node to that item's output/class label. Decision rules determine the

ISSN 2394 - 7780

results of the contents of the leaf node. In principle, rules are 'If condition 1 and condition 2 are true, but condition 3 is wrong, the result is false'. This decision tree makes it easy to understand and analysis and enables the insertion of additional scenarios, making it easier to establish the worst, best and expected values for diverse situations.

Random Forest: Random Forest is a technique of regression and classification. It is a group of decision tree classifiers. A fraction of the training sample is sampled altered so that each node splits all the exercises in a single tree and then one decision tree by random subset. Also, it is remarkably quick even for large-scale sets of training and data in random forests with every tree being trained independently. This technique provides a good evaluation of the generalisation error and resists overfitting. The relevance of variables may be determined naturally in the Random Forest using a random forest in a regression or classification task.

METHODOLOGY

The initial data is derived from the data source and the validation is done on the data set, where the redundancy is removed, empty spaces are filled into columns and the required variable is converted in factors or classes. The K-fold crosses are now validated and randomly separated into k sub-samples of the same size. The validation of the model is retained as a subsample, while the rest of the k sub-samples are used as training data. Logistic regression, decision-making, SVM and random forest models will be developed, and precision will be tested, and a comparison will be made. Sensitivity will then be evaluated.

The data set comes from a public repository that is updated for peer-to-peer financial services by Lending.com. The dataset contains information of 15000 individuals with their financial history. The data set is severely imbalanced, and 0.18 percent of the data is skewed to the negative class. It comprises solely numerical (continuous) input variables that are transformed into 12 main components according to the Principal Component Analysis (PCA). And in this study a total of 8 input functions are used. A variable in each profile usage, indicating customer financial situation combined with days of month, hours of days of day, geographical sites or type of the merchant in whom the transaction takes place is the typical behavioural of the individual. Confidentiality problems cannot provide the specifics and context of characteristics. The time function saves the seconds between each transaction and the first transaction in the dataset. The transaction value is the 'amount' feature. Feature 'class' is the binary class target class and takes value 1 in positive (failure) and value 0 in negative (fail) cases (non-fraud).

Four classification models were trained in this study based on logistic regression, SVM, decision-trees and Random Forest. 80% of the data set is utilised for training to assess these models, whilst 20% are used for validation and testing. The performance of the four classifiers is assessed using accuracy, sensitivity, specificity, precision. In every set of a sample the true positive, true negative, false positive and false negative rates are represented in the table below and a confusion matrix format is also shown. The precision and specificity ratings of several true negatives are inaccurately high in the table.

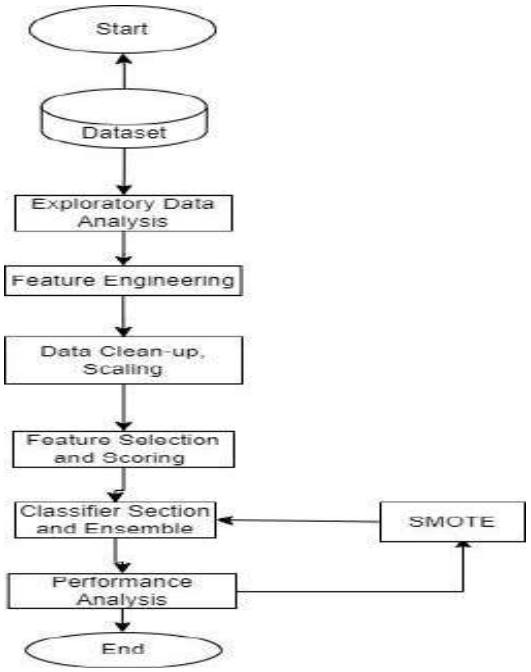


Figure 1: Architecture

ISSN 2394 - 7780

RESULTS

From the studies, it has come to the knowledge that the logistic model is 97.7 percent accurate, while the SVM is 97.5 percent accurate as well as the decision tree is 95.5 percent accurate, however, the Random Forest with highest outcomes have achieved. 98.6 percent precision. Interpreting form different model performance metrics, it comes to light that model was overfitting the training data because of bias inherited form the dataset. After SMOTE was applied, Model performance was seen to be improved. Random Forest was seen to be the best performer on the dataset.

Table 1: Performance Metrics

Metrics	Logistic Regression	SVM	Decision Tree	Random Forest
Accuracy	0.977	0.975	0.955	0.986
Sensitivity	0.965	0.973	0.949	0.991
Specificity	0.923	0.912	0.893	0.982
Precision	0.996	0.995	0.963	0.994

CONCLUSION

Though there are many identity verification methods available today none is able to identify all frauds entirely while they are actually occurring, they generally detect it until the fraud has been perpetrated. This happens because a very minuscule number of transactions from the total transactions are actually fraudulent in nature. With more learning information, the Random Forest Algorithm will do faster, but velocity will be impaired in experimentation and implementation. It would also assist to implement more pre-processing methods. The support vector machine software already comes from unbalanced data sets issue and needs a higher preliminary processing rate to achieve superior outcomes at the outcomes as seen by Support vector machine. The requisite to develop a successful hybrid system is to combine costly training techniques with incredibly precise and exact outcomes with an enhancement method to reduce system costs and rapidly train the machine. The selection of hybrid methods depends on how the fraud sensing device works and the workplace

REFERENCES

[1] Raj S.B.E., Portia A.A., Analysis on credit card fraud detection methods, Computer, Communication and Electrical Technology International Conference on (ICCCET) (2011), 152-156.

[2] Jain R., Gour B., Dubey S., A hybrid approach for credit card fraud detection using rough set and decision tree technique, International Journal of Computer Applications 139(10) (2016).

[3] Dermala N., Agrawal A.N., Credit card fraud detection using SVM and Reduction of false alarms, International Journal of Innovations in Engineering and Technology (IJIET) 7(2) (2016).

[4] Phua C., Lee V., Smith, Gayler K.R., A comprehensive survey of data mining-based fraud detection research. arXiv preprint arXiv:1009.6119 (2010).

[5] Bahnsen A.C., Stojanovic A., Aouada D., Ottersten B., Cost sensitive credit card fraud detection using Bayes minimum risk. 12th International Conference on Machine Learning and Applications (ICMLA) (2013), 333-338.

[6] Carneiro E.M., Dias L.A.V., Da Cunha A.M., Mialaret L.F.S., Cluster analysis and artificial neural networks: A case study in credit card fraud detection, 12th International Conference on Information Technology-New Generations (2015), 122-126.

[7] Hafiz K.T., Aghili S., Zavarsky P., The use of predictive analytics technology to detect credit card fraud in Canada, 11th Iberian Conference on Information Systems and Technologies (CISTI) (2016), 1-6.

[8] Sonepat H.C.E., Bansal M., Survey Paper on Credit Card Fraud Detection, International Journal of Advanced Research in Computer Engineering & Technology 3(3) (2014).

[9] Varre Perantalu K., Bhargav Kiran, Credit card Fraud Detection using Predictive Modeling (2014).

[10] Stolfo S., Fan D.W., Lee W., Prodromidis A., Chan P., Credit card fraud detection using meta-learning: Issues and initial results, AAAI-97 Workshop on Fraud Detection and Risk Management (1997).

[11] Maes S., Tuyls K., Vanschoenwinkel B., Manderick, B., Credit card fraud detection using Bayesian and neural networks, Proceedings of the 1st international naiso congress on neuro fuzzy technologies (2002), 261-270.

ISSN 2394 - 7780

[12] Chan P.K., Stolfo S.J., Toward Scalable Learning with NonUniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection, In KDD (1998), 164-168.

[13] Rousseeuw P.J., Leroy A.M., Robust regression and outlier detection, John wiley & sons (2005).

[14] Wang C.W., Robust automated tumour segmentation on histological and immunohisto chemical tissue images, PloS one 6(2) (2011).

[15] Sait S.Y., Kumar M.S., Murthy H.A. User traffic classification for proxy-server based internet access control,IEEE 6th International Conference on Signal Processing and Communication Systems (ICSPCS) (2012), 1-9.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

HIV/AIDS INFECTION PREDICTION USING MACHINE LEARNING

Manisha Divate and Govind Thakur
Usha Pravin Gandhi College of Arts, Commerce and Science, Vile Parle, Mumbai, Maharashtra, India

ABSTRACT
HIV disease transmission very fast in human body .the focus of this paper improve the artificial intelligence using machine learning method or algorithm in health sector to learn the pattern of virus to predict in human being .what is the effect are causing on our health. HIV/AIDS Treatment .In the first

Thing we analysis the genetic diversity using classification of machine learning labelled data supervised learning. KNN (k-neareast neighbour algorithm),Naïve based classifier. The result is matching with its each strain check its infection . HIV/AIDS infection patients blood sample data are available. To check its virus strain to detect its his/her infected lineage.And also trying to find its recovery rate who infected with this virus based on their past immune recovered from critical stage.

Keywords: HIV Prevention, HIV/AIDS Infection, Virus,KNN,SVM, AI Technique, Machine learning.

I. INTRODUCTION
Human immunodeficiency infection/AIDS (HIV/AIDS) was begun from monkeys in the United States in 1981. Helps is a persistent and possibly most compromising irresistible sickness caused by human immunodeficiency infection in the 21st century. 78 million individuals were assessed to be experiencing HIV/AIDS and 35 million individuals have kicked the bucket since the beginning of the pandemic year yet 36.7 million individuals were revealed as HIV/AIDS tainted and 1.1 million individuals have kicked the bucket in 2015 all around the world; 2.1 million individuals were viewed as recently HIV/AIDS contaminated all around the world.

Eastern and Southern Africa have a greatest increment since practically the beginning of the pandemic year. HIV/ Helps routinely wrecked the number of inhabitants in Africa displayed in fi gure1 (HIV/AIDS 2016). The current commonness of HIV/AIDS is 0.8% the around the world. 18.2 million individuals were getting to antiretroviral treatment in June 2016 (Alkema et al. 2016).

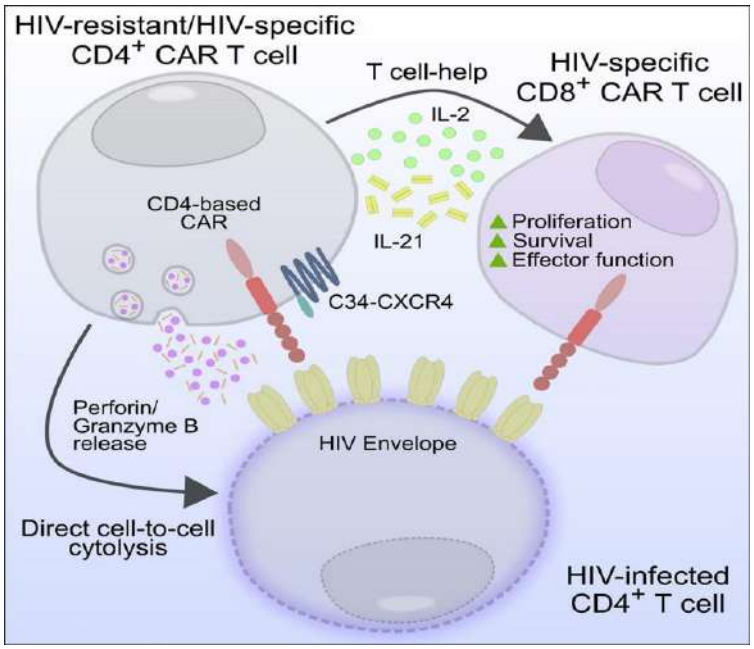


Fig-1 HIV/AIDS CD4 Cells

STAGES OF HIV/AIDS:
There are basically three periods of HIV: Primary infection (Intense HIV), Clinical latent sickness (Chronic HIV) and Early demonstrative HIV illness. In fundamental illness, an influenza like affliction made inside a couple of months after the fact the disease entering the grouping of people, counting signs and appearances like fever, headache, muscle harms, joint misery, rash, sore throat and broadened lymph organs essentially on the neck. In any case, the signs of the primary period of HIV tainting are for the most part concealed, how much contamination in the viral weight or dissemination framework spreads significantly at the present time, achieving dispersing the HIV defilement more productively during fundamental infection than the

ISSN 2394 - 7780

accompanying stage. In the clinical dormant, not really set in stone extending of lymph center points occurs and HIV stays in the human body yet with next to no signs and after effects.

COMPLICATIONS OF HIV/AIDS

The heaviness of HIV is for the most part consequence of pollutions (tuberculosis (TB), cytomegalovirus, candidiasis, cryptococcal meningitis, toxoplasmosis and cryptosporidiosis), cancers (Kaposi's sarcoma and lymphomas) likewise others (wasting condition, neurological snares and kidney dis How ease) (How). The extending in the frequencies of the HIV/AIDS and their outcomes to the extent tumbling down the amount of CD4 (Cluster of division 4) receptors and coreceptors lead to hurt the safe game plan of the human.

Infections common to HIV/AIDS

Tuberculosis (TB): Tuberculosis coinfection is related with increment viral replications.

Cytomegalovirus: This herpes virus is transmitted in human body fluids such as saliva, urine, blood, breast milk and semen.

Lymphomas: Lymphomas cancer originates in white blood cells and usually appears in lymph nodes. Painless swelling of the lymph nodes in neck, armpit or groin is most common early sign.

Kaposi's sarcoma: Kaposi's sarcoma is a blood vessel walls, a very rare tumor in HIV-negative people but a very common in HIV-positive people. It usually appears as pink, purple or red lesions on the skin and mouth of the HIV infected people.

Other complications:

Squandering condition: Aggressive treatment regimens have diminished the quantity of instances of squandering condition, yet it actually influences numerous HIV/AIDS tainted individuals. It

Kidney sickness: HIV-related nephropathy (HIVAN) is an inflammation of the little filters in kidneys that eliminate.

II. LITERATURE REVIEW

[1]A portion of the literary works are with respect to of medication improvement, hostile to viral specialists advancement (Kirchmair et al. 2011), antiretroviral reaction forecast (Zazzi et al. 2012 Prosperi 2011 and Prosperi et al. 2009), antiretroviral opposition expectation (Zazzi 2016) [2]Riemenschneider 2016a Riemenschneider 2016b Heider et al. 2013 and Kijsirikul 2008), antiretroviral unfavorable impacts expectation (Adroveret al. 2015) has been investigated which

[3]A review of the use of machine learning approaches in studying HIV/AIDS infection was previously published . Te paper by Lee et al. used machine learning approaches in classifying patients with and without the toxicity of biomarkers of mitochondrial in HIV. [4]Recently, Orel et al. used machine learning techniques on the Demographic Health Survey of 10 countries to identify HIV Positive individuals.

III. METHODOLOGY

Machine learning entails the utilisation of computational and statistical algorithms to determine hidden associations of data that might increase predictions through relaxation of the modelling postulates advanced by standard approaches. Among the recent advances in prediction tools and identifcation techniques in HIV statistical data, machine learning offers greater capability in processing huge amounts of data.

DATASET

We use the dataset to analysis of based on age structure of the dataset population HIV Impact assestment that comprasis of cross sectional of family evaluated of family HIV related key wellbeing marker overseas and carries out the project Project is surveying projects of HIV in nations upheld AIDS by public family studies. We take data included individual tested for HIV in our prediction to take ananlysis report survey of data.

HIV/AIDS positive cases respectively growing day by day it is not only dangerous(harmful) physically but also mentally characteristics of the dataset are use to show in structured format we consider to HIV test outcome for respondent positive and negative require to construct of binary using machine learning

K-closest neighbors (KNN)

The k-nearest neighbors (KNN) algorithm is a **simple, supervised machine learning algorithm** that can be used to solve both classification and regression problems. It's easy to implement and understand, but has a major drawback of becoming significantly slows as the size of that data in use grows.

ISSN 2394 - 7780

The KNN Algorithm

1. Load the data
2. Initialize k to your chosen number of neighbors
3. For each example in the data
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.
5. Pick the first K entries from the sorted collection
6. Get the labels of the selected K entries
7. If regression, return the mean of the K labels.
8. If classification, return the mode of the K labels.

NAIVE BASE CLASSIFIER

It is basically based on Byes theorem, which gives a mathematical framework for describing the probability of an event that might have been result of two or more causes:

21 (/)() (/) ( ) pb a pa pa b p b = This equation describes the probability p for state a existing for a given state b. The importance of Bayesian theorem is that probabilities of occurring new things depends upon existing knowledge. This is frequently used in chemo informatics both generally for predicting biological rather than physicochemical properties, prediction of toxicity of compound.

MODEL VALIDATION

Our machine learning task was structured to solve a binary classifcation problem. Our dataset comprises healthy individuals labelled negative in one class while the infected individuals are labelled positive in the other class.

We haphazardly picked from a matrix 50 arrangements of control upsides of the learning system (hyperparameters), and these were utilized in preparing also approval of information utilizing every one of Elastic Net (EN), k-Nearest Neighbors (KNN)],Random Forest (RF),Support Vector Machine (SVM), XG-Boost also Light Gradient Boosting (LGBT) calculations, Not really set in stone the normal scores of f1 for each of these 50 sets with a five-overlay get approval plan over the approved examples and the most impressive arrangement of hyperparameters were picked, Fig.1, stage 2. f1 score is a metric that is the most-utilized individual from the parametric group of the f-measures, named after the boundary esteem β=1, where beta is a variable of review significance than accuracy. It is characterized as the consonant mean of accuracy and review.

IV. CONCLUSION

We take historical to predict the HIV infection in human being to solve basic social issue of HIV/AIDS .The technology can involve everyone current days Artificial Intelligence growing and in each and every sector. Solve the problem Every sector like manufacturing , Automation ,Medical sector, using machine learning and some historical dataset to solve HIV/AIDS one of the major issue.

Machine learning cutting edge technologies would provide sound effect in development as comparative analysis is done in seconds. Studies shows KNN prediction is good then others.

V. REFERENCES

[1] ‘UNAIDS data 2020’. https://www.unaids.org/en/resources/documents/2020/unaids-data. Accessed 29 Oct 2020.

[2] ‘Zambia UNAIDS’. https://www.unaids.org/en/regionscountries/countries/zambia. Accessed 10 Nov 2020.

[3] ‘Fast-track commitments to end AIDS by 2030 | UNAIDS’. https://www.unaids.org/en/resources/documents/2016/fast-track-commitments.vAccessed 30 Nov 2020.

[4] ‘2016 United Nations Political Declaration on Ending AIDS sets world on the Fast-Track to end the epidemic by 2030’. https://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2016/june/20160608_PS_HLM_PoliticalDeclaration. Accessed 29 Oct 2020.

[5] Jewell BL, et al. Potential efects of disruption to HIV programmes in subSaharan Africa caused by COVID-19: results from multiple mathematical models. Lancet HIV. 2020;7(9):e629–40. https://doi.org/10.1016/S2352- 3018(20)30211-3.

[6] Dorward J, et al. The impact of the COVID-19 lockdown on HIV care in 65 South African primary care clinics: an interrupted time series analysis. Lancet HIV. 2021;8(3):

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

RISE AND IMPACT OF DIGITAL PAYMENTS IN INDIA DURING PANDEMIC

[1]Dr. Neelam Naik and [2]Shubh Gopani
[1]Assistant Professor and [2]MSc I.T, Sem I Student, Usha Pravin Gandhi College of Arts, Science & Commerce, Mumbai

ABSTRACT
India has started to become a cashless economy with the governments digital India program and with the launch of Unified Payment Interface (UPI). It has become easy for a user to receive and pay anyone with the help of their smartphones. The Covid-19 has pushed people to make more use of digital payments. However, the closedown of business in lockdown had put some break on the growth of digital payments in India. In this paper the author has studied was the growth in the digital payments due to pandemic or whether there were other reasons which gave digital payments a boost in India, the relation between consumption growth and digital payments with the help of time series analysis and has also discussed the value to digital transaction to Gross domestic product ratio in India.

Keywords: Digital payments Covid-19, UPI, Mobile wallets.

I. INTRODUCTION

Digital payments are transactions which takes place through online mode via Internet, with no hard money involved in it. This means that both the parties use electronic mode to exchange their money. No hard money is engaged with the digital payments. All the exchanges between payee and payer are finished on the web. Digital payments play an important role for the economy in this pandemic. In the pandemic situation where there were lockdowns and people were forced to stay at their home, maintain all the precautionary measures. Digital payments got a boost due to this. All the non – essential small shops were closed. To avoid dealing with cash people started paying through digital payments methods. Digital payments increased fivefold**,** from 1,004 in 2016-17 to 5,554 cores in 2020-21. Till mid-November this fiscal, the total number of digital transactions stood at 4,683 crores as per moneycontol.[6] An exponential spurt in online shopping and the pandemic also gave digital transactions a boost.

The digital India program of the government was launched by Indian Prime Minister Narendra Damodardas Modi on 1 July 2015. The campaign was launched to make government programs available electronically through the improvement of online infrastructure. and increased Internet connectivity, with the vision of transforming India into a digitally empowered society and knowledge economy.

The Digital India program focuses on three key fields of vision.

a. Digital Infrastructure as a Core Utility to Every Citizen

b. Governance & Services on Demand

c. Digital Empowerment of Citizens

Prime Minister Narendra Modi had announced the demonetization of old Rs 1,000 and Rs 500 banknotes on November 8, 2016, with the goal of reducing black money and encouraging people to use digital payments. Demonetization is also one of a reason people came to know about digital payments and started using it.

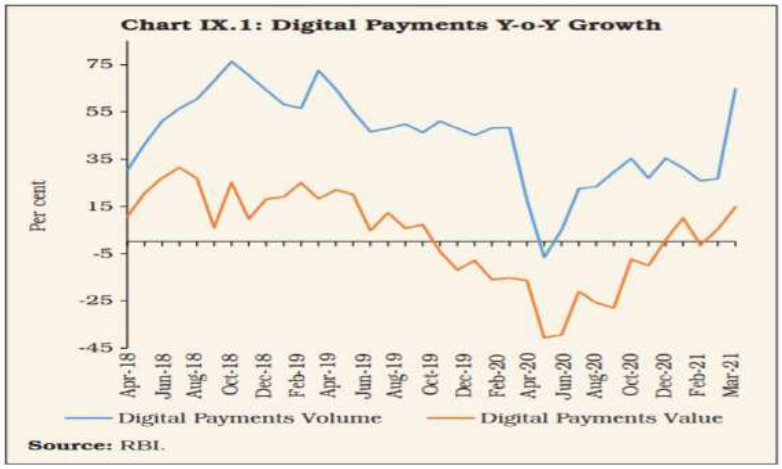


Fig 1: Digital payment year on year growth. [10]

ISSN 2394 - 7780

In figure 1 shows us that between April 2020 to June 2020 at the time of start of pandemic in March 2020 there was a dip in digital payments but soon after that there is a sharp rise in value as well as volume of digital payments.

During the early stages of the country wide lockdown due to the COVID-19 epidemic, payments decreased. With the progressive easing of the lockdown, however, the value and number of payments increased. In recent years, the industry has been changed by the Digital India program, demonetization, a slew of creative FinTech businesses, and government efforts.

It can be deduced from the foregoing that there was an increase in digital transactions during the epidemic, but it cannot be concluded that if this was due only to the virus or if there were other variables at play.

II. PROBLEM STATEMENT

Since Demonetization and start of pandemic, the Government of India has been emphasizing on going cashless, making transactions smoother and more transparent, and to eliminate the risk of Covid 19. Though the number of digital payments is increasing it cannot be said whether it is due to pandemic or there is any other reason. The main reason for taking this research is understand how much increase is there in digital payments due to pandemic.

III. LITERATURE REVIEW

The author of the paper [12] has shown innovative types of digital payments, the usage and importance of digital payment services. The examination found that the installment framework activities taken by the Govt. also, RBI have brought about more noteworthy acknowledgment and more profound entrance of non-money installment modes in India.

The author [1] has shown impact of Unified Payment Interface on payments system, Impact of UPI on business, Expansion of UPI, its international expansion in Singapore. The study found that UPI is the most progressive payment system in the world. UPI works on a safe, secure and healthy platform with ample safety features to make it more secure than any existing payment systems. UPI can be a great technology for financial institutions in India and enable a huge set of population to be a contributor of digital economy.

The author [13] has shown various types of digital payments and shown the year-on-year growth of transaction for particular payment mode. The study found that the trend of converting to cashless systems which is somewhat exponentially increasing, it has rapidly increased due to the onset of COVID-19 pandemic. It also stated that rise in digital payments has increased the chances of frauds.

The author [2] investigations disclose that digital wallets are quickly becoming the primary form of online payment. Customers adopted advanced wallets at the end at an incredibly fast pace, largely because of comfort and convenience.

The author [3] has shown importance of digital payments, the study found that digital payments across India grew at a compound annual growth rate (CAGR) of 55.1 percent over the past five years between the financial year (FY) 2015-16 and 2019-20, Its value has risen from INR 920.38 lakh Cr to INR 1623.05 lakh Cr during this time. Clipping at a 15.2 percent compounded annual rate.

The author [16] has shown discusses different digital payment methods used in the event of a pandemic based on primary data by gathering data from 220 respondents. The study found that it is too early to conclude what the changes might look like in each cultural, demographic, and institutional context, it was further said, it is already supporting existing tendencies toward more payment digitization.

The author [11] in this study has revealed that the traditional system of cash transaction cannot completely be replaced by card or e-payment system. The trust is the main factor affecting users' satisfaction directly and it impacts on many user's intention to adopt mobile wallets.

The article [5] states that after demonetization, many people started electronic payments for their transactions. Everyone from the small merchant to neighboring vegetables vendor is embracing digital payment solution. After demonetization, alternative payment methods like as cards, net banking, and mobile banking became more or less equivalent. The act of demonetization pushes merchants to adopt digitization to some degree. After Demonetization people knew the use of digital payments, so due to the pandemic it gave rise to use of digital payments as people were aware about its usage.

The author [14] states that UPI is indeed a revolution in the Indian economy. How-ever its success depends on various factors. Financial inclusion, or access to financial services, is required for the rise of UPI, which is aided

ISSN 2394 - 7780

by the Pradhan Mantri Jan Dhan Yojana (PMJDY) and the growing use of smartphones. The capacity of banks to capitalize on client trust, promote UPI successfully, and compete with mobile wallets will all be decisive considerations. UPI may collapse if banks fail to establish a successful front-end platform.

In his article [4] revealed about the situation at the time of demonetization. The researcher attempted to investigate the impact of demonetization on financial technology companies. During the demonetization era, the researcher also examines the payment service industry.

Fast Moving Consumer Goods (FMCG) firms have prolonged their credit cycles to solve the liquidity crisis, and some FMCG companies have given credit to distributors using RTGS. From a technological standpoint, digital payment is the best bet in the mobile internet sector.

The above papers help the author to understand the importance of digital payments in India. The impact of UPI on digital payments and its increase in usage in the time of pandemic. The papers also says that there was and an increase in digital payments due to pandemic, which helps the author to say that there was an increase due to pandemic and need to find how much increase was there due to pandemic.

IV. RESEARCH METHODOLOGY

In order to study the rise and impact of pandemic on digital payments, secondary data from Reserve bank of India's (RBI) website [7] has been studied and analyzed.

ABOUT DATASET

The data has been collected from November 2019 to October 2021. It contains three attributes which are

a. Year and Month

b. Volume of Transactions (Lakh)

c. Value of Transactions (₹ Crore)

Each volume and value are total of

a. CCIL Operated Systems

b. Credit Transfers – Retail

c. Debit Transfers and Direct Debits

d. Card Payments

e. Prepaid Payment Instruments

f. Paper-based Instruments

The data contains total 24 records each for value and volume according to their months.

V. Method of analyzing the data

The data has been analyzed with help of excel and line chart to understand whether there was an increase in digital payments because of pandemic.

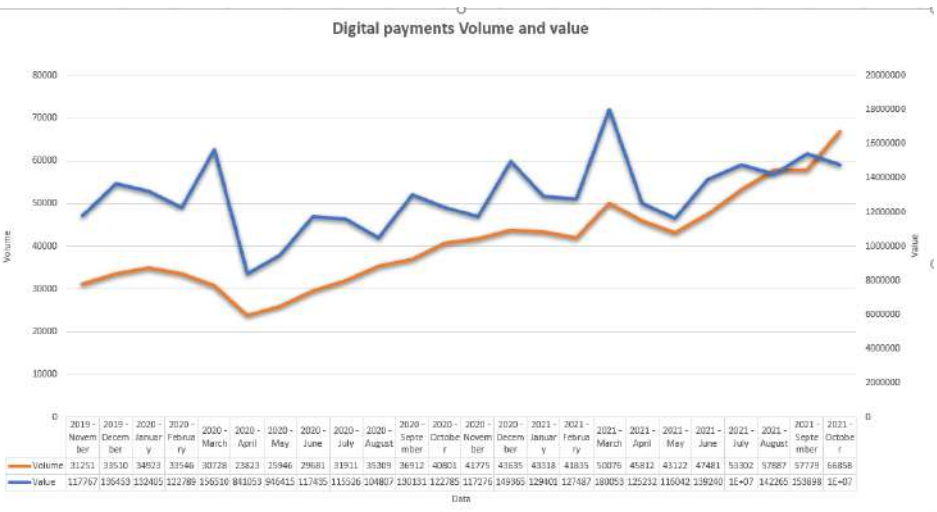


Fig 2: Digital payments volume and value

ISSN 2394 - 7780

In the figure 2 the red line is used to show the volume of payments, whereas the blue line is used to show the value of transactions. Below the line graph the data has been shown which has been used for drawing the line chart.

On the basis of the above chart, it can be said that there was a dip in payments during the start of the pandemic but soon after it regained its pace.

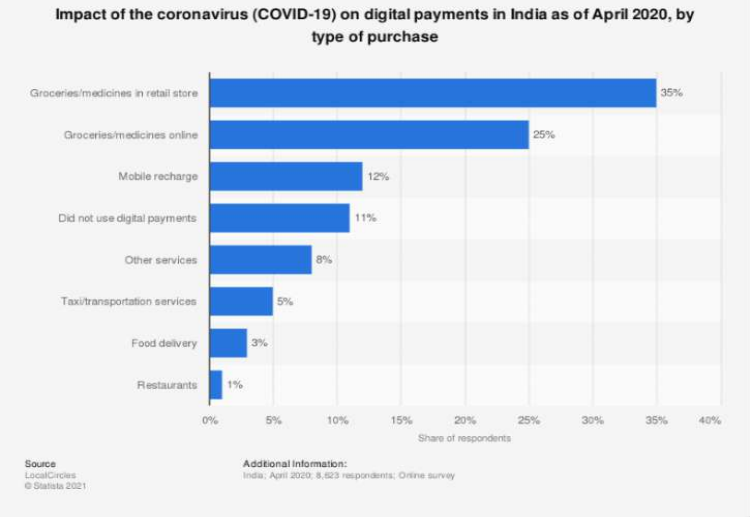


Fig 3. Impact of the Covid-19 on digital payments in India by type of purchase. [8]

In the figure 3 it shows the results of an online survey which was done by taking inputs from 8623 respondents on where were digital payments made to purchase products during the time of pandemic. Around 35% of digital payments were made to Groceries/Medicine stores. 25% of them were made to order Groceries/Medicine online. 12% for Mobile recharges as all the local shops were closed. There were also 11% of respondents who did not opt for digital payments. 8% of respondents paid for other services. While 5%,3%,1% of respondents used it for transportation service, food delivery and at restaurants respectively.

VI. `What is Time Series analysis?

Time series analysis is a method for studying a collection of data points over a period of time. Instead of capturing data points sporadically or arbitrarily, analysts use time series analysis to capture data points at constant intervals throughout a specified length of time. This form of study, on the other hand, is more than just gathering data over time. The ability to depict how variables change over time differentiates time series data from other types of data.

Time series analysis can be used be used to find whether increase in consumption also leads to increase in digital payments. The author in his study [9] has shown a relation between growth in consumption with growth in digital payments.

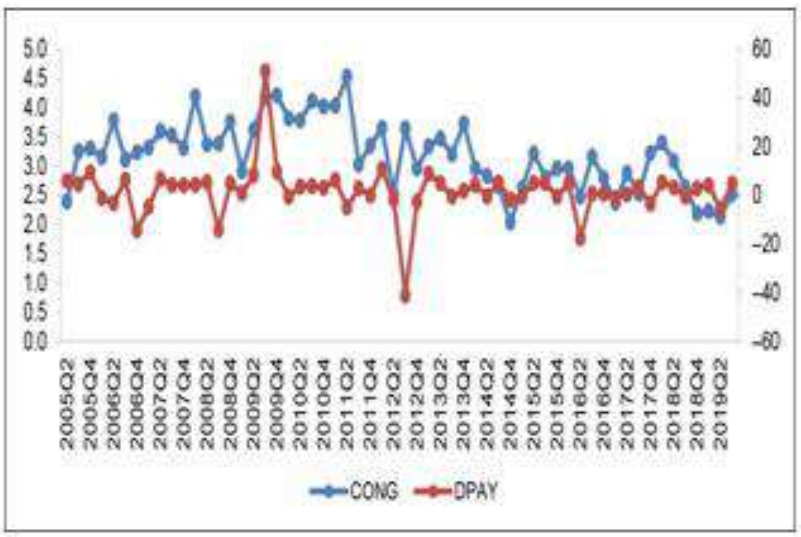


Fig 4. Consumption Growth and Digital Payment in India. [9]

ISSN 2394 - 7780

In the figure 4 which has been generated with time series analysis. The blue line represents the consumption growth whereas the red line represents the digital payments in the India.

In the year 2009 quarter 2 there was a sharp increase in consumption growth which can also be seen in case of digital payments. Whereas in the year 2014 quarter 2 there was dip in consumption which leaded to decrease in digital payments.

From the above observations it can be interpreted that when there is a rise in consumption growth there is a rise in digital payments. Which shows as people start making more consumption there will a rise in digital payments in the country?

VII. Value of digital transactions to Gross domestic product

Gross domestic product is the total worth of products and services generated within a country's geographic limits during a given time period, usually a year. The GDP growth rate is a key indication of a country's economic performance.

"Value of Digital Transactions (VDT) to Gross domestic product ratio" is an indicator growth of digital payments in terms of value. It is calculated by dividing the total amount of digital payments in a given year by the GDP of that same nation in that same year.

Particulars	31st March 2014	31st March 2015	31st March 2016	31st March 2017	31st March 2018	31st March 2019
Number of digital transactions per capita per annum	2.38	4.06	5.44	10.73	13.15	22.42
VDT to GDP Ratio	542%	561%	579%	644%	726%	769%
CIC to GDP Ratio	11.4%	12%	12%	9%	11%	10.7%
Average value of retail digital payment per transaction (Rs)	16,878	15,180	13,446	12,510	13,295	13,447

Fig 5. Number of digital transactions per capita per annum, VDT to GDP Ratio, and CIC to GDP Ratio in India since March 2014. [15]

In the figure 5 It can bee seen that there is an increase in VDT to GDP Ratio year on year. The value of digital transactions as a percentage of gross domestic products has increased from 542 percent in 2013-14 to 769 percent in 2018-19. It demonstrates the extent and potential of the Indian digital payments sector, which is open to market participants.

VIII. CONCLUSION

From figure 2 it can be interpreted that during the start of pandemic in India that is in April 2020 there was dip in digital payments as whole country was in Lockdown. There was 30% drop in digital payments during the start of the lockdown. But after April 2020 till August 2020, there can be seen a rise and recovery in Digital payments. There was growth of 23% within 3 months of the lockdown in the country, which indicates that people had started switching to digital payments during the pandemic. As there was ease in lockdown and consumption growth started to get increase the volume and value of digital payments of digital payments also started to increase. Though there is rise in digital payments post during the pandemic, it cannot conclude that the rise was only because of covid-19. The value of digital transactions as a percentage of gross domestic product has increased from 542 percent in 2013-14 to 769 percent in 2018-19. But it can surely conclude that covid-19 had some impact on the rise of digital payments during the pandemic.

IX. REFERENCES

[1]. Akshay Bhargava, Mohammad Ubaid, Younus Khan, Prabhat Chandra Gupta," Expansion of Unified Payment Interface", Annals of R.S.C.B, volume:25, Issue 6, 2021.

[2]. Dr. Stitch Shewta Rathore (2016) "Appropriation Of Cashless transactions By Consumers" Issue 4, December 2007.

[3]. Dr Nirmala M, Parvathi S," THE IMPACT OF PANDEMIC ON DIGITAL PAYMENTS IN INDIA", Journal of the Maharaja Sayajirao University of Baroda, Volume-55, No.1 (I) 2021.

ISSN 2394 - 7780

[4]. Dr.V.Sornaganesh and Dr.M.Chelladurai (2016), "Demonetization of Indian currency and its impact on business environment" International Journal of Informative and Futuristic Research Vol-4, Issue-3 November2016.

[5]. G.sudha,M.thangajesu sathish, "Demonetization: the way to increase the digital cash",International journal of creative research thoughts, Volume 8, Issue 5 May 2020.

[6]. https://www.moneycontrol.com/news/business/markets/digital-payments-script-robust-growth-story-in-india-cards-in-race-7788161.html

[7]. https://www.rbi.org.in/scripts/BS_ViewBulletin.aspx.

[8]. https://www.statista.com/statistics/1111084/india-coronavirus-impact-on-digital-payment-purchases/

[9]. https://www.adb.org/sites/default/files/publication/696286/adbi-wp1249.pdf

[10]. https://m.rbi.org.in/Scripts/AnnualReportPublications.aspx?Id=1322

[11]. M. Thangajesu Sathish, R. Sermakani, G.Sudha,"A Study on the Customer's Attitude toward the E-Wallet Payment System", IJIRT, Volume 6 Issue 12.

[12]. Rashi Singhal,"Impact and importance of digital payment in india",International journal of multidisciplinary educational research,volume:10, issue:2(3), february:2021.

[13]. Rajat Rajesh Narsapur, Apurva Parasar," An Analysis on the Rise in Digital Transactions in India".

[14]. Roshna Thomas and Abhijeet Chatterjee (2017) 'Adoption of Digital Wallet by, Consumers'. BVIMSR's Journal of Management Research, Vol. 8 Issue -1 April : 2016.

[15]. Ravikumar T, Murugan N, Suhashini.J, Rajesh R," Digital Payments Diffusion in Emerging and Developed Economies", International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-9, Issue-4, February 2020.

[16]. Sudha.G, Sornaganesh.V, Thangajesu Sathish.M, Chellama A.V," Impact of Covid-19 Outbreak in Digital Payments", International journal for innovative research in multidisciplinary field, Volume - 6, Issue - 8, Aug – 2020

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

SAP PROFITABILITY AND PERFORMANCE MANAGEMENT (PAPM)

[1]Neelam Naik and [2]Ramesh Patel
[1]Assistant Professor and [2]Student, MScIT, Part II, Usha Pravin Gandhi College of Arts, Science and Commerce. Mumbai, India

ABSTRACT
SAP Profitability and Performance Management (PaPM) is a unique solution that enables the business to run complex calculations and perform certain functions. SAP Profitability and Performance Management (PaPM) is not new in market as it was earlier known as FS-PER. It is just a rebranding of FS-PER with some upgraded capabilities. Before PaPM allocation of data was done manually by business which was very time consuming as well as inefficient. PaPM solves this issue as it works as an automation tool for allocation, etc. The purpose of this article is to explore PaPM as a tool used for Allocation and Simulations.

Keywords: SAP, Performance Management, Simulation

I. INTRODUCTION

"SAP Profitability and Performance Management (PaPM) is new technology solution which can be used to maintain and execute complex facts calculation, allocation, simulation and to optimize the overall performance of the business". PaPM application does not require their own data model. It can implement and reimplement existing data and information models from SAP, Non-SAP, or third-party applications. PaPM is developed by MSG-Global along with SAP as a successor of SAP Profitability and Cost Management (PCM).

"SAP Profitability and Performance Management (PaPM) is implemented on in-memory SAP HANA platform". An in-memory database stores all the data in the main memory/RAM of the computer. Earlier databases use to retrieve data from disk drive and then use it. However, with in-memory concept databases have evolved and there is no need to retrieve data from disk anymore as it can store data in RAM and use it for on the go calculations making the process faster. Since PaPM uses in-memory database its processing speed increases rapidly, which enable it to process a huge amount of data in relatively less time. SAP Profitability and Performance Management is built for business and provides an instant insight by using a single source of code, which provides real-time results. PaPM can be deployed each within the cloud and on-premise.

SAP Profitability and Performance Management (PaPM) is used to allocate business data which makes it easier for month end close activities and Audit. Using PaPM profitability results users can make future business decisions easily. Literature is not available on this topic.

II. FUNCTIONALITIES OF SAP PROFITABILITY AND PERFORMANCE MANAGEMENT

DATA AGGREGATOR

The data aggregator of SAP Profitability and Performance Management (PaPM) has the abilities that allows the integration of software programs and data warehouses at high velocity with little or no data duplication. Data aggregator can integrate SAP with non-SAP databases and programs. For example, Databases – Oracle, Teradata, HADOOP, etc.; Software programs- SAP S/4 HANA, SAP BW, SAP BPC, etc. SAP Profitability and Performance Management uses the official software interfaces from the SAP or non-SAP application to perform data read and write access activities smoothly.

CALCULATION ENGINE

The calculation engine of SAP Profitability and Performance Management enables business users to develop and execute data models by configuring and combining them across functions. The calculation engine can be considered as the brain of PaPM. With the help of calculation engine PaPM can process thousands of records in few minutes. This is achieved because PaPM calculation engine process data on thousands of cores in parallel. Since thousands of cores run in parallel the data which needs to be processed get segmented and assigned to respective core for processing. As it can process such huge amount of data in such a short time along with its write back capability PaPM can be used as a tool for planning. However, the main use of calculation engine comes into play for allocation purpose.

SIMULATION

The simulation ability of SAP Profitability and Performance Management enable the business to perform what-if scenarios. This gives the users insights on how the business is performing in terms of profit and performance in the market. Simulation provides high level drill down capabilities so that users can get details about business at a very detail level. Hence gives a transparent view of business as it offers auditable and

ISSN 2394 - 7780

traceable information. It allows non-SAP and SAP business intelligence tools like webi, AO to access data in real-time. Due to its simulation capability PaPM can be used as a data analysis tool. We can create queries in PaPM and have reports generated on top of those queries for better business understanding.

III. STATUS OF THE RESEARCH/KNOWLEDGE IN THE FIELD AND LITERATURE VIEW

SAP PaPM application are implemented by many. Every developer has their own way of designing the models. No literature is available on the internet with respect to this. This is a hypothetical scenario in which I'll mention how the model was implemented and how we could have done it in a better way.

The model was designed with a lot of functions and too many calculations done in different functions. This results in performance degradation of the model. When we develop something we make sure it is working as expected but at the same time it should be very good in terms of performance as well.

In this above scenario what could have done to improvise the model was, instead of creating multiple functions one should make sure to use less no. of function as possible. Because the more the number of functions the more time for them to process.

Also, the calculation's implemented in the functions takes time to process so, always try to implement calculations in on function if possible. This will let the calculation engine to calculate everything in one go rather than going to different functions.

There are many such ways in which we can build a better performance optimized model.

Figure 1: Model designed with multiple functions

IV. SAP PROFITABILITY AND PERFORMANCE MANAGEMENT AS A SOLUTION

SAP Profitability and Performance Management is a native digital performance management solution that helps in maintaining and executing complex calculations, rules, and simulations. As it is implemented on SAP HANA, it provides real-time business data aggregation capabilities for SAP and non-SAP systems. PaPM offers solutions for industry-specific content as well as cross-industry content.

To achieve all the above functionalities, we need to build PaPM functions and data models. All the major configuration activities of PaPM are implemented in an environment. An Environment consist of the information and calculation unit of PaPM. Environment is a shell that hosts all the other PaPM functions. Functions are the basis building blocks of PaPM. For example, consider a book – the book is the environment and the pages in it are the functions. Each function has its unique task and they can be reused. Functions can be independent or can be dependent on other functions. The functions which are dependent uses other functions as their data source. Each function has its own structure which comprise of – Header, Input, Signature, Lookups, Rules and Checks.

There are different types of functions in PaPM with each having their own functionalities.

- Allocation - This function allows to perform direct and indirect allocations in PaPM
- Calculation – This function allows to perform complex logical and mathematical operations
- Calculation Unit - These functions encapsulates a group of functions and make them available for reuse.
- Description - This function is used to describe processes and topics used for the documentation of models
- Environment - This function is used to register all required fields and the connection to the database
- File Adapter - This function provides automated access to files
- Funds Transfer Pricing - This function can perform funds and liquidity transfer pricing calculations
- Join – This function is used to perform collections, joins, unions, and lookups.
- Model Table – This function allows to read and write access to a local or remote data table

ISSN 2394 - 7780

- Model View - This function allows to read access to a local or remote data table or view
- Model RDL - This function allows to read and write access to a local FRDP Results Data Layer
- Model BW - This function allows to read and write access to a local BW InfoSource like ADSOs
- Writer – This function can store data in a model table, model RDL or model BW
- View – This function is used project or aggregate data, including filtering options and formulas
- Remote Function Adapter – This function is used to perform an ABAP-based remote function call (for example, a call to a remote FI-GL posting BAPI)
- Query – It is a reporting function that allows the output and input of data.
  All these functions work together to perform data retrieval from ACDOCA, processing the retrieved data, performing calculation if required, allocating the data, and posting it back in ACDOCA post allocation. We can also perform data analysis, create reports, etc. When we execute the RFA (Remote Function Adapter) it performs allocation process by triggering all the preceding function that are in use. Once RFA is executed successfully the data is Allocated.

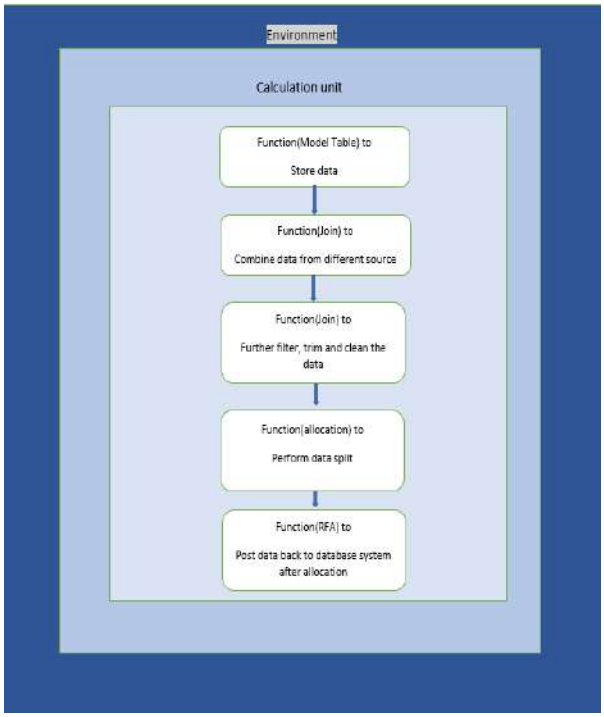


Fig 2. Model data flow

PaPM can be easily integrated with SAP BPC and SAP SAC. Business planning and consolidation (BPC) is an SAP tool used for planning business strategies and decision based on currents data analysis. BPC analyze data at a very granular level. The data fetched in BPC is real-time giving a clearer picture to business users of their market performance. SAP Analytics cloud (SAC) is a cloud data virtualization tool used for reporting, planning and data analysis. By integrating them with PaPM we can get real-time allocated data which can be used for data analysis and accordingly business users can take financial decision.

Apart from BPC and SAC users can get their business performance via PaPM as well. It has a query function which enables it to generate reports, graphs, and tables of post allocated data by RFA (Remote Function Adapter). The query function uses RFA function as and input and provides the same result as RFA in graphs and table form making it easier for business users to understand.

V. FEATURES OF PAPM

- **DEPLOYMENT**

PaPM can be deployed on premise i.e. on one central system or on cloud

- **CONNECTIVITY**

PaPM can be connected to SAP as well as non-SAP systems. It can use both SAP and non-SAP as its data source for data processing and allocation.

ISSN 2394 - 7780

- **PERFORMANCE**

As thousand of cores run in parallel for data processing, it can process complex calculation within minutes or hours. Since calculation is done fast, the allocation process is also improved significantly.

- **WRITEBACK**

With PaPM as a tool we can write our data back to the source system post allocation.

- **FRONTEND**

PaPM is a tool which is totally web based i.e. whenever we launch our environment it is opened as a web service makes it easier to work on.

- **INTEGRATION WITH BPC AND SAC**

With PaPM being integrated with BPC and SAC, users can make better business decision based on the quick data analysis due to fast data processing.

- **OTHERS**

PaPM provides Transparency, traceability, and auditability of data.

PaPM implements Cross-industry and industry-specific content.

VI. CONCLUSION

PAPM is at the verge of turning into a effective tool with its new iterations bringing advanced balance and integrations with numerous different systems. It maybe a perfect Analytical tool for allocation of costs for a business model which will compliment SAC within the mere future.

VII. REFERENCES

1. SAP Profitability and Performance Management, https://en.wikipedia.org/wiki. (Retrieved 05 November 2021)
2. SAP PaPM, https://www.mindtree.com/insights/blog/sap-papm-powerful-new-solution-or-just-another-etl-tool (Retrieved 07 November 2021)
3. PaPM | SAP Blogs, https://Blogs.sap.com (Retrieved 07 November 2021)
4. SAP PaPM ETL tool, https://Sap.com (Retrieved 08 November 2021)

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

DETECTION OF CHRONIC KIDNEY DISEASE USING MACHINE LEARNING MODELS

[1]Smruti Nanavaty and [2]Pranav Kateliya
[1]MScIT Coordinator and [2]M.Sc. Student, Information Technology Department, Usha Pravin Gandhi College of Arts, Science and Commerce

ABSTRACT
The kidney is considered to be one of the most important parts of the human body. Its main function is to filter extra fluid and remove wastes from the body. Chronic kidney disease is a condition of the kidney where the kidney loses its filtration capacity and does not perform as it should. Sometimes the patient is not aware of the presence of this disease until it becomes severe. Chronic means the damage to the kidney rises slowly over a long period. This slow and gradual rise in chronic disease leads to the process of dialysis or transplantation of the kidney. Kidney transplantation becomes more costly or the chances for the patient of survival become less. The field of bioscience has become advanced and has generated a large number of electronic health records. The use of machine learning in the field of medicine can aid to detect the presence of chronic kidney disease at an early stage and a proper medication diet provided to the patient can help to recover. This paper uses data preprocessing, data transformation, and various classifiers to predict CKD. This paper uses the K-NN algorithm and Random Forest algorithm machine learning models for the classification of the patient who is suffering from CKD or Not.

Keywords: Chronic Kidney Disease, K-NN, Random forest, Machine learning, Dialysis.

I. INTRODUCTION

Chronic kidney disease (CKD) is a condition in which your kidneys have been damaged and are no longer able to filter blood properly. Because the damage to the kidneys happens gradually over time, the condition is called "chronic." Waste may build up in your body as a result of this harm. It can also lead to other health issues. Many patients have no symptoms until their renal disease has progressed to the point that they can no longer function. The only way to tell if you have kidney disease is to have a blood or urine test. Chronic kidney failure is a condition in which the kidneys have been damaged and are no longer capable of filtering blood in the same way that a healthy kidney does. Excess fluids and blood waste accumulate in the body as a result, which can lead to various health issues like heart disease and stroke.

A. BLOOD TEST:

1) A GFR blood test: This test is used to determine how well the blood is filtered, which is referred to as GFR. Glomerular filtration rate (GFR) is an abbreviation for glomerular filtration rate.

TEST RESULTS:

i. The kidneys are functioning normally if the glomerular filtration rate (GFR) is 60 or above. The normal range is defined as a glomerular filtration rate (GFR) of 60 or above.

ii. If the patient's glomerular filtration rate is less than 60, they have renal disease. The health care professional should recommend the proper medicine.

iii. Kidney failure is diagnosed when the GFR is 15 or below. People with a GFR of less than 15 must either have a kidney transplant or stay on dialysis.

2) Creatinine: Creatinine is a waste product produced when your body's regular muscles break down. The kidneys remove creatinine from the blood and excrete it from the body. The quantity of creatinine in your blood is used by doctors to determine your GFR. Creatinine levels grow as the renal disease progresses.

B. Albumin urine test: This is a test that checks for albumin in the urine. Albumin in the urine is a clear indicator that the patient's kidneys have been damaged. The protein albumin is present in the blood. Albumin cannot pass through healthy kidneys into the urine. Some albumin is able to pass through the injured kidney and into the urine. The lower the albumin concentration in the urine, the better. Albuminuria is the presence of albumin in the urine.

Albumin in Urine Results:

i. If 30 mg/g or less is considered normal.

ii. Exceeding 30 mg/g might indicate renal disease.

Diabetes and hypertension are the two most common causes of chronic kidney disease, accounting for up to two-thirds of cases. When blood sugar levels are excessively high, diabetes develops, causing damage to

ISSN 2394 - 7780

numerous organs in the body, including the kidneys and heart, as well as blood vessels, nerves, and the eyes. When blood pressure rises against the walls of blood arteries, it is known as hypertension. High blood pressure, whether uncontrolled or poorly regulated, is a primary cause of heart attack, stroke, and chronic renal disease. Hypertension can also be a symptom of chronic renal disease.

II. LITERATURE REVIEW

"Early detection and prevention of chronic kidney disease". Author : Maithili Desai[2]. In many of these chronic conditions, data mining techniques can be utilized to uncover significant metrics that can be used to determine the disease's suffering. This can be done via data prediction, which is a low-cost, high-accuracy solution for folks who can't afford to run repeated experiments. Instead of doing every test, the algorithm assists in identifying a selection of them that will yield comparable results, saving you time and money. Because Boruta's study is free, it aids in a potentially costly medical diagnosis.

"Chronic Kidney Disease Detection Using Machine Learning Techniques", By N. Vanitha & S.V. Sendhura[1], Several machine learning methods can be used to detect chronic renal disease. We will test all of the aspects of machine learning approaches in the future to obtain the best accuracy. The suggested SVM algorithm effectively extracted the features and classified the samples with a 98.40 percent accuracy. Deep learning techniques can be used to attain high accuracy.

"Early prediction of chronic kidney disease using machine learning supported by predictive analytics". Authors,Ahmed J. Aljaaf1, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, Thar Baker, Abir J. Hussain, and Jamila Mustafina[7]. The RPART model, which is a classification and regression tree, produced fairly acceptable results. With these data, no information on any type of medication was gathered. The value of several metrics may be influenced by the drugs that have been prescribed. The MLP model had the best AUC and TPR, while the RPART model had the maximum TNR of 1.00, according to the data.

"Comparative analysis of machine learning methods to detect chronic kidney disease". Madhusree Sankar Roy, Ritama Ghosh, Diyali Goswami, Karthik R[3]. Comparing all models, Random Forest Classifier and Extra Tree Classifier provide the highest accuracy of 99.36%.

"A novel approach to predict chronic kidney disease using machine learning algorithms". Bhavya Gudeti, Shashi Mishra, Shaveta Malik, Terrance Frederick Fernandez, Amit Kumar Tyagi, Shabnam Kumari[5]. To carry out the CKD research, the support vector machine, logistic regression, and K-NN are investigated. The accuracy of the algorithms was used to determine their performance. Within the confines of this medical scenario, our findings demonstrated that the support vector machine algorithm predicts chronic kidney disease better than logistic regression and nearest neighbors.

"Prediction of chronic kidney disease using machine learning models". S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh[6]. Research results showed that the Random Forest Classifier model better predicts CKD compared to decision trees and support vector machines.

"Chronic Kidney Disease Prediction and Recommendation of Suitable Diet plan by using Machine Learning". Akash Maurya, Rahul Wable, Rasika Shinde, Sebin John, Rahul Jadhav, Dakshayani. R[4]. Data preprocessing, feature extraction, creating zones based on blood potassium level, and diet advice module are the four primary modules of the proposed system. A suitable statistical method, such as regression, extracts very efficient characteristics in the choice of CKD detection from the dataset, which comprises 25 primary features.

III. MACHINE LEARNING MODELS

Machine learning is a subfield of artificial intelligence that is described as a machine's capacity to emulate intelligent human behavior in a wide sense. Artificial intelligence systems are utilized to carry out complicated tasks in the same manner that people do.

A. Machine learning is divided into three categories:

1) **Supervised Machine Learning:** Models are trained on labeled datasets, which enables them to learn and improve over time. For instance, an algorithm may be trained on photographs of dogs and other objects, all of which were classified by humans, and the machine would learn how to recognize images of dogs on its own. The most popular sort of machine learning nowadays is supervised machine learning.
2) **Unsupervised Machine Learning:** The algorithm searches for patterns in data that haven't been labeled. Unsupervised machine learning can uncover patterns or trends that individuals aren't aware of. An unsupervised machine learning software, for example, may monitor online sales data and identify various sorts of clients making purchases.

ISSN 2394 - 7780

3) **Reinforcement machine learning:** Establish a reward system to train robots to do the optimal action through trial and error. Reinforcement learning may be used to teach models to play games or autonomous vehicles to drive by notifying the car when it has made the appropriate judgments, allowing it to understand what actions it should perform over time.

B. K-NN Algorithm

The K-Nearest Neighbors (KNN) method is a supervised machine learning technique that may be used to solve classification and regression issues. However, it is mostly employed in the business for the categorization of predicted issues. KNN is well defined by the following two characteristics:

1) **Lazy learning algorithm:** KNN is a lazy learning algorithm since it does not have a dedicated training phase and instead uses all of the data for classification training.

2) **Non-parametric learning algorithm:** Because it makes no assumptions about the underlying data, KNN is also a non-parametric learning algorithm.

C. Random forest Algorithm

Random Forest is a frequently used supervised machine learning technique for classification and regression tasks. It creates decision trees from several data, using a majority vote for classification and the mean for regression. One of the most essential characteristics of the Random Forest technique is its capacity to cope with data sets comprising both continuous and categorical variables, as in regression and classification. It leads to more accurate categorization results.

D. GINI INDEX

A Gini index, also known as a Gini impurity, calculates the likelihood of an attribute being erroneously scored when randomly picked. It is said to be pure if all of the elements belong to the same class.

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

Fig. 1: Formula for Gini index.

IV. DATASET AND ALGORITHM

A. Attribute Information

There are 25 characteristics in the dataset, 11 of which are numeric and 14 of which are nominal. The whole dataset's 400 instances are utilized in the training of machine learning algorithms. In 400 instances, 250 are diagnosed with chronic kidney disease (CKD) and 150 with non-chronic kidney disease (NCKD). The attributes present in the data set are bacteria, sodium, age, hemoglobin, diabetes mellitus, classification, appetite, coronary artery disease, blood pressure, pus cells, anemia, foot edema, sugar, white blood cell count, hypertension, red blood cells. Count, Potassium, Specific Gravity, Pus Cells, Packed Cell Volume, Albumin, Serum Creatinine, Red Blood Cells, Blood Urea, and Random Blood Glucose. The data set is separated into two groups: one for testing samples and the other for training samples. The proportions of test and training data are 30% and 70%, respectively.

Table II: Dataset Description.

Sr no.	Attribute	Description about the attribute
1	Bacteria(nominal)	ba (present / not present)
2	Sodium(numerical)	sod in mEq/L
3	Age (numerical)	Person's Age in Years
4	Haemoglobin (numerical)	Hemo in grams
5	Diabetes Mellitus (nominal)	dm (yes / no)
6	Class (nominal)	class (ckd / notckd)
7	Appetite (nominal)	appet (good / poor)
8	Coronary Artery Disease(nominal)	CAD (yes / no)
9	Blood Pressure (numerical)	BP in mm/Hg
10	Pus cell (nominal)	PC (normal / abnormal)
11	Anemia (nominal)	ane (yes / no)
12	Pedal Edema (nominal)	pe (yes / no)

ISSN 2394 - 7780

13	Sugar (nominal)	su (0/1/2/3/4/5)
14	White Blood CellCount (numerical)	Wc in cells/cumm
15	Hypertension (nominal)	htn (yes/no)
16	Red Blood Cell Count (numerical)	Rc in cells/cumm
17	Potassium (numerical)	Pot in mEq/L
18	Specific Gravity (nominal)	Sg - (1.005/1.010/1.015/1.020/1.025)
19	Pus Cell clumps (nominal)	pcc (present / notpresent)
20	Packed Cell Volume (numerical)	P cv
21	Albumin (nominal)	al (0/1/2/3/4/5)
22	Serum Creatinine(numerical)	Sc in mgs/dl
23	Red Blood Cells (nominal)	RBC (normal/ abnormal)
24	Blood Urea (numerical)	Bu in mgs/dl
25	Blood Glucose Random (numerical)	BGR in mgs/dl

B. Classification Accuracy

The created classifier model's accuracy may be determined using the following equation:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

Fig. 2: Formula for finding Accuracy

Where,

TP = Observation is positive an predicted is also positive.

TN = Observation is negative and predicted is also negative.

FP = Observation is negative but predicted is positive.

FN = Observation is positive but predicted is negative.

C. ALGORITHM

Input: chronic kidney disease dataset

Outputs: Finding classification accuracy.

Step 1: Enter the data

Step 2: Pre-process the data

Step 2.1: Convert nominal values to binary values.

Step 2.2: Replace the missing values with Mean in the respective columns.

Step 3: Build Classifier Models.

Step 3.1: Practice with the Gini Index.

Step 3.2: Train Using the Random Forest Algorithm.

Step 3.3: Train using the K-NN algorithm.

Step 4: Predict the test using a genetic index.

Step 5: Check the accuracy of the models built using a confusion matrix.

Step 6: Determine the best classification model for CKD.

V. RESULTS AND DISCUSSION

A. Accuracy of Gini index

The Confusion Matrix was generated by the Gini index fortest data (120 instances) with class (values: CKD, NON CKD) as the target variable is given by TABLE III. The confusion matrix clearly says that 4 instances are not correctly classified, 116 instances have been accurately classified, and the accuracy of this classifier model is 96.66%.

ISSN 2394 - 7780

Table III: Confusion Matrix of Gini index.

	NON CKD	CKD
NON CKD	39	1
CKD	3	77

B. Accuracy of Random Forest

Confusion matrix generated by Random Forest model of test data (120 cases) with category (values: CKD, NON CKD) as the target variable is given by TABLE IV. The confusion matrix clearly says that 120 instances have been accurately classified and the accuracy of this classifier model is 100%.

Table IV: Confusion Matrix of Random forest.

	NON CKD	CKD
NON CKD	40	0
CKD	0	80

C. ACCURACY OF K-NN

The Confusion Matrix was generated by the K-NN model to

Test data (120 cases) by category (values: CKD, NON CKD) as a target variable was introduced by TABLE V. The confusion matrix shows that 20 cases were not correctly classified 100 cases were classified with the accuracy and accuracy of this classifier model 83.33%.

Table V: Confusion Matrix of K-NN.

	NON CKD	CKD
NON CKD	33	7
CKD	13	67

VI. CONCLUSION

The models used for the classification of CKD for the given dataset achieve high accuracy. The use of these machine learning models in the field of bioscience can help in the early detection of CKD and can also help the patient for early treatment at the earliest. Proper medication and adequate diet major risk of kidney damage can be avoided. The random forest algorithm achieves the highest accuracy in classifying the presence of CKD or NOTCKD.

VII. REFERENCES

[1] N. Vanitha & S.V. Sendhura, "Chronic Kidney Disease Detection Using Machine Learning Techniques", AR J Med Sci; Vol-2, Iss-2 (Mar-Apr; 2021): 127-133.

[2] Maithili Desai, "Early Detection and Prevention of Chronic Kidney Disease", 978-1-7281-4042-1/19/$31.00 ©2019 IEEE.

[3] Madhusree Sankar Roy, Ritama Ghosh, Diyali Goswami , Karthik R, "Comparative Analysis of Machine Learning Methods to Detect Chronic Kidney Disease", doi:10.1088/1742-6596/1911/1/012005.

[4] Akash Maurya, Rahul Wable, Rasika Shinde ,Sebin John ,Rahul Jadhav,Dakshayani.R, "Chronic Kidney Disease Prediction and Recommendation of Suitable Diet plan by using Machine Learning", 978-1-5386-9166-3/19/$31.00 ©2019 IEEE.

[5] Bhavya Gudeti, Shashvi Mishra, Shaveta Malik, Terrance Frederick Fernandez, Amit Kumar Tyagi, Shabnam Kumari, "A Novel Approach to Predict Chronic Kidney Disease using Machine Learning Algorithms", IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1.

[6] S.Revathy, B.Bharathi, P.Jeyanthi, M.Ramesh, "Chronic Kidney Disease Prediction using Machine Learning Models", ISSN: 2249 – 8958, Volume-9 Issue-1, October 2019.

[7] Ahmed J. Aljaaf, Dhiya Al-Jumeily, Hussein M. Haglan, Mohamed Alloghani, Thar Baker, Abir J. Hussain, and Jamila Mustafina, "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics", 978-1-5090-6017-7/18/$31.00 ©2018 IEEE.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

DEVOPS IN CLOUD GIANTS: A COMPARATIVE STUDY

[1]Sunita Gupta and [2]Diti Majithia
[1]Assistant Professor and [2]M.Sc. (IT), UPG College of Arts, Science & Commerce

ABSTRACT
DevOps is an evolving practice to be followed in the Software Development life cycle. The name DevOps shows that it's an amalgamation of the Development and Operations team. It is gaining fame because of its continuous approach – continuous integration (CI), continuous deployment (CD), and software delivery. This paper studies the building blocks of DevOps, the application of DevOps in cloud giants, and their comparison.

Indexterms: DevOps, Cloud Computing, DevSecOps, AWS, Azure

I. INTRODUCTION
Cloud computing is software architecture built on apps that store data on remote servers accessible over the internet. There are two types of cloud computing: front-end and back-end. Using an internet browser or a cloud computing application, a user can access data stored in the cloud via the front end. The backend, on the other hand, is an essential aspect of cloud computing because it is in charge of securely storing data and information. It includes servers, computers, databases, and central servers. The central server eases operations by following a set of protocols. [1]

DevOps is a set of methodologies for communicating and collaborating between developers and operations to provide software and services more quickly, reliably, and with greater quality. The term 'DevOps' was coined by combining the words 'Dev' and 'Ops.' From development to deployment and support, DevOps is the division of jobs and responsibilities among a team empowered with complete accountability for their service and its underlying technology stack. DevOps emphasizes the automation of change, configuration, and release procedures in order to handle continuous releases. [2]

DevSecOps is a mindset that emphasizes "everyone is responsible for security" when it comes to IT security. It entails incorporating security procedures into the DevOps pipeline of a company. The goal is to integrate security into the software development process at every level. DevSecOps means you're not saving security for the end of the SDLC, which is in direct opposition to previous development approaches. [3]

Azure DevOps enables teams to organize work, collaborate on code development, and create and deliver apps using developer services. Azure DevOps promotes a culture and set of protocols that unite developers, project managers, and contributors to collaborate on software development. [4]

AWS offers services that assist you in implementing DevOps at your firm and are designed specifically for use with AWS. These solutions let teams manage complicated systems at scale, automate tedious processes, and keep engineers in charge of the high velocity that DevOps allows. [5]

II. REVIEW OF LITERATURE
The author's main purpose in writing this paper was to analyze and identify whether software quality gets improved when DevOps is practiced. Online questionnaires and interviews with DevOps specialists in the software development business were used to obtain data. Culture, automation, monitoring, and sharing all have an impact on product quality, according to research findings. As a result, if DevOps practices are followed appropriately, software quality will improve. [2]

This survey examines recent research on quality-aware software engineering tools to aid DevOps. The research context is examined, and some gaps in domains such as continuous integration/continuous delivery (CI/CD), incremental verification, and infrastructure-as-code are identified (IaC). [9]

In this research article, the authors have discussed the limitations in DevOps methodology and see how a missing security model can cause the problem of insecure development. They have also discussed the various techniques of DevOps and DevSecOps can support the development process. They have observed the challenges that were faced during the process without DevSecOps. Their explanation also provides the stages where DevSecOps can be adopted. The result shows that with DevOps one can develop a secure application ensuring the constraints of DevOps methods like speed. They have stated that their future work will be to implement the same concept on a project with a larger scale and scope. [10]

The fundamental role of DevOps in the digital transformation of service delivery is outlined in this paper as a collection of techniques and practices that complement IT as a Service and software product delivery. The

ISSN 2394 - 7780

application evaluation framework is presented as a means of preparing the application landscape for the DevOps transition. [11]

III. COMPARATIVE STUDY

DevOps is a methodology that combines development and operations, while DevSecOps is a subsection that emphasizes security. Although the two notions are not mutually exclusive, their objectives are rather different.

The differences in DevOps and DevSecOps can be classified into the following aspects:

Features	DevOps	DevSecOps
Philosophy	The collaboration of the Development and Operations teams helps increase their productivity.[6]	The aim of the DevSecOps team is to find inventive solutions by bringing down walls between development teams and IT, removing silos so that both parties can collaborate.[6]
Purpose	DevOps being significantly involved in the engineering process on a daily basis has its main purpose as speed.[6]	The main objective of DevSecOps is to offer premium security along with pertaining faster speed of process, accessibility, and scalability.[6]
Goal	DevOps focuses on bridging communication gaps between teams, aims to reduce risk while delivering high-quality software faster by focusing on collaboration, continuous integration, and automation.[6]	DevSecOps aims to create a safe and secure environment for sharing security choices while preserving the highest levels of security, speed, and control.[6]
Emphasis	DevOps emphasizes on Software Development.[6]	In order to reduce downtime and data loss, DevSecOps emphasizes the significance of developers writing secure and compliant code as their primary job.[6]
Security begins	The concept of security in DevOps begins right after the development pipeline.[6]	The application of security in DevSecOps begins during the build process.[6]

Both AWS and Azure aim to automate the software development lifecycle.

Features	AWS DevOps	Azure DevOps
Infrastructure	AWS's strategy has been to stay ahead of the innovation curve and provide services that IT operations can easily comprehend and use to meet their on-demand compute and storage require ments.[7]	Azure was designed as a holistic platform and has been geared toward businesses from the beginning. It originated as a platform as a service (PaaS), allowing developers to develop applications without focusing on the servers on which they're running.[7]

ISSN 2394 - 7780

Service Integration	AWS DevOps helps users to easily integrate AWS services like EC2, S3, and Beanstalk in just a few clicks.[7]	Azure DevOps enables customers to integrate Azure services such as Azure VM, Azure App Services, and SQL databases, as well as third-party tools such as Jenkins, thanks to their extensive ecosystem of third-party editions.[7]
Managing Packages in a Software	If you need to manage a package in AWS DevOps, it is required to integrate an external software or third-party software like Artifactory.[7]	Azure offers a package manager tool called Azure Artifacts that can handle the packages like Nuget, Maven, etc.[7]
Best Feature	AWS DevOps can simply automate a complete code deployment with AWS services. [7]	Azure DevOps has Kanban boards, workflows, and a huge extension ecosystem.[7]

IV. CONCLUSION

The building components of DevOps have been explored in this paper. The Cloud is an essential component of DevOps, as it aids with integration, continuous delivery, security, and feedback collection. Many architecture frameworks have been discussed, and each of these models has addressed some difficulties or enhanced existing DevOps benefits. Apart from all of these considerations, the team must be skilled and adaptable enough to deal with cultural shifts.

REFERENCES

[1] https://www.hcltech.com/technology-qa/how-does-cloud-computing-work

[2] P. Perera, R. Silva and I. Perera, "Improve software quality through practicing DevOps," 2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer), 2017, pp. 1-6, doi: 10.1109/ICTER.2017.8257807.

[3] https://www.plutora.com/blog/devsecops-guide

[4] https://docs.microsoft.com/en-us/azure/devops/user-guide/what-is-azure-devops?view=azure-devops

[5] https://www.edureka.co/blog/aws-devops-a-new-approach-to-software-deployment/#AWSDevOps

[6] https://www.bunnyshell.com/blog/devops-vs-devsecops

[7] https://k21academy.com/amazon-web-services/aws-devops-vs-azure-devops/

[8] M. Gokarna and R. Singh, "DevOps: A Historical Review and Future Works," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021, pp. 366-371, doi: 10.1109/ICCCIS51004.2021.9397235.

[9] A. Alnafessah, A. U. Gias, R. Wang, L. Zhu, G. Casale and A. Filieri, "Quality-Aware DevOps Research: Where Do We Stand?," in IEEE Access, vol. 9, pp. 44476-44489, 2021, doi: 10.1109/ACCESS.2021.3064867.

[10] Z. Ahmed and S. C. Francis, "Integrating Security with DevSecOps: Techniques and Challenges," 2019 International Conference on Digitization (ICD), 2019, pp. 178-182, doi: 10.1109/ICD47981.2019.9105789.

[11] L. Georgeta Guşeilă, D. -V. Bratu and S. -A. Moraru, "DevOps Transformation for Multi-Cloud IoT Applications," 2019 International Conference on Sensing and Instrumentation in IoT Era (ISSI), 2019, pp. 1-6, doi: 10.1109/ISSI47111.2019.9043730.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

LAND AND BOUNDARY SURVEYING USING GNSS TO DEMARCATE PROPERTY BOUNDARIES

[1]Smruti Nanavaty and [2]Zayed Lalljee
[1]Associate Professor and [2]M.Sc. I.T Research Student, Usha Pravin Gandhi College of, Arts, Science and Commerce, Mumbai

ABSTRACT
Numerous instances of pressing issues in rural areas such as unlawful land grabbing, encroachments, illegal occupancy of forest areas for cultivation and other commercial activities, land disputes without sufficient evidence to support the claims, measuring areas that were affected by floods and famines, etc. makes the need for land and boundary surveying a priority. Land Surveying and Land Boundary is avoided by a large mass because of the overhead cost of infrastructure and the lack of knowledge of the latest calculation tools. The current data available in hand is outdated or has been rustically assembled- manual drawing with the hand. The landmass moves on an average of 0.6 inches a year. This indicates that after 5 years, the data can be inaccurate and false claims can be made. This emphasizes on the need for a Land and Boundary Surveying. This research paper aims to bring forth a solution more accessible to the common man; so that there would be sufficient evidence and there would be a much simpler and feasible solution.

I. INTRODUCTION

Land surveying is the art and science of determining the exact location. Land surveying is a detailed examination or study that is carried out after obtaining information by observation, measurement, study and research of legal documents, data analysis, and other methods.

There are 3 major types of land surveys- **Construction or Engineering, Geodetic & Boundary or Land. Construction or Engineering** studies changes in property lines, the location of buildings, road topography and grade; **Geodetic** uses satellite and aerial imaging to measure large portions of the earth and **Boundary or Land** determines where properly lines are located.

There are **several equipments** that can be used for **Land Surveying**. However, the purchasing or renting cost of the equipment is not feasible and it is complex to use. There is a much **simpler and feasible procedure** that can be adopted with the same **high-level accuracy**. The procedure involves the use of a **simple GNSS device.**

Global Navigation Satellite System (GNSS) is a general term used to describe any satellite constellation that provides **Positioning, Navigation, and Timing (PNT)** services on a **global or regional basis**. GNSS, by definition, gives worldwide coverage. The Galileo navigation satellite system in Europe, the NAVSTAR Global Positioning System (GPS) in the United States, the Global'naya Navigatsionnaya Sputnikovaya Sistema (GLONASS) in Russia, China's BeiDou Navigation Satellite System, Indian Regional Navigation Satellite System (IRNSS) / Navigation Indian Constellation (NavIC) and Japan's Quasi-Zenith Satellite System (QZSS) are all examples of GNSS.

The Global Navigation Satellite System (GNSS) is based on the idea of **trilateration** and uses satellite constellations. Simply said, GNSS receivers compute their own location by calculating the distance between four or more satellites. These satellites would have originally all come from the same GNSS, although multi-GNSS receivers are now popular.

For a long time, the only GNSS options were **GPS and its Russian-owned equivalent (GLONASS)**. Because GPS was the more reliable of the two systems during this time — GLONASS was in disrepair and GPS became the most extensively used GNSS, and it remains so to this day.

With the rebirth of GLONASS, as well as the introduction of Europe's Galileo system and China's BeiDou, users and developers now have access to a wider variety of signals and all the benefits it entails, such as More satellites are available at any given moment, resulting in greater reliability; Greater precision can help offset impacts like air interference on GNSS precision by combining signals and frequencies; Multi-GNSS receivers are less vulnerable to mistakes produced in the space segment and are more difficult to spoof/jam.

The **GPS (Global Positioning System)** is a satellite-based navigation system with at **least 24 satellites**. With no subscription fees or setup charges, GPS works in any weather condition, anywhere in the world, 24 hours a day. The satellites were launched into orbit by the US Department of Defense (USDOD) for military purposes, but they were later made accessible for civilian usage in the 1980s.

ISSN 2394 - 7780

The GPS traditionally provides **Latitude, Longitude, Elevation, Speed and Time**. Based on the extraction of this data, it offers a host of application, on land, in the air or at sea. Because of its **parallel multi-channel construction**, today's GPS receivers are exceptionally accurate. When our receivers are first turned on, they quickly lock onto satellites. The accuracy of Garmin GPS receivers (which will be used in the project) is typically within 10 meters, thus making it highly reliable. On the water, accuracy is even better. The biggest names in GPS industries are Google, Garmin, Magellan, OnStar and TomTom.

Garmin is a well-known navigation system. They're very popular because of their application in various fields suggest and guide your movements as you travel. You can set your car to pre-determined destination and tailor your route. Their horizon is expanded to Aviation, Marine, Fitness, Treking, etc.

Garmin GPS uses **GPX file format** to store its data. A GPX file is a GPS data file recorded in the GPS Exchange format, which is a widely used open standard. It includes waypoints, routes, and tracks, as well as longitude and latitude location data. GPX files are saved in XML format, which makes it easier to import and read GPS data from a variety of programs and web services. GPX Stores **A waypoint** which contains a point's GPS coordinates; **A route,** which is a set of track points that lead to a destination. Track points are waypoints for turn or stage points and **Track** which is a path is described by a list of points.

II. LITERATURE REVIEW

The first use of surveying can be dated back to 1400BC when early Egyptians used the precision for the division of lands into plots. Furthermore, in 1200BC, the innovation of geometry was introduced by the Greeks, which helped in the accurate division of land and standardize the procedures for conducting surveys. They also have the distinction of being the first to develop an instrument (Diopter) for conducting surveys. Surveying became more important in the 1800s for the development of public infrastructure such as canals, highways, and railroads. The Industrialization ushered in the development of geodetic and plane science, which in turn ushered in the creation of increasingly complex equipment. [1]

As rightly pointed out by researchers, A complete physical or manual land survey is the only efficient and reliable way of documenting the physical boundaries of the property and identifying the improvements on it. Land surveys, on the other hand, are frequently disregarded and are one of the most misunderstood aspects of a real estate transaction. A surveyor's plotted area is frequently of any irregular shape. [1]

The researchers have identified the existing technology of surveying i.e. Chain Surveying, Compass Surveying, Theodolite Surveying and Photographic surveying and how weather and seasonal changes affect the same. [1]

In the traditional method, land surveying consists of ten stages that take more than two weeks to complete in order to register a property with the appropriate offices. Many issues over ownership, land boundaries, and locale eventually develop. As a result, the issue has been handled by using a GPS. [2]

The ownership of land is one of the oldest and most essential aspects of human society, and adastral surveys deal with it. They are the surveys that establish, mark, define, retrace, or rebuild the public lands' boundaries and subdivisions.[3]

Traditional land use planning is unable to match the demands of the times, and planning theory and practise are showing signs of deterioration. As a result, land use planning should continue to incorporate the most recent scientific and technological advances, as well as develop and improve land use planning concepts and methodologies. [4]

Land is the most essential resource for national development and a major regulatory mechanism for the macro economy. Land management is linked to a country's economic, food, and ecological security, and is essential for the country's healthy and long-term development. The current land use status in the country as a whole is unclear, and land use management levels in different provinces are inconsistent. [5]

A Geographic Information System (GIS) is a specialised system for managing and analysing geographical data. It can handle all types of geographical data according to geographic coordinates and locations and explore an object's spatial relationships with the help of hardware. [6]

The Digital Land Management System, or DLMS for short, is a model system for digitally managing land regions in any country, such as India, using current technology. [7]

Crop field borders have been created on the cadastral map for centuries through land surveying. The recent development of survey tools such as Global Positioning System (GPS) devices, theodolites, and total stations has enhanced the mapping of field borders. [8]

ISSN 2394 - 7780

Investigators collect traditional land parcel field data using GPS positioning and record it in a standard record chart that includes geodetic coordinates for the four boundaries of one sampling land parcel as well as other data. Furthermore, the data from the record chart was entered into the office computer, and the space vector and characteristics data were processed to obtain land parcel data. The article uses network technology, database technology, and system integration technology to construct a fully digital multi-field data collection, processing, and data management technology system in light of our urgent need for stringent land management. [9][10]

III. METHODOLOGY

As mentioned earlier, there are several methods and equipments that can be used for Land Boundary and Land Surveying. However, they may be unfeasible and complex. So, a simpler and suitable method which can be adopted for Land Boundary Surveying is using GNSS Device.

Considering that GPS is the most popular GNSS, and Garmin is one of the most widely used GPS handheld devices, the same shall be used for the Study.

The hardware used to initiate the procedure will include a Garmin Handheld GPS, a Laptop/Desktop and USB to GPS Cable. Softwares used will include the OEM software, Garmin Basecamp, GIS software (QGIS, ArcGIS, etc.) and Python IDE (in this Study, Spyder IDE shall be used)

A. Data Collection and Exploration

The gathering of the Data involves using the Garmin GPS. IN our example we shall use a Garmin Oregon 550 handheld GPS.

Firstly, start the Garmin GPS. Wait till the satellites are acquired. The range icon in the bottom centre shall point out once the GPS signals are caught. When the icon turns green, it implies that the range has been acquired.

On acquiring the signal, start by marking waypoint by clicking on ‘mark waypoint’ icon, and then naming your waypoint and saving the same.

Now simply walk/drive with your GPS device around the plot. Mark Waypoints at extreme ends if necessary. Make sure you come back at the same spot. AS soon as you complete a revolution turn off the GPS device.

Now connect the GPS to the laptop. Navigate to the ‘GPX’ folder and search for current.gpx file. The last gpxfile is saved as current.gpx. Copy that GPX file at the desired location.

GPX files are GPS files that store information such as Latitude, Longitude, Speed, Elevation, Tracks etc. in XML Format. The format is shown below.


<?xml version="1.0" encoding="UTF-8" standalone="no" ?><gpx xmlns="http://www.topografix.com/GPX/1/1" xmlns:gpxx="http://www.garmin.com/xmlschemas/GpxExtensions/v3" xmlns:wptx1="http://www.garmin.com/xmlschemas/WaypointExtension/v1" xmlns:gpxtpx="http://www.garmin.com/xmlschemas/TrackPointExtension/v1" creator="Oregon 550" version="1.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.topografix.com/GPX/1/1 http://www.topografix.com/GPX/1/1/gpx.xsd http://www.garmin.com/xmlschemas/GpxExtensions/v3 http://www8.garmin.com/xmlschemas/GpxExtensionsv3.xsd http://www.garmin.com/xmlschemas/WaypointExtension/v1 http://www8.garmin.com/xmlschemas/WaypointExtensionv1.xsd http://www.garmin.com/xmlschemas/TrackPointExtension/v1 http://www.garmin.com/xmlschemas/TrackPointExtensionv1.xsd"><metadata><link href="http://www.garmin.com"><text>Garmin International</text></link><time>2022-01-07T15:21:39Z</time></metadata><trk><name>Current Track: 05 JAN 2022 23:14</name><extensions><gpxx:TrackExtension><gpxx:DisplayColor>Black</gpxx:DisplayColor></gpxx:TrackExtension></extensions><trkseg><trkpt lat="19.0941458544" lon="72.8292098845"><ele>24.65</ele><time>2022-01-05T17:44:56Z</time></trkpt><trkpt lat="19.0941384295" lon="72.8292116895"><ele>24.65</ele><time>2022-01-05T17:44:57Z</time></trkpt><trkpt lat="19.0941088647" lon="72.8292016312"><ele>21.29</ele><time>2022-01-05T17:45:02Z</time></trkpt><trkpt lat="19.0940795280" lon="72.8291622363"><ele>21.77</ele><time>2022-01-05T17:45:37</time></trkpt><trkpt lat="19.0940887481" lon="72.8291445505"><ele>21.77</ele><time>2022-01-05T17:45:45Z</time></trkpt><trkpt lat="19.0940248780" lon="72.8289313149"><ele>20.81</ele><time>2022-01-05T17:46:05Z</time></trkpt><trkpt lat="19.0938921925" lon="72.8288568836"><ele>21.29</ele><time>2022-01-05T17:46:28Z</time></trkpt><trkpt lat="19.0937864967" lon="72.8288466576"><ele>21.29</

ISSN 2394 - 7780

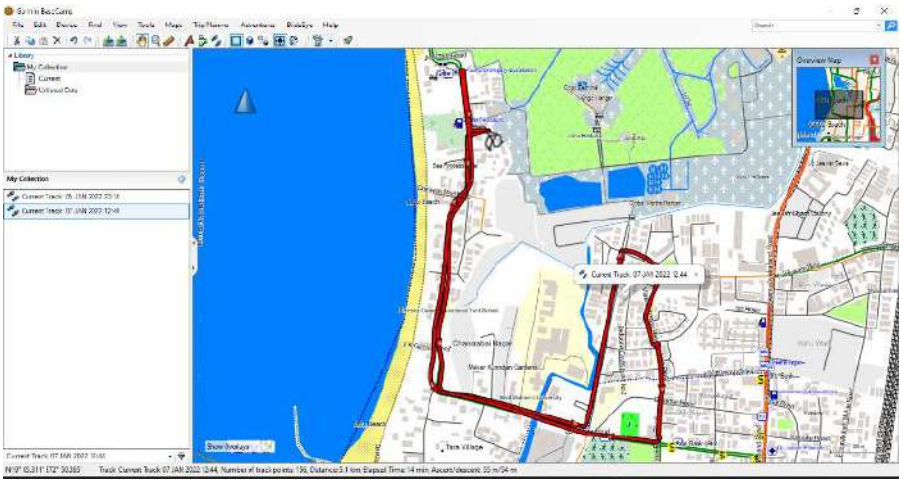
B. Analysis of Data

After extracting the GPX file from the GPS and storing it in a desired location, the analysis can be carried out on it. There are 4 approaches that can be used for analysis.

Approach 1- Basic Approach

The **first approach** that can be adapted is by using **Garmin's basecamp application**. This is a desktop application provided by the equipment provider, Garmin. To use this, open basecamp on laptop, click on maps and select OSM Maps. OSM maps are the free open-source maps that are made available to the public.

After enabling OSM, click on file, and then import into "My Collection". Navigate to the GPX file and add it. After importing the GPX, select on the relevant time stamp and it will display the area on the map



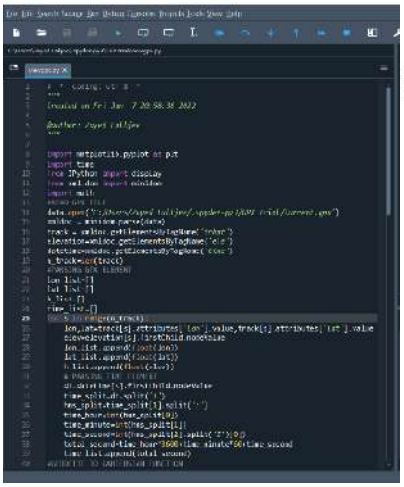
The red highlighted region is the path followed which marks the land boundary and will help land surveying.

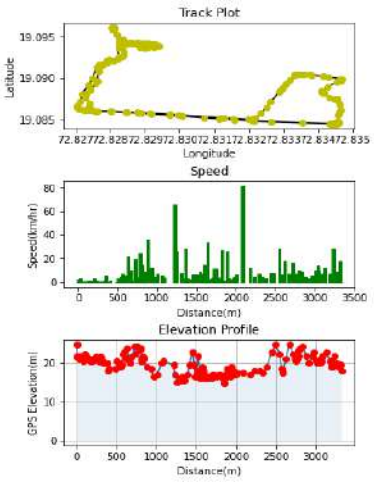
Approach 2- Programmers Approach:

The Second approach is using **Python Programming language.** Using Python programming language, user specific data can be called from the GPS and can be displayed. This approach can be used by prgrammers to specifically develop software to suit their needs.

Some modules like matplotlib, time, IPython, xml dom and math shall be used for the same. To open the GPX file and parsing the same, Minidom is used. After the calculation speed, distance and elevation using various functions, plot them on a graph.

The Output of the python code will be as follows:



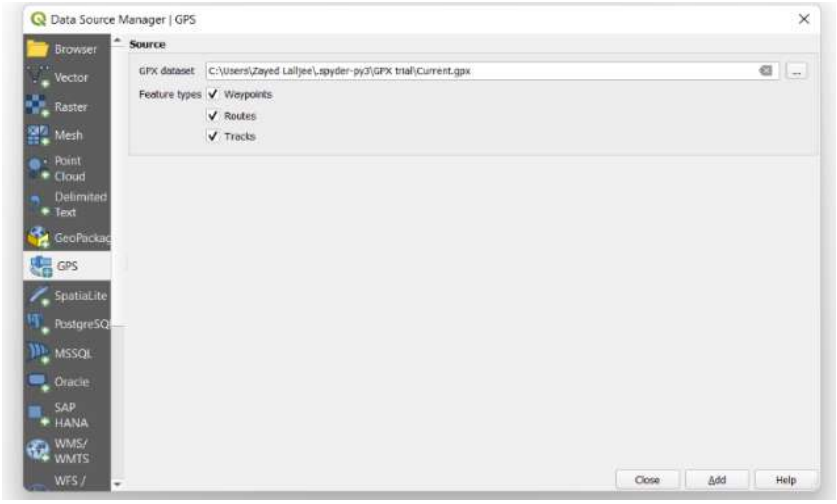


Approach 3- Surveyors Approach

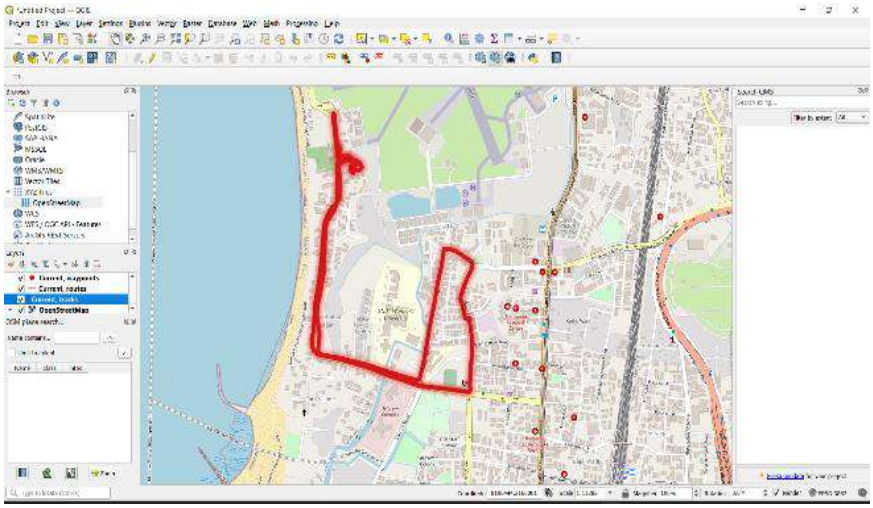
The third approach can be adapted by a **GIS Analyst**. This will aid his analysis on for land. For this procedure, one requires a GIS software such as ArcGIS, one of the most popular GIS tools, QGIS, etc. In this case, we will use **QGIS**.

Start by creating a new project, click on Layer-> Add Layer -> Add Vector Layer. Select the GPS option. Navigate to your gpx dataset and click add.

ISSN 2394 - 7780



Add OpenStreetMap to the QGIS Map after installing OSM Plugins. Add to current layer The output will be as follows:

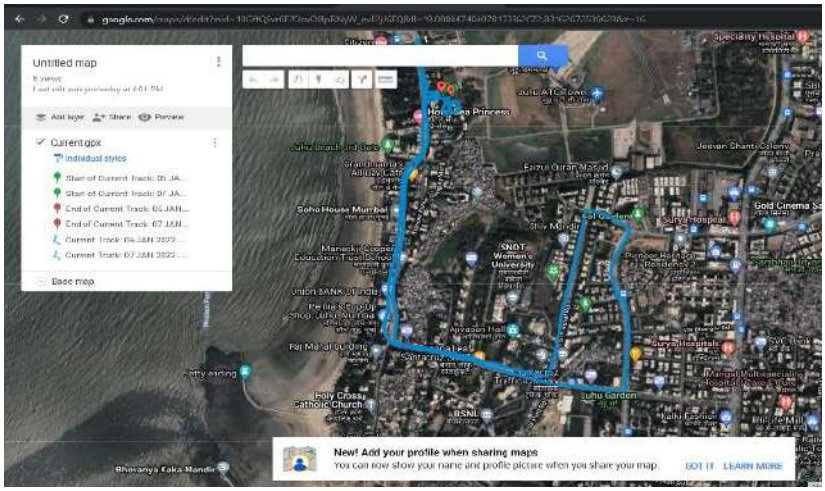


The red highlighted region is the path followed, which marks the land boundary and will help land surveying.

Approach 4- Simplified Approach:

The Fourth approach can be achieved by using **Google My Maps**. It is the simplest method, and this can be achieved by a common man.

For this, user has to log onto https:\\mympas.google.com, click on create new map and then clicking on import. Add the gpx file and select the type of map. In this case Satellite images map was used.



ISSN 2394 - 7780

C. Comparison between the approaches
The comparison of the four methods is as follows:

	Method 1	Method 2	Method 3	Method 4
Who Uses it	GPS Users	GIS Analysts	Programmers	Layman
Advantages	Easy to understand, Intermediate Analysis, OEM provided and simultaneous up dation of GPS	Advanced Analytics can be performed especially by GIS Analyst	Tailor Made softwares that address specific needs can be developed	Basic analysis, easy availability and not much infrastructure required
Infrastructure	Computer with basic processing speed and GPS	Computer with advanced processing speeds	Computer with advanced process speeds and IDE.	Latest Web Browser and good Internet connection
Importing of Maps	Only specific maps can be imported and read	A wide variety of maps can be imported and read	Maps protected by IPR, can be afforded. Hence, scope for professionals to find this more dependable.	Various types of maps preloaded for example Google and Ma

IV. CONCLUSION

As technology improves, it reduces the chances of Jamming and degradation of signals and other means hacking of the signals. The use of GPS for land surveying is a cost friendly and accurate method for defining boundaries. The use of heavy infrastructure is not needed. Garmin GPS has around 10 metre accuracy and with the induction of the Indian Satellites, once can expect submeter accuracy, The use of DGPS and other equipment can be used in places where receiving the GPS signal may be an issue and where GPS signals are weak, additional equipments may be needed to improve the precision. The trilateration of GPS has been improved over time making it much more accurate. Now with the new concept of multi-GNSS introduced, and the number of GPS Satellites increased and with the Indian constellation, including GAGAN, making a serious inroad into the accuracy of the GPS.

GPS Surveying is ideal for areas who have irregular boundaries. With the use of waypoints, the accurate position of the outermost position can be calculated. Now with the use of satellite imagery, the tracklogs obtained from the GPS can be superimposed thus providing a visual representation.

GPS can be very effectively deployed in situations of natural calamities such as floods, and hurricanes to estimate the extent of impact. This can be done with the least amount of heavy infrastructure and permits post processing of the data. It will have even more applications in the future.

REFERENCES

[1] R. S. Abhi Krishna and S. Ashok, "Automated Land Area Estimation for Surveying Applications," 2020 International Conference for Emerging Technology (INCET), 2020, pp. 1-5, doi: 10.1109/INCET49848.2020.9154042.

[2] R. Gowtham, J. Alfred Daniel, S. Karthik and R. C. Vignesh, "Revamping Digital Land Survey using GPS and Internet of Things (IoT)," 2018 International Conference on Soft-computing and Network Security (ICSNS), 2018, pp. 1-9, doi: 10.1109/ICSNS.2018.8573672.

[3] R. Ghazali, Z. Latif, A. R. Rasam and A. M. Samad, "Integrating Cadastral GIS Database into GPS Navigation System for Locating Land Parcel Location in cadastral surveying," 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, 2011, pp. 469-473, doi: 10.1109/CSPA.2011.5759924.

[4] L. Chang, M. Yang and R. Song, ""Anti-Planning" Methodology and the Application on the Plans for the Land Utilization," 2010 International Conference on Multimedia Technology, 2010, pp. 1-3, doi: 10.1109/ICMULT.2010.5629694.

[5] G. Yang and G. Qiao, "Data Processing Method for Current Land Use Using GIS Technology," 2010 Second International Workshop on Education Technology and Computer Science, 2010, pp. 511-514, doi: 10.1109/ETCS.2010.336.

[6] Z. He-bing and Z. Su-xia, "Study on Urban Land Information System Based on GIS," 2009 International Forum on Information Technology and Applications, 2009, pp. 672-674, doi: 10.1109/IFITA.2009.291.

ISSN 2394 - 7780

[7] S. K. Talukder, M. I. I. Sakib and M. M. Rahman, "Digital land management system: A new initiative for Bangladesh," 2014 International Conference on Electrical Engineering and Information & Communication Technology, 2014, pp. 1-6, doi: 10.1109/ICEEICT.2014.6919031.

[8] M. S. Rahman et al., "Crop Field Boundary Delineation using Historical Crop Rotation Pattern," 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), 2019, pp. 1-5, doi: 10.1109/Agro-Geoinformatics.2019.8820240.

[9] Zongjian Lin, Chaofeng Ren and Na Yao, "Precison land survey system and application technology," 2011 International Conference on Remote Sensing, Environment and Transportation Engineering, 2011, pp. 4261-4263, doi: 10.1109/RSETE.2011.5965271.

[10] H. Zhang and S. Zhao, "Development and Application of Land-Use Planning Management Information System Based on ArcGIS," 2010 International Forum on Information Technology and Applications, 2010, pp. 64-67, doi: 10.1109/IFITA.2010.65.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

CAPTURE THE UNSEEN - THE ROLE OF IOT IN SECURITY SYSTEM

[1]Prashant Chaudhary and [1]Manpreet Kaur Matta
[1]Assistant Professor and [2]M.Sc.IT Student, UPG College

ABSTRACT

The Internet of Things (IoT) is an intensifying tendency in today's world and has proved to be a game-changer in the field of technology. The technology or belief for this paperis IOT based which is a thing of present and future. The primary purpose of this paper is to make our homes smarter and more secure. In today's world, where every person wants to maintain their valuables safe and secure, video surveillance for observing a particular area has become the need of the hour. To deal with this difficulty, the safest solution is a smart surveillance system for certain places like bank vaults, homes where the human presence is not available.

Based on Arduino, a security system simulation ispresented in this paper. The consequences are known from the glow of LED's connected to the board. It is a distinctive way of keeping the system secure with current technology. This paper researches the building blocks of IOT In Security systems, challenges in adopting it, Models to improve practices, and Future works on Security systems related to IOT.

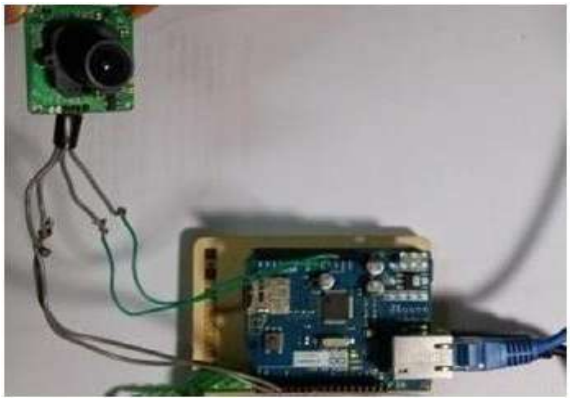
Keywords: Internet of Things, Arduino, Security system,Surveillance.

I. INTRODUCTION

The Internet of Things is ultimately described as anetwork of physical objects. Networks before used tobe just computers linked together, but nowadays it has evolved into a more advanced network. Nowadays most devices are connected to the Internet; moreover, these devices collect data through sensors. These devices can range from smallto big, such as smart fridges or cars. The problem here is that all these devices collect any data from their sensors, but don't have a shared platform to share it. The Internet of Things aims to combine all their data and connect all these devices to form a huge network of things, where they can communicate together and analyse patterns. "The internet of thingsis an internet of three things, people to people, machine to people and machine to machine" (Salazar,C.,2016). devices together to be able to achieve new heights in technology and provide new services and applications. There is an increasing thrust on reducing manpower and human labour to give way to machine efforts and digitization. In this regard, home automation is perhaps the most populararea of research, as it offers a wide variety of use casesand avenues for introducing machine learning. The market is flooded with different kinds of devices thatenable home automation. While most of them focus on appliance control through apps, or more recently, voice control, there is a lot more that can be accomplished through automation. The internet of things aims to accomplish the unity of multiple Security at living species has become necessary for current lifestyles, asmonitoring and surveillance 24x7 is difficult to maintain manually by a person.

The latest technology of IoT applications is one of the best solutions. By using IoT we can have access to information about security threats, damage alerts, danger alerts and additional controls over home appliances for convenience and in automation and home surveillance.

Arduino is an electronic board of uses micro controllers. It is programmed with a hex code. Integrated Development Environment (IDE) is software running on a 'C' program or assembly program. In our daily life, securing anything is made the main concern. Security is inevitable and wanted by everyone. In a security chain, the well manageablecontrol system is formed as a vital link [1]. Arduino UNO based security system by using a PIR sensor is proposed in this paper.



ISSN 2394 - 7780

II. LITERATURE SURVEY

The paper "Capture the Unseen - The Role of IOT in Security System" extracts useful data from various research papers that examined smart home safety and security systems using the Arduino platform. This paper gives us a deep insight into the types of motion sensors and the frequency of their use in security systems and talks about the various types of alert mediums used in these systems. Further, this paper also details the use of various Arduino boards in thesesystems, while also comparing their architectures. This data has then been analyzed to extract useful information regarding the advantages and disadvantages of these systems and how future research could enable better implementation and use of these systems.

IIA. COMPARISON OF COMPONENT

1. Arduino UNO R3 and Raspberry Pi 3B

Arduino UNO R3	**Raspberry Pi 3B**
The control unit ofArduino UNO R3 is From the Atmega family.	The control unit ofRaspberry Pi 3B is from the ARM family.
Arduino is a microcontroller.	Raspberry Pi is amicroprocessor.
It has an 8-bit CPU.	It has an 64-bit CPU
RAM 2KB.	RAM 1GB.
It is cheap in the market.	It is more expensive in the market.
Power consumption is less.	Power consumption is more.
There is no USB support.	There is 4 USB support available.
It can store dataonboard.	Raspberry Pi 3B cannotstore data onboard, it is providedSD card.
Arduino UNO R3 usesC/C++.	Raspberry Pi 3B usesPython, C/C++ and Ruby.

2. **PIR Sensor and Microwave Sensor**

PIR Sensor	**Microwave Sensor**
PIR sensor works onInfrared signals.	Microwave sensors work on Dopplersignals.
It operates on heatdetection.	It operates on a frequency band of 3.2 GHz.
Angle covered by PIRsensors is 110 degrees	Angle covered by Microwave sensors is360 degrees but not accurate
3-7 m range, PIR sensors is good in detecting shorter range.	3-7 m range, not good.
Power consumption is 4-28 volts.	Power consumption is 4.2-12 volts.

3. ESP8266 and ESP32 Camera Module

ESP8266	**ESP32 camera Module**
Released in 2014.	Released in 2016.
Xtense single-core 32-bit L106 microcontroller.	Xtense single/dual-core32-bit Lx6 microcontroller.
The clock frequency is 80 MHz	The clock frequency is 160/240 MHz
It supports 160 kb SRAM.	It supports 520 kb SRAM.
WIFI available (802.11.b/g/n).	WIFI available (802.11.b/g/n).
No Bluetooth support.	Bluetooth support is available (4.2 version).
There is no camera init.	There is a camera on it.
SD card slot is not there.	There is an SD card slot.
Power consumption is20uA.	Power Consumption is less than ESP8266 10uA (deep sleep).

4. Arduino Boards

Sr. No	Board Name	Processor Type	Clock Speed	Digital I/O	Analog Inputs	PWM
1	Uno	ATmega328P	16 MHZ	14	6	6
2	Mega	ATmega2560	16 MHZ	54	16	15
3	Due	ATSAM3X8E	84 MHZ	54	12	12
4	Yun	ATmega32U4	16 MHZ	20	12	7
		AR9331 Linux	400 MHZ			
5	Nano	ATmega168	16 MHZ	14	8	6

ISSN 2394 - 7780

		ATmega328P				
6	Leonardo	ATmega32U4	16 MHZ	20	12	7

I. SOFTWARE AND HARDWAREREQUIREMNETS

Software Requirements:

- Arduino (IDE)
- Telegram App

Hardware Requirements

- Arduino UNO
- PIR sensor
- Buzzer
- Relay Module
- Bulb
- Esp32 Camera Module
- Breadboard
- Jumper Wires
- TTL programmer
- Power Supply

II. System Architecture

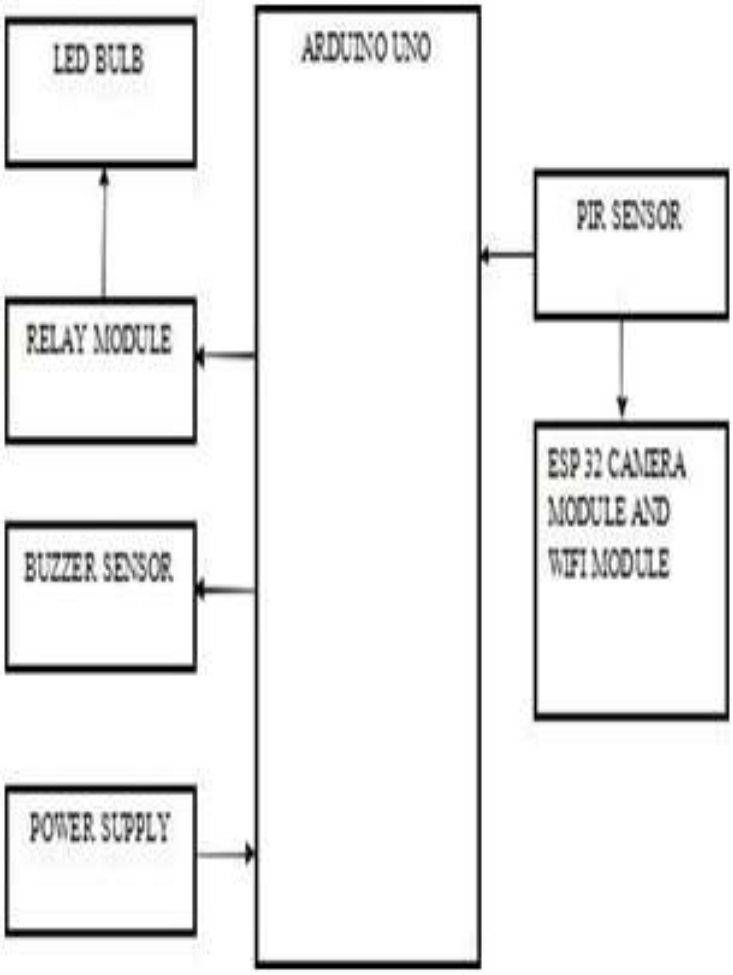


Fig 1. Block Diagram

ISSN 2394 - 7780

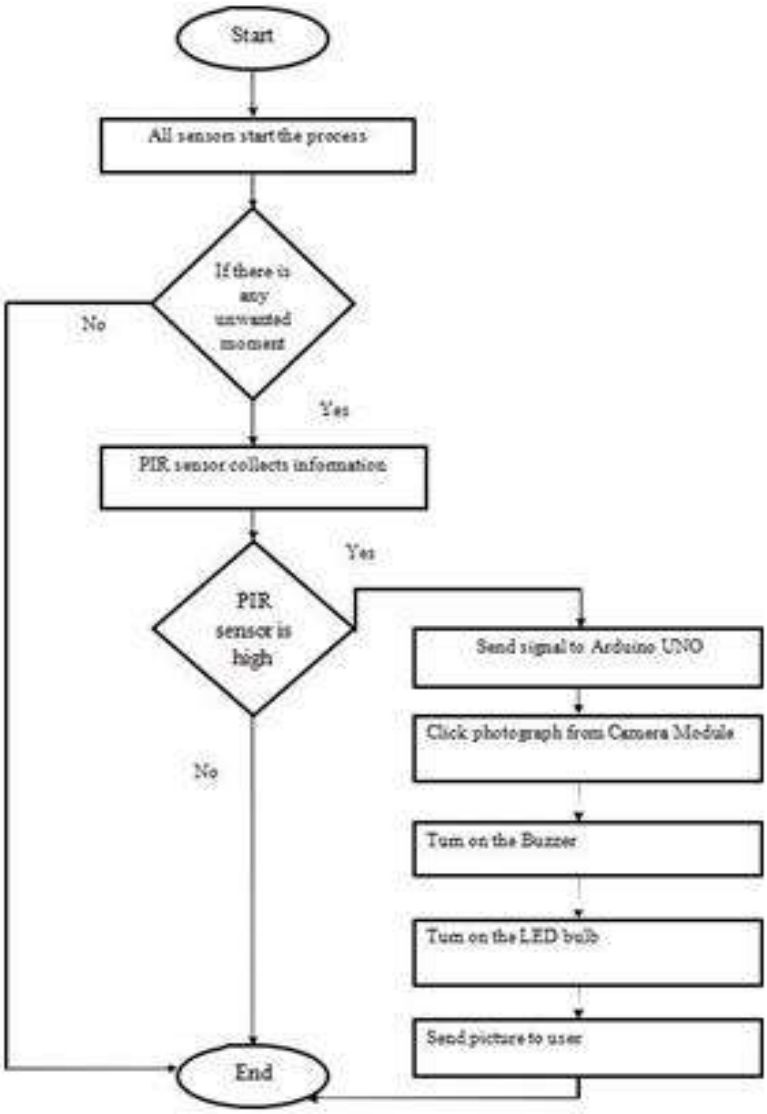


Fig 2. Flow Chart

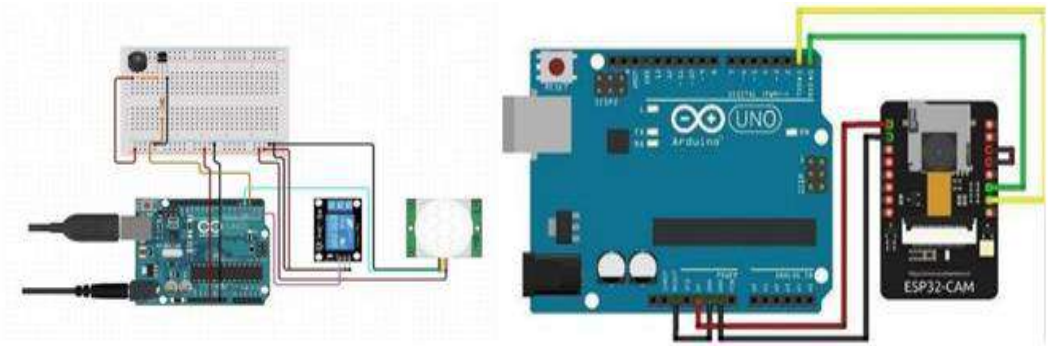


Fig 3.A **Fig 3**.B

Fig 3A & B. Circuit Diagram

III. APPLICABILITY

- This device can be used by anyone as per their needs.
- As the device work with less power, the user does not need to worry about power consumption.
- The device has a normal look because of that the intruder won't be able to recognize it easily.
- It has minimal upkeep and user friendly.
- Modern security systems now allow you to remotely monitor what's happening in your place from your phone when you're not there.
- Adding a security system to your home gives you an extra layer of protection from the thief/intruder.

ISSN 2394 - 7780

IV. CONCLUSION

This IOT based device surveillance and control system are solely used to keep surveillance on the electric devices operating condition and to controlthe on/off functionality from a central remote location. The proposed system works effectively for both indoor and outdoor lighting. On the one hand, it augments the effectiveness of the system by transmitting an alert signal in situations of anydefect and on the other hand it severely diminishes the electric energy utilization by providing central control over the appliances. Thegraphical App-based mobile operating delivers a user-friendly and effortlessly accessible platform to the user. This system can be installed as an energy-efficient system to control streetlamp that requires a lot of energy and needs manual intervention.

V. FUTURE SCOPE

As this is IOT Based system it can be furtherused to enhance to monitor the complete traffic system and add function, componentslike:

- Reading Number plates of vehicles: -MATLAB or Open CV can be used to further enhance this system to automatically read Number Plates of Vehicles.
- Depending upon the requirement of using multiple cameras and PIR sensors can be added.
- Challan the vehicles for over speeding: - In case of a traffic violation or over speeding, Challan can be automatically issued via camera monitoring and recording clips can be saved forfuture reference.
- Trespasser's detection: - Trespassers can be traced if found guilty.
- Real-time implementation of sensors to plan and employ HMIS [Health Care Management Information System]
- Live video feedback to traffic control centre: - Live video streaming can be screened on to theapplication to a central monitoring team.

VI. REFERENCES

[1] www.educba.com

[2] https://en.wikipedia.org/wikilInternet_of_Things.

[3] https://a.arduino.cc/hc/en-us

[4] http://playground.arduino.cc/.

[5] http://www.cisco.com/web/solutions/trends/iot/ov erview.htmI.

[6] Honeywell Automation: https://homesecurity.honeywell.com

[7] T. Ming Zhao, Chua, "Automatic face and gesture recognition, 2008. fg '08. 8th ieee international conference on," pp. 1–6, September2008.

[8] https://www.cleverism.com/smart-home-intelligent-home-automation/

[9] Arka Das, O.V. Gnana Swathika, Dr. S. Hemamalini, "Arduino Based Dual Axis Sun Tracking System", American Scientific Publishers, 2015.

[10] K Vinay sagar and S Kusuma, "Home Automation Using Internet of Things", International Research Journal of Engineering and Technology, vol. 02, no. 03, June2015.

[11] R. Piyare and M. Tazil, "Bluetooth based home automation system using cell phone", Consumer Electronics (ISCE) 2011 IEEE15th International Symposium on, pp. 192-195, 2011.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

DIAGNOSIS OF PNEUMONIA, TB AND COVID THROUGH MACHINE LEARNING AND DEEP LEARNING ALGORITHMS: A REVIEW

[1]Sunita Gupta and [2]Mohammed Raihan Makba
[1]Assistant Professor and [2]MScIT Student, UPG College of Arts, Science and Commerce, University of Mumbai, India

ABSTRACT
Machine learning & Deep learning are often two branches/fields which have been significantly researched over the last decade under the umbrella term-Artificial intelligence. Leveraging these applications of AI in the field of classification and data extraction, this paper emphasises on how machine learning and deep learning algorithms can be applied together with medical imaging for the purpose of prediction and diagnosis, specifically, with the case of Pneumonia, TB and Covid as their medical interpretation is challenging within short period of time. This paper further talks about how machine learning algorithm like neural networks can be used to replicate the behaviour of human brain and how that could outperform in diagnosis, detection and predictions. Further, it briefs about the relation between AI-ML-DL and covers about the basics of different algorithms and learning types which can be applied with medical imaging depending upon the need and further consolidates future scope and improvements.

Keywords: *Machine Learning, Deep Learning, Neural Networks, Radiology, Diagnosis, Image Processing, Artificial Intelligence.*

I. INTRODUCTION

Artificial intelligence has dramatically impacted the efficiencies of our workplaces by augmenting human capabilities. The use of AI technologies in healthcare and medical industry has the ability to potentially assist the healthcare providers in one of the many aspects like, patient care, and administrative processes, etc. Most healthcare and AI technologies have strong relevance to the healthcare field, with varying underlying methodologies. The use of AI in medical imaging can perform just as well or better than humans at certain procedures in diagnosing disease from previous learning/training data sets.

What is Machine Learning?

Machine learning is a branch of artificial intelligence (AI) which focuses on the use of - data and algorithms to imitate the way that humans learn and thereby gradually improving its accuracy. In technical terms, it provides the systems with the ability to learn (automatically), and to improve from the experiences without being explicitly programmed, where the learning phase comprises of methods that automatically detect patterns in the training data set, and utilize the uncovered patterns to predict future data or enable decision making.

What is Deep Learning?

Deep learning is a subset of machine learning which consists of single/multi-layer neural network (three or more layers). These neural networks maker an effort to simulate the behaviour of the human brain and hence allowing it to "learn" from large data sets. A neural network with a single layer is capable of approximating predictions, however, with additional hidden layers, the algorithm can be further optimized and refined for accuracy.

Relation between AI - ML - DL

AI is an umbrella term which covers everything related to making machines smarter, and machine learning is a subset of AI which provides the systems with the ability to learn (automatically), and to improve from the experiences without being explicitly programmed. By the definition, ML refers to an AI system that can self-learn based on the algorithm (supervised/unsupervised) and hence the systems get smarter, and smarter over a period of time without human intervention is ML. Deep Learning (DL) is a subset of machine learning (ML) which usually deals with large data sets with multiple hidden layers forming a neural network thereby imitating the behaviour of human brain.

ISSN 2394 - 7780

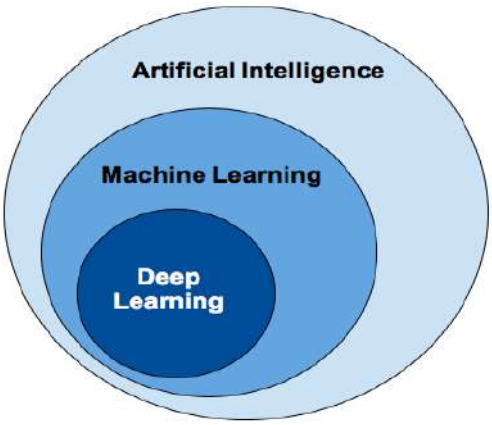


Fig. 1: Relation between AI -ML -DL [3]

In reference to the above Fig.1, Artificial intelligence is a large set of different algorithms with distinct abilities, out of which the major distinctive domains being machine learning and deep learning. However, since AI being vast, in order to imitate human intelligence in a machine, different sets of machine learning algorithms are being used (to achieve artificial intelligence), and different while deep learning techniques are used to implement machine learning [3].

Confining the scope of this paper, machine learning algorithms can further be classified into 3 types:
1 - Supervised Learning

2 - Unsupervised Learning

3 - Reinforcement learning

Supervised learning
Supervised learning, as the word emphasises 'Supervised', has the presence of a supervisor which directs the behaviour of the machine by providing a set of labelled data.

Unsupervised learning
Unsupervised learning is a type of learning where the machines consume data which is neither classified nor labelled and allows itself (algorithm) to act on information without guidance. The machine groups unclassified /unlabelled /unsorted data ap per their observed similarities, patterns, and differences without any prior training.

Reinforcement learning -
Reinforcement Learning is feedback-based learning technique in which an agent learns to behave in an environment by evaluating its own actions which are rated depending upon the behavioural execution. The agent gets positive feedback/score for each right action negative feedback/penalty for each wrong action.

II. MACHINE LEARNING IN MEDICAL IMAGING

As discussed earlier that machine learning is a technique for recognizing patterns, this when combined with medical imaging/radiology helps in early diagnosis and predictions. It typically begins with ML algorithm computing image features which are believed to be important for specific diagnosis. The system (algorithm) then identifies the best combination of these image features for classifying the image or computing some metric for the given image [4].

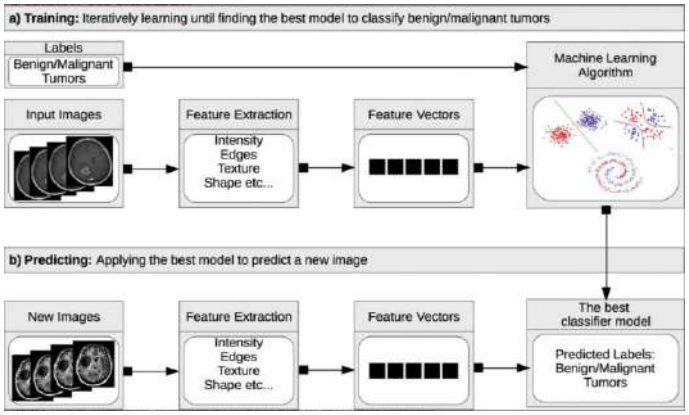


Fig. 2: ML in medical imaging [4]

ISSN 2394 - 7780

In fig 2, For training, the machine learning algorithm uses a set of input images to identify the image properties thereby creating a knowledge base which would be further used for the correct classification of test data - that is, identification/prediction of diseases - as compared with the supplied labels for these input images. (b) Once the system has learned how to classify images, it can be used for its purpose predictions and early diagnosis. This trained model is further applied to new set of images to further assist and increase the accuracy of predictions/diagnosis.

III. LITERATURE REVIEW

Several research articles and papers have been reviewed to obtain background knowledge on this problem and to explore what has been done already. In 2019, [1] a study has been conducted where the use of artificial neural networks as one of the machine learning algorithms was developed to go through the feature extraction process, however ANN suffered from overfitting and vanishing gradient problems for training deep networks. One of the key tasks for radiologist was to differentiate appropriate data for diagnosis and to feed in for image classification, followed by object detection and categorizing it.

Another study conducted in 2018, [2] mentions about the applications of ML and image processing so that MRI images can be further graphically enhanced and deep study can be conducted. This study had explained about the neural networks which consists of a number of connected computational units, called neurons, arranged in layers. There's an input layer where data enters the network, followed by one or more hidden layers transforming the data as it flows through, before ending at an output layer that produces the neural network's predictions. The network is trained to output useful predictions by identifying patterns in a set of labelled training data, fed through the network while the outputs are compared with the actual labels by an objective function. During training the network's parameters – the strength of each neuron – is tuned until the patterns identified by the network result in good predictions for the training data. Once the patterns are learned, the network can be used to make predictions on new, unseen data, i.e., generalize to new data.

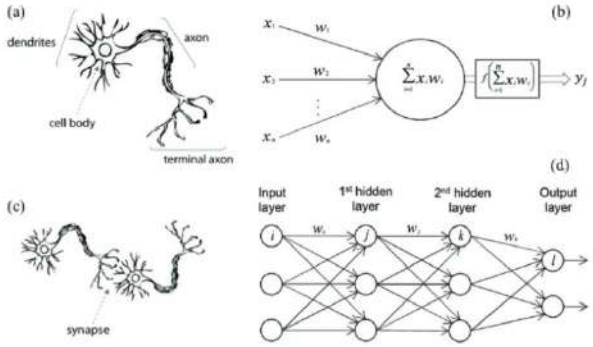


Fig. 3: ANN – Artificial neural network (a) human neuron; (b) artificial neuron; (c) biological synapse; and (d) ANN synapses [3]

Fig 3, briefs and compares ANN with human nephron.

A study conducted by Beijing University of Technology in 2019 [3], briefs about the different ML algorithms and how are they classified. The diagram depicting the different stages of ML and DL which are used in medical imaging domain is shown in fig 4.

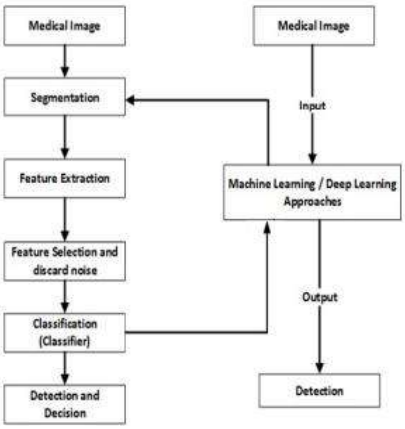


Fig. 4: ML and DL workflow [3]

ISSN 2394 - 7780

The authors have further explained the process of segmentation followed by feature extraction for specific diagnosis and discarding of noise. These features are further classified with the help of classifier and the whole process is carried for fewer iterations/cycles till the features distinctly identified. Parallelly, a test data is provided as input which is further analysed based on the previous knowledge or training data set and then appropriate predictions/detections are made.

IV. CONCLUSION

Machine learning field has been evolving over the past few years and is currently being used by many industries for demand-supply, stock predictions, sentiment analysis, etc. However, medical imaging is still climbing up the ladder where radiology could be integrated with deep learning to pre-diagnose the early symptoms of diseases, identify fractures, tumours, etc. This research ensures significant reduction of diagnosis turnaround time and provides better care to patients. In medical imaging, machine/deep learning algorithms helps to categorize, classify, and enumerate disease patterns which further extends the analytical goals and generates strong prediction model. The current existing models are able to work upon single set of data and diagnose one specific disease/characteristic at a time.

FUTURE ENHANCEMENTS

The future scope of this research is to develop a model which is trained on vast data set comprising of x-rays related to Pneumonia, TB and Covid. This model should be able to distinguish and classify the input (patients) Xray into one of the three mentioned diseases so that the early symptoms could be identified and treated accordingly, thereby reducing the turnaround time.

REFERENCES

[1] Mingyu Kim, Jihye Yun, Yongwon Cho, Keewon Shin,Ryoungwoo Jang, Hyun-jin Bae, and Namkug Kim1, Deep Learning in Medical Imaging., December 2019

[2] Alexander Selvikvåg Lundervold, Arvid Lundervold, An overview of deep learning in medical imaging focusing on MRI

[3] Jahanzaib Latif, Chuangbai Xiao, Azhar Imran, Shanshan Tu, Medical Imaging using Machine Learning and Deep Learning Algorithms: A Review

[4] Bradley J Erickson, Panagiotis Korfiatis, Machine Learning for Medical Imaging

[5] Chuansheng Zheng, Xianbo Deng, Qiang Fu, Deep Learning-based Detection for COVID-19 from Chest CT using Weak Label

[6] Aiken, A., Moss: A system for detecting software plagiarism. http://www/.cs.berkeley.edu/~aiken/moss.html, 2004.

[7] Doi, K., Computer-aided diagnosis in medical imaging: historical review, current status and future potential. Computerized medical imaging and graphics, 2007. 31(4- 5): p. 198-211.

[8] Mahesh, M., Fundamentals of medical imaging. Medical Physics, 2011. 38(3): p. 1735-1735.

[9] Jannin, P., C. Grova, and C.R. Maurer, Model for defining and reporting reference-based validation protocols in medical image processing. International Journal of Computer Assisted Radiology and Surgery, 2006. 1(2): p. 63-73.

[10] Michalski, R. S., Carbonell, J. G., & Mitchell, T. M. (Eds.). (2013). Machine learning: An artificial intelligence approach. Springer Science & Business Media.

[11] Norris, D.J., Machine Learning: Deep Learning, in Beginning Artificial Intelligence with the Raspberry Pi. 2017, Springer. p. 211-247.

[12] Jankowski, N. and M. Grochowski. Comparison of instances seletion algorithms. Algorithms survey. In International conference on artificial intelligence and soft computing. 2004. Springer.

[13] Schmidhuber, J., Deep learning in neural networks: An overview. Neural networks, 2015. 61: p. 85-117.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

DATA ANALYTICS FOR ACHIEVING IMPROVEMENT IN SMALL BUSINESS

[1]Neelam Naik and [2]Niyati J. Shah
[1]Assistant Professor and [2]Student (MSc.IT, Part-II), Department of IT, Usha Pravin Gandhi College of Arts Science and Commerce,

ABSTRACT
Until recently, sales and marketing has been a guessing game, one strategy follows another until something sticks. To know exactly what's going on with the sales and marketing, one needs to embrace the power of analytics. Data analytics can help all businesses to grow and develop more by utilizing the data they already have. The study will tell how online business should utilize data and start analyzing their business, predictions of customers' behavior. Survey has also been done on the small business not using data analytics and why they should start using for effective growth in the business. Various classification algorithms are used in the present study to predict how frequently discounts have to be given to the customers. It is found that SVM algorithm performs best among applied classification algorithms.

Keywords: Data Analytics, prediction, SVM, Naïve Bayes.

I. INTRODUCTION

In today's world, everyone wants to grow their business. When it comes to the success of the business, online business plays a key role. Unfortunately, far too many firms regard reporting as the primary goal of installing an analytics system. Reports provide data rather than insights. Although data is essential, insights are priceless. Insights from analysis may be used to drive actions or adjustments, transforming a good business into a great one.

It's no surprise that CEOs are irritated when they discover that their analytics budgets are being squandered on standing around Setup rather than investing time obtaining relevant insights and maximizing for success.

Reports are only effective if they assist any firm to go forward by serving as a springboard for inquiries, ideas, and analysis.

When someone markets the business and promotes the products & services, online business is the best option. Online business is nothing but a boon for any business. Data Analytics has escalated in e-commerce cultural in recent years. But the concept survey has been poor. Now, lots of Companies have widely embraced the use of analytics to streamline operations and improve processes. Data analytics will give information that allows businesses to make better decisions, especially in the sale department. Analytics are starting to play a major role in almost every aspect of life, not just in business. Research of any type is a method to discover information. Within analytical research articles, data and other important facts that pertain to a project is compiled after the information is collected and evaluated, the sources are used to prove a hypothesis or support an idea. Using critical thinking skills (a method of thinking that involves identifying a claim or assumption and deciding if it is true or false) a person is able to effectively pull-out small details to form greater assumptions about the material. Some researchers conduct analytical research to find supporting evidence to current research being done in order to make the work more reliable. Other researchers conduct analytical research to form new ideas about the topic being studied.

Analytical research is conducted in a variety of ways including literary research, public opinion, scientific trials and Meta-analysis. Anyone can refine stream line and improve every area of their business, through the use of data. Many believe that only large enterprise has the resources to utilize big data and analytics, yet small business can easily take advantage of it as well. Making sense of this information can be overwhelming, but once they discover a way to integrate it into their own decision-making process, they will quickly realize the myriad of new sales opportunities that are now at their disposal.

II. PROBLEM STATEMENT

The aim of this paper is to study the usefulness of data analytics in business. The present study performs comparative analysis of classification algorithms to predict frequency of offering discounts to the customer in small scale businesses.

III. LITRATURE REVIEW

D. P. Acharjya in his paper [2] did the following study. Big data analysis is a current area of research and development. The basic objective of the paper was to explore the potential impact of big data challenges, open research issues, and various tools associated with it. As a result, the article provides a platform to explore big

ISSN 2394 - 7780

data at numerous stages. Additionally, it opens a new horizon for researchers to develop the solution, based on the challenges and open research issues.

Hiba Alsghaier, Mohammed Akour, Issa Shehabat, Samah Alibaba [3] In this research they have highlighted some aspects of big data and its importance on organizations, business performance and how companies and how the companies can use the famous open-source platform Hadoop to process the data and gain the competitive advantage.

Xiaoli Ren [7] In this paper, they survey the state-of-the-art studies of deep learning- based weather forecasting, in the aspects of the design of neural network (NN) architectures, spatial and temporal scales, as well as the datasets and benchmarks. Then analyze the advantages and disadvantages of DLWP by comparing it with the conventional NWP, and summarize the potential future research topics of DLWP.

Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody [6] in the paper they have researched that, Big Data (BD), with their potential to ascertain valued insights for enhanced decision-making process, have recently attracted substantial interest from both academics and practitioners. Big Data Analytics (BDA) is increasingly becoming a trending practice that many organizations are adopting with the purpose of constructing valuable information from BD. The analytics process, including the deployment and use of BDA tools, is seen by organizations as a tool to improve operational efficiency though it has strategic potential, drive new revenue streams and gain competitive advantages over business rivals. However, there are different types of analytic applications to consider.

Saurabh Malgaonkar, Sanchi Soral, Shailja Sumeet, Tanay Parekhji [5] in their paper they have researched that, Data Analytics is the trending domain that analyses data to observe patterns and predict future outcomes. The outcomes are based upon analysis of past and current trends and behaviors. Data analytics deals with both descriptive and predictive analyses of data. Descriptive Data Analytics summarizes the data, it's behavior and draws useful conclusion from it. Predictive Data Analytics is the branch of data analytics that predicts future outcomes based on the current and historical data. These future predictions are drawn by observing patterns followed for past data and outcomes for the past events for similar scenarios. In this paper, various branches of data analytics have been discussed. Big data analytics architecture gives an overview of the various tools and system structure involved in big data analytics. Big data analytics is closely related to data mining and hence, implements data mining algorithms. Latter part of the paper covers machine learning algorithms and neural networks for training the dataset to recognize patterns for the modeled data and predict outcomes based on the training and pattern recognition. Modeling of data using neural networks helps in generating accurate and exhaustive outcomes.

Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao & Athanasios V. Vasilakos [1] in their paper they have deeply discussed the issue with a brief introduction to data analytics, followed by the discussions of big data analytics. Some important open issues and further research directions will also be presented for the next step of big data analytics.

Patrick Mikalef, Ilias O. Pappas, John Krogstie and Paul A. Pavlou [4] in the paper they have researched that, "As organizations strive to achieve a competitive edge over their rivals, big data and business analytics are now playing an increasingly important role in realizing performance gains (H. Chen, Chiang, & Storey, 2012a). This has signaled an increased interest in the domain by both researchers and practitioners especially over the past few years, with data being now regarded as one of the most valuable organizational resources. Recent studies have begun to empirically demonstrate the value that big data and business analytics have on organizational-level outcomes, such as agility (Ashrafi, Ravasan, Trkman, & Afshari, 2019), innovation (Lehrer, Wieneke, vom Brocke, Jung, & Seidel, 2018), and competitive performance (Côrte-Real, Ruivo, & Oliveira, 2019; Mikalef, Krogstie, Pappas, & Pavlou, 2019). Nevertheless, a recurring finding of these studies is that in order to derive value from big data, firms must develop the organizational capacity to identify areas within their business that can benefit from data-driven insight, strategically plan and execute data analytics projects, and bundle the resource mix necessary to turn data into actionable insight (Gupta & George, 2016; Vidgen, Shaw, & Grant, 2017). While there is significant variation with regards to the term used to denote this capacity, there is overall consensus about the key resources needed to develop a big data analytics or business analytics capability (Gupta & George, 2016). "

IV. NEED OF DATA ANALYTICS IN BUISNESS

A. Predict customers buying behavior:

Think about how one visits the supermarket every week to get the same groceries that they always buy. The same way, customers are not different. They will always buy the similar products form any e-commerce site. They will

ISSN 2394 - 7780

buy the products, run out of them and then buy more similar products. The key is to analyse when they buy, how frequently they buy, how much they buy, what they buy.

Being armed with such data/information one can predict their future needs which will help them to better time management, having enough stock, having regular customers, prioritization of key accounts and a realistic sale forecast.

B. Sale Forecasting
Applying forecasting techniques to business decisions can help companies predict consumer's behavior. Use relevant data that focusses on the factors impacting the retail business industry. Use of predictive analytics for consumer prediction: Predicting consumer's behavior patterns based on customers interactions and transactions is extremely important in digital era.

C. Analyzing strong and week products
Analysing product sales and trends from previous transactions and orders is a great way to spot which product is picking up or being declined. One can set product prizing accordingly, if any product is not being sold since very long and want to get rid of it, they can decrease the profit margin and put a sale on such products whereas, they can increase profit margins for products that are picking up faster.

D. Spot slipping customers
As recommended above tracking and analyzing Customer's behavior will help them find which customers have stopped visiting and buying products, who were frequent buyers in previous months? This way they can also target one time visitors and customers those are not frequently visiting and know their requirements with which they can enhance their products strategy and make more and more customers.

E. Analyze and monitor customer engagement
Understanding customer, when last they visited e-commerce business, latest order, when was the last time they contacted the customer support team. Customer's engagement will help the business understand customers better and what exactly expect from the business and to make products more valuable for non-buying customers as well. Support team queries is most effective way to reach the customer and analyze them.

F. Better segmentation
Generating excel sheets and factorizing data is the best way of segmentation. Take the data business holds and divide it up by grouping similar data set together based on the chosen parameters so that one can use it more efficiently within marketing and operations. The more data business has more it will allow to really segment the data into useful action points.

G. Analyze old data:
The former data holds the most valuable information that can be leveraged to improve sales. Identify which tactics work best and gradually they can approach and improve each time. In order of this, they can have a lead and have opportunities to detect sales. They can improve tactics by using some marketing techniques. Example: As these two products go hand in hand and are similar they could offer a 10% discount when they buy both together.

V. RESEARCH METHODOLOGY

A. DATA COLLECTION

A collection of questions was supplied to small business in the form of a Google form, using that database, a data set was generated in .csv format and various classification algorithms are applied on this dataset.

The questionnaire was been generated looking at different surveys and keeping lots of online business points in mind. Mainly online based small business entrepreneurs have been targeted as respondent to this questionnaire. Almost had received 43 responses through the survey, selling products related to clothing, food or accessories.

All questions were about business, such as the age range targeted, marketing techniques, product advantages, voucher code strategies, advertisements performed, purchase behavior, the influence of unfavorable reviews, and company loans.

The.csv file's data was cleaned up so that it could run successfully, Categorical variables are converted into numerical variables also textual data were converted to binary values using Hot Encoding, which generates new columns for each of the categories and assigns binary values: 0's or 1's.

The frequency of Discount offered is the last and target attribute of the dataset. Where, the labels of the categories were once per quarter, once a year, once a month.

ISSN 2394 - 7780

Once the final sorted data file was ready, multiclass classification algorithms are used to predict the frequency of discount offered based on the data such as, KNN Algorithm, Decision Tree Algorithm, SVM Algorithm, Naïve Bayes Algorithm. Further in the paper these classifiers are explained in detail.

b. Multiclass Classification algorithms used

Multiclass classification is a popular problem solving method that falls under supervised machine learning category. Given a dataset of m training examples, each of which contains information in the form of various features and a label. Each label corresponds to a class, to which the training example belongs. In multiclass classification, have a finite set of classes. Each training example also has n features. In a multiclass classification, train a classifier using our training data and use this classifier for classifying new examples.

1) **Decision Tree algorithm:**

A decision tree classifier is a method for multiclass classification that is systematic. It asks the dataset a series of queries about its attributes/features. A binary tree may be used to depict the decision tree classification technique. A query is presented on the root and each of the internal nodes, and the data on that node is further separated into individual records with various features. The classes into which the dataset is divided are represented by the tree's leaves. Train a decision tree classifier in scikit-learn in the following code snippet.

2) **SVM (Support vector machine) algorithm:**

When the feature vector is high in dimension, SVM (Support vector machine) is an effective classification approach. One may define the kernel function in sci-kit learn (here, linear). For more information on kernel functions and SVM, see Kernel operator or sci-kit learn and SVM.

3) **KNN or k-nearest neighbors' algorithm:**

The simplest classification algorithm is KNN, or k-nearest neighbors. This categorization algorithm is not affected by the data structure. When a new example is met, the training data's k nearest neighbors are reviewed. The Euclidean distance between two samples can be defined as the distance between their feature vectors. The class for the encountered case is assumed to be the majority class among the k nearest neighbors.

4) **Naive Bayes algorithm:**

Naive The Bayes classification technique is based on the theorem of Bayes. It is called 'Naive' because it implies independence between each pair of characteristics in the data. Let (x1, x2, ..., xn) be a feature vector and y be the class label associated with this feature vector.

VI. ANALYSIS

A. Descriptive Analysis

Here is the summary of few questions asked to the businesses. After the collection of data, this raw data was converted into visualization, pasted below with other details.

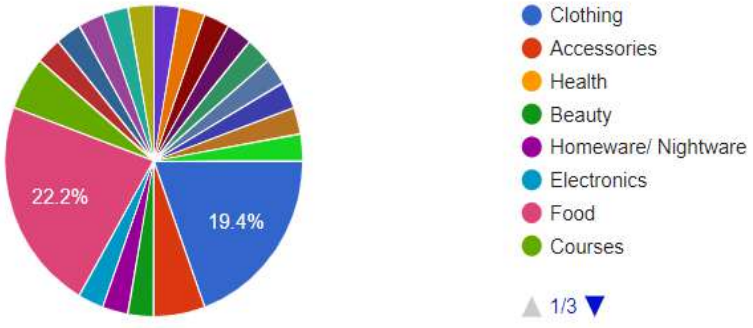


Figure 1: Nature of the product

. The following pie chart illustrates the result of a survey in which the question was asked to refer what products are being sailed the most online. Entrepreneurs are predominantly selling Food (22%) and Clothing (19%), with just 3% difference between those two products. Other products have very fine line of competition with 1% or 2% of scores for each.

In conclusion, since Food and Clothing products are most popular choices, it is clear that many entrepreneurs have businesses and can easily use data analytics to improve the sale of their products more, by predicting customer's behavior

ISSN 2394 - 7780

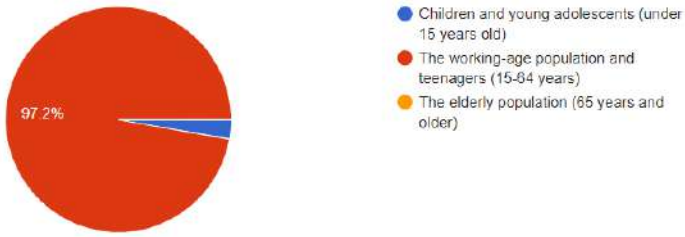


Figure 2: Age groups of customers

The above graph shows the result of a survey in which the question was asked to check that online businesses are targeting which age group the most.

From the pie chart it is clear that majority of entrepreneurs prefer targeting teenagers and the working-age population which is 15years to 64 years with 97% whereas, just around 3% business prefer age group under 15years targeting the children and young adolescents.

In conclusion, the final analysis is that 15-64 years of age groups is the main source of business.

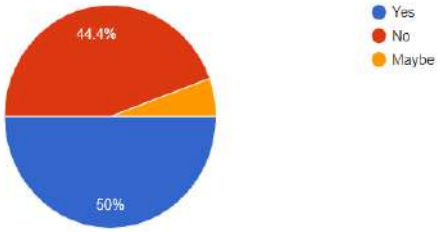


Figure 3: Customer support Team

The chart illustrates the survey to check how any businesses have customer support team. Form the graph it can be seen it's a close difference of 4% between Yes and No. with Yes at 50% and No at 44.4%.

In conclusion, the 44.4% of entrepreneurs (not having a support team) should develop a customer support team for better business strategies.

B. Predictive Analysis

In this analysis, classification algorithms have been used. As, this model is the simplest of the several types of predictive analytics model. It puts data in categories based on what it learns from historical data.

Here will use different multiclass classification methods such as, KNN, Decision trees, SVM, etc. And will compare their accuracy on test data. Also, will perform all this with sci-kit learn (Python).

C. Steps to preform multicalss Classification:

1. Load dataset from the source.

2. Split the dataset into "training" and "test" data.

3. Train Decision tree, SVM, and KNN and Naïve Bayes classifiers on the training data.

4. Use the above classifiers to predict labels for the test data.

5. Measure accuracy and visualize classification.

VII. RESULT AND DISCUSSION

Various mutliclass classification algorithms are applied on the dataset. The Table 1 shows the percenatge accuarcy of the each algorithm.

Table1: Accuracy of multiclass classifiers

Accuracy	Multiclass Classifier
78%	Decision Tree Algorithm
85%	SVM Classifier
65%	KNN Algorithm
69%	Naïve Bayes Classifier

ISSN 2394 - 7780

According to the analysis and the output recevied after running the code for each algorithm, SVM Algorithm performs the most effective and provides the most accurate data.

VIII. CONCLUSION

In the past few years, data analytics has taken a seat from both academic and the e-commerce industry. The online business that use data analytics have higher productivity than their competitors. According to this study, there are many algorithms used for classification in machine learning but SVM is better than most of the other algorithms used as it has a better accuracy in results.

SVM Classifier in comparison to other classifiers have better computational complexity and even if the number of positive and negative examples are not same, SVM can be used as it has the ability to normalize the data or to project into the space of the decision boundary separating the two classes.

REFERENCES

1. Chun-Wei Tsai, Chin-Feng Lai, Han-Chieh Chao & Athanasios V. Vasilakos, "Big data analytics: a survey", Journal of Big Data volume 2, Article number: 21 (2015).
2. D. P. Acharjya, Kauser Ahmed P, “A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016
3. Hiba Alsghaier, Mohammed Akour, Issa Shehabat, Samah Alibaba, “The importance of big data analytics in business: A case study”, American Journal of Software Engineering and Applications Vol. 6(No. 4):pp. 111-115 DOI: 10.11648/j.ajsea.20170604.12
4. Patrick Mikalef, Ilias O. Pappas, John Krogstie and Paul A. Pavlou, "Bigdata and business analytics: A research agenda for realizing business value", 1. Norwegian University of Science and Technology, Department of Computer Science, Sem Sælandsvei 9 7491, Trondheim, Norway 2. SINTEF Digital, Department of Technology Management, S. P. Andersens veg 15B, 7031 Trondheim, Norway 3. University of Agder, Department of Information Systems, Universitetsveien 25, 4604, Kristiansand, Norway 4. University of Houston, Bauer College of Business, Melcher Hall, TX 77204-6021, Houston, USA
5. Saurabh Malgaonkar, Sanchi Soral, Shailja Sumeet, Tanay Parekhji, "Study on big data analytics research domains", 2016 5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), DOI: 10.1109/ICRITO.2016.7784952
6. Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody, "Critical analysis of Big Data challenges and analytical methods", Journal of Business Research Volume 70, January 2017, Pages 263-286.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

PREDICTING THE NUMBER OF HASHTAGS REQUIRED TO PROMOTE BUSINESS USING CLASSIFICATION ALGORITHMS

Neelam Naik, Rutvik Thakrar and Yash Vora
Usha Pravin Gandhi College of Arts, Science and Commerce, Vile Parle, Mumbai, University of Mumbai, India

ABSTRACT
Along with the development of Internet and technology users in the world, one of the marketing approaches that can be use by company to approach customers is digital marketing, while social media and email marketing are digital marketing tools that is most suitable for marketing nowadays. Beside of the easiness and robustness of social media and email marketing the digital marketing tools that have correlation and effectively building customer engagement. A company that can build a good engagement with customers will have a higher purchase intention compare with company who do not have a good customer engagement. On the other hand, by using digital marketing tools, it will provide several benefits for a company such as easy access to promote their product and build relationship with customers, suppress the expenditure, also effecting on the increase of sales volume. This study will tell how hashtags will help to grow their business with digital marketing. Survey has been done on the small businesses which uses digital marketing as their marketing strategy. Various classification algorithms are used in the present study to predict how many hashtags will help them to target relatable audiences. It is found that decision tree algorithm performs best among applied classification algorithms.

Keywords: Decision Tree, KNN classifier, Naïve Bayes Classifier, Support Vector Machine, Digital Marketing

I. INTRODUCTION

Digital Marketing is defined as marketing or advertising in each aspect of the internet. It's also called online marketing. It's the elevation of products or brands to connect with implicit guests using the internet and other forms of digital communication. It's described as advertising on different platforms like social media, websites, search engines, emails, etc. With the help of the Digital Marketing business proprietor can get a large niche audience and can promote their product at a distinct position at the same moment. Digital marketing can be extensively broken into 7 main types including Search Engine Optimization, Pay-per- Click, Social Media Marketing, Content Marketing, Email Marketing, Mobile Marketing, Marketing Analytics.

In 2021, 91.90% of small businesses like Digital Marketing over Traditional Marketing. Small and medium enterprises applying digital marketing techniques will have 3.3 times better chances of developing their workforce and business. Digital Marketing is more favourable due to its good ROI (Return of Investment). Digital Marketing gives good ROI compared to Traditional Marketing. In survey it's estimated that Digital Marketing campaign can be done in minimal $12000 in other side Traditional Marketing take $8000 for only set up the campaign and $8000 to $10000 for run that campaign. So, Traditional Marketing is expensive for small and medium businesses. In Digital Marketing we can reach to more people rather than in Traditional Marketing.

As shown in figure 1.1, it's visible that costing for 2000 audience reach is$ 1000 in Traditional Marketing but on the other hand it'll cost$ 125 in Digital Marketing. Alternate benefit of using Digital Marketing is that people now a days spending nearly 10-12 hours on internet out of 24 hours.

In the world of Marketing or Advertising, engagement is the only success that businesses can get. Business owners which selected Digital Marketing over Traditional marketing has higher chance of Engagement due to its niche audience and large amount of sample population. In 2019 a survey was conducted regarding which strategy small businesses want to choose, 35% of businesses chooses Digital Marketing over Traditional Marketing. 23% businesses use both as their marketing strategy and 31% businesses uses only traditional marketing. This figure of Digital Marketing will be expected to reach 45% to 48% in 2022. 90% from this 35% businesses received good ROI (Return on Investment). It shows that Digital Marketing can give us good result. In Digital Marketing, businesses can predict their outcome by seeing things like engagement of audience, reach of audience through hashtags, etc. At last, we can say Digital Marketing is more effective compare to Traditional Marketing.

ISSN 2394 - 7780

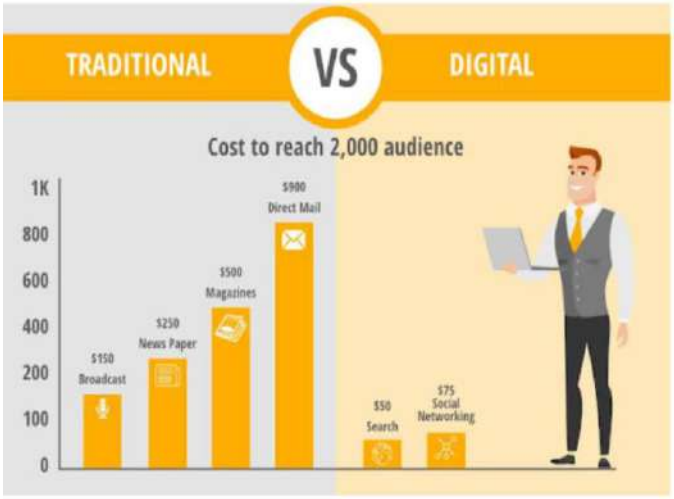


Figure 1.1 cost for 2000 audience in Digital Marketing vs Traditional Marketing

II. PROBLEM STATEMENT

"Many businesses are aware of Digital Marketing but there is lack of knowledge of Digital Marketing" The current project aims to give an output to the businesses that can boost their marketing via use of various strategies related to Digital Marketing. So, this project will give them a guidance about how they can promote their product in the world of Digital Marketing by using various Digital Marketing Strategies.

III. LITERATURE REVIEW

In the paper [1] the authors are pointing towards many events in the world which are accompanied by the Hash-Tag trends on social media. Whole idea behind this is to create such an application that would help in marketing of products and services over social media platforms. The technique is known as Social Media Marketing and is a sub-set of Digital Media Marketing. The algorithm which they have used is Multi-layer Perceptron (MLP) which can be viewed as a logistic regression classifier where the input is first transformed using a learnt non-linear transformation where this transformation projects the input data into a space where it becomes linearly separable. This intermediate layer is referred to as a hidden layer.

In the paper [2] the authors are trying to determine the impact of digital marketing especially social media and email marketing on how the tools give an effect to customer engagement, which will also give an impact to purchase intention. They have used the method of Quantitative approach which are used with associative method to elaborate the objective of the research, cross sectional is used for time horizon of this research.

In the paper [3] the authors of this study focus on exploring the Search Engine Optimization from the managerial perspective as well as identifying the role and importance of SEO in increasing the profitability of firms. It also analyses the present scenario and future scope of Search Engine Optimization in Digital Marketing. They have compared the different types of SEO like on page and Off page SEO, mobile vs Desktop SEO.

In the paper [4] the authors are focusing on the selection and adoption of the ML-driven analytical tools by three distinct groups: marketing agencies, media companies, and advertisers. They behave used the method of Qualitative research, using this method of in-depth interviews, was used to gain insight into the way ML is practically used in marketing.

In the paper [5] authors propose a framework that provides a competitive advantage for an organization's marketing by providing analysts with a visibility into the tag behaviour on their organization's site in real-time. They propose a methodology of providing a framework that provides a competitive advantage for an organization's marketing by providing analysts with a visibility into the tag behaviour on their organization's site in real-time.

In the paper [6] authors are trying to deal with a new approach that will optimize web technologies for the evolution of user trends, and therefore, will be of academic and professional use for marketing managers and web solution developers. The Delphi technique was used in this research. This technique is one of the research methods used for prospective research. Delphi is a prospective method that is considered suitable for analysing opinions made about any topic

ISSN 2394 - 7780

In the paper [7] authors are proposing to study visually research mapping and research trends in the field of digital marketing on an international scale. This study used the methodology of bibliometric techniques with secondary data from Scopus. Analyse and visualize data using the VOS Viewer program and the analyse search results function on Scopus.

In the paper [8] authors are setting the goal of the paper to introduce the new technique for the start-ups and reduce the effort for offsite SEO and work has classified low competition keywords and generate quicker results and revenue. This paper will compare results of content written based on identification of low competition keywords given by keywords tools and those which qualify the criteria of low competition by the analysis of results generated by all intitle operator.

IV. RESEARCH METHODOLOGY

A. DATA COLLECTION

A google form was created to gather the data in which there were many questions were asked to gather the data such as

- Name and Details of Product?
- Which social media do you prefer?
- What would be the best time zone to post your products?
- What age group of customers are you targeting?
- Gender of the customer?
- Number of Hashtag used in post?
- Which month of year you prefer for this product to market?
- Has Digital Marketing given you expected result? Or getting boost from it?
- Any other type of marketing does you prefer if yes, then which?

The form was distributed to various known small and medium scale business, and also to various clients of Digital Marketing service provider firms.

B. Pre-Processing of Data

Data collected is cleaned and pre-processed to make it ready to use for ml algorithms and models. Data gathered from the google forms was imported to a excel file which was a human readable data but for using ml algorithm I was pre-processed to numerical data for that each column was divided into multiple columns and it was separated using 1's and 0's for example: Which social media do you prefer was the column and the answer were Instagram, LinkedIn, Facebook, Twitter. So, the new four column were inserted and the data was divided. The target column was how many hashtags are used for predicting the success of digital marketing so it was divided into four parts which were labelled as follows.

1-9 is labelled as one

10-15 is labelled as two

16-25 is labelled as three

26 to 30 is labelled as four

C. Analysing Different Algorithms

Different Multiclass Classification Algorithms such as Random Forest, k-Nearest Neighbour, Decision Tree, Naive Bayes are analysed to find the best suitable for the data which has been collected.

Decision tree classifier – A decision tree classifier is a systematic approach for multiclass classification. It poses a set of questions to the dataset (related to its attributes/features). The decision tree classification algorithm can be visualized on a binary tree. On the root and each of the internal nodes, a question is posed and the data on that node is further split into separate records that have different characteristics. The leaves of the tree refer to the classes in which the dataset is split. [9]

SVM (Support vector machine) classifier – SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). SVM maps training examples to

ISSN 2394 - 7780

points in space so as to maximise the width of the gap between the two categories. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. [9]

KNN (k-nearest neighbours) classifier – KNN or k-nearest neighbours is the simplest classification algorithm. This classification algorithm does not depend on the structure of the data. Whenever a new example is encountered, its k nearest neighbours from the training data are examined. Distance between two examples can be the Euclidean distance between their feature vectors. The majority class among the k nearest neighbours is taken to be the class for the encountered example. [9]

Naive Bayes classifier – Naive Bayes classification method is based on Bayes' theorem. It is termed as 'Naive' because it assumes independence between every pair of features in the data. Let $(x_1, x_2, \ldots, x_n)$ be a feature vector and **y** be the class label corresponding to this feature vector. [9]

V. ANALYSIS

The pre-processed data is analysed using four multiclass classification algorithms. These algorithms are Decision tree classifier, support vector machine classifier, KNN classifier and Naïve Bayes classifier. The procedure of applying these algorithms is as shown in figure 2.

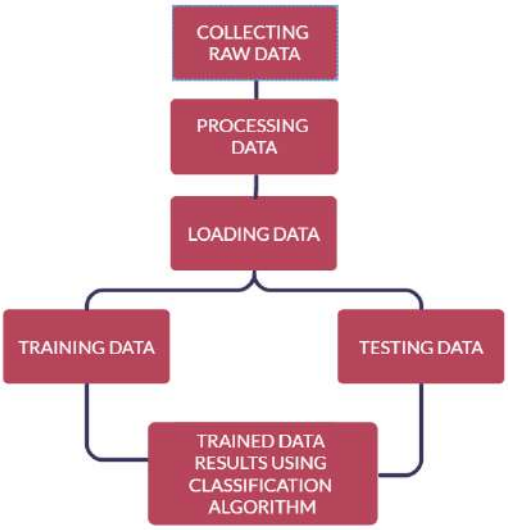


Figure 2: Procedure of applying multi-classifiers

Steps to preform multicalss Classification

1. Decision tree, SVM, KNN and Naïve Bayes classifiers, from this one algorithm is selected to get the accuracy of the data.

2. After this package are imported in respect to selected algorithms

3. Load dataset from the source- The .CSV in which the raw data was preprocessed into numerical form is loaded in the program.

4. Then the range of data values is assigned as per .csv file. X is assigned for all the columns except the target column and Y is assigned for the target column only

5. Then split the dataset into "training set" and "testing set"

6. Then Measure accuracy and visualize classification.

VI. RESULTS AND DISCUSSIONS

Various multiclass classification algorithms are applied on the dataset. The Table 1 shows the percentage accuracy of the each algorithm.

Table1: Accuracy of multiclass classifiers.

Algorithm	Accuracy (%)
Decision Tree Classifier	78.23
SVM	65.45
KNN Classifier	56.89
Naïve Bayes	61.72

ISSN 2394 - 7780

According to the analysis and the output received after running the code of each algorithm, Decision Tree algorithm perform the most effective and provides the most accurate results.

VII. CONCLUSION

Due to increase in digital marketing the requirement of the knowledge of number of hashtags to be used while marketing your product on social media. this study will let you identify the number of hashtags to be used According to this study, there are many algorithms used for classification in machine learning but Decision Tree Classifier is better than most of the other algorithms used as it has a better accuracy in results. Decision Tree Classifier is better than other classifiers in this research because it does not require data normalization and scaling of data. Also the missing values and sparse data present in the dataset does not affect the outcome of the decision tree algorithm.

There may be more enhancement in future where this study can be extended with increasing number of algorithms to train the data. Dataset can be also increased as this study has a minimal data and the number of features in the data set can be also improvised. By using large dataset, it may give a better result and accuracy.

VIII. REFERENCES

1. Harsh Namdev Bhor, Tushar Koul, Rajat Malviya and Karan Mundra "Digital Media Marketing using Trend Analysis on social media" IEEE Xplore
2. Aryo Bismo, Sukma Putra and Melysa "Application of Digital Marketing (social media and email marketing) and its Impact on Customer Engagement in Purchase Intention" PT. Soltius Indonesia
3. Mr. Himanshu Matta, Dr. Ruchika Gupta and Dr. Sumit Agarwal "Search Engine Optimization in Digital Marketing: Present Scenario and Future Scope" 2020 International Conference on Intelligent Engineering and Management (ICIEM)
4. Andrej Miklosik, Martin Kuchta, Nina Evans and Stefan Zak "Towards the Adoption of Machine Learning-Based Analytical Tools in Digital Marketing" Digital Object Identifier 10.1109/ACCESS.2019.2924425
5. Andy Bengel, Amin Shawki and Dippy Aggarwal "Simplifying Web Analytics for Digital Marketing" 2015 IEEE International Conference on Big Data (Big Data)
6. Juan Jose Lopez Gracia, David Lizcano, Celia MQ Roamos and Nelson Matos "Digital Marketing Actions That Achieve a Better Attraction and Loyalty of Users" 26 April 2019 MDPI
7. Agung Purnomo, Nur Asitah Nuzula, Firdausi Syaifurrizal, Wijaya Putra, Moch. Khafidz and Fuad Raya "A Study of Digital Marketing Research Using Bibliometric Analysis" 14 September 2021 IEEE
8. Asad Nadeem, Musa Hussain and Ahsan Iftikhar "New Technique to Rank Without Off Page Search Engine Optimization" November 2020 IEEE
9. Arik Pamnani [Geeks for Geeks,(Sept, 2021)] https://www.geeksforgeeks.org/multiclass-classification-using-scikit-learn/ . (Retrived on 25 December 2021)

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

STOCK PRICE PREDICTION USING MACHINELEARNING

[1]Prashant Chaudhary and [2]Omkar Gharat
[1]M.Sc. [I.T.] Assistant Professor, [2]M.Sc. [I.T] Student, Usha Pravin Gandhi College, Vile parle (W), Mumbai

ABSTRACT
One of the most difficult machine learning issues is stock price prediction. It is determined by a variety of factors that influence variations in supply and demand. With the emergence of artificial intelligence and increasing processing power, programmed methods of prediction have shown to be more effective in predicting stock prices. We will showand discuss a more practical method for predicting a company's next "n" days closing price in this article. The dataset of stock market values from the previous year is the first factor we considered. For real-time analysis, the dataset was pre-processed and fine-tuned. As a result, we will concentrate on data pre- processing of the raw dataset in our article. Second, after pre-processing the data, we'll lookat how Linear Regression and Decision Tree Regression were used on the dataset and the results they produced. Stock market institutions will benefit greatly from good stock prediction since it will bring real-world solutions to the challenges that stock investors encounter.

Keywords: Machine Learning, Data Pre- processing, Dataset, Stock price, Stock Market.

I. INTRODUCTION
The stock market is essentially a collection of different stock buyers and sellers. In general, astock (also known as shares) represents ownership claims on a corporation by a single person or a group of people. A stock market forecast is an attempt to predict the future value of the stock market. The prediction is anticipated to be accurate,reliable, and time-saving. The system must be designed to work in real-world conditions and should be well-suited to them. The system is also supposed to account for all potential influences on the stock's value and performance. The method of predicting stock market prices for short time periods appears to be random. Over a long period of time, the price of a stock usually forms a linear curve. People like to invest in equities that are predicted to rise in value in the near future. People avoid investing in equities because of the stock market's instability. As a result, there is a requirement to precisely predict the stock market ina real-world scenario. Time series forecasting, technical analysis, machine learning modelling, and predicting the variable stock market are some of thestrategies used to anticipate the stock market. The stock market prediction model's datasets contain information such as the closing price, opening price, data, and a variety of other variables that are required to predict the object variable, which is theprice on a given day. Linear regression is used in business, science, and almost any other industry where predictions and forecasts are important. It aids in the identification of links between one or more independent variables and a dependentvariable. The term "simple linear regression" refers to the use of a feature to predict an outcome. The application of linear regression to stock market forecasting seems appealing. These analyses can be implemented in a few lines of code using modern machine learning packages such as scikit-learn. A decision tree is a tool that draws a tree structure with various decisions and their possible outcomes.It works seamlessly with both continuous and output variable inputs.

II. LITERATURE SURVEY
The literature survey is an important aspect in maintaining consistency. It refers to the steps that must be taken during the development process. Software development necessitates the legitimacy of resources as well as their availability. This section aids in the discovery of previously worked-on content, as well as the application and implementation of the same in today's world. The economy and the strength of the product are the most important factors in development. The support and resource flow must be monitored and computed once the innovation has gone through thebuilding phase. This is also known as the Research phase, and it is here that all of the research is incorporated and completed in order to keep the flow going.

The approaches and established facilities for exploitation, particularly of financial news, social media data, and analytical findings, are provided. A prediction model has been developed that employs big data analytical capabilities, social media analytics, and machine learning to predict stock market trends on a regular basis.

A. Survey of Stock Market Prediction UsingMachine Learning Approach :
In the current day, stock market forecasting has become increasingly crucial. Technical analysis is one of the strategies used, but such procedures may not always produce reliable results. As a result, it's critical to create ways for making more accurate predictions. Generally, investments are made basedon projections derived from the stock price after taking into account all possible influences. In this case, regression was used as the method of choice. Because financial stock exchanges create massive amounts of data at any given time, a significant

ISSN 2394 - 7780

amount of data must be analysed before a prediction can be formed. Each of the approaches covered under regression has its own set of benefits and drawbacks in comparison to its peers. Linear regression was cited as one of the remarkable strategies.

B. Stock Market Prediction: Using HistoricalData Analysis

The stock market forecasting process is fraught with risk and can be influenced by a variety of factors. As a result, the stock market has a significant impact on business and finance. The emotional analysis procedure is used to perform technical and fundamental analysis. Because of its widespread use, social media data has a significantimpact and can be useful in anticipating stock market trends. Machine learning algorithms areused to do technical analysis on historical stock price data. The strategy usually entails gathering various social media data and news in order to derive individual sentiments. Other information, such as past year's stock prices, is also taken into account. The relationship between different data points is taken into account, and a prediction is generated based on these data points. The model was able to forecast stock prices in the future.

C. Stock Market Prediction via Multi-SourceMultiple Instance Learning

Predicting the stock market accurately is a difficult undertaking, but the current web has shown to be a valuable tool in making it easier. It is possible to extract specific sentiments due to the interconnected style of data, making it easier to build links between various variables and generallyscope out an investing pattern. Investment patterns from multiple firms exhibit signs of similarity, and using these same consistencies amongst data sets is the key to accurately predicting the stock market. Stock market information may be successfully anticipated by analysing more than just technical historical data, and by employing other tools such as sentiment analyzers to derive a crucial link between people's emotions and how they are influenced by certain stock investments.[1]

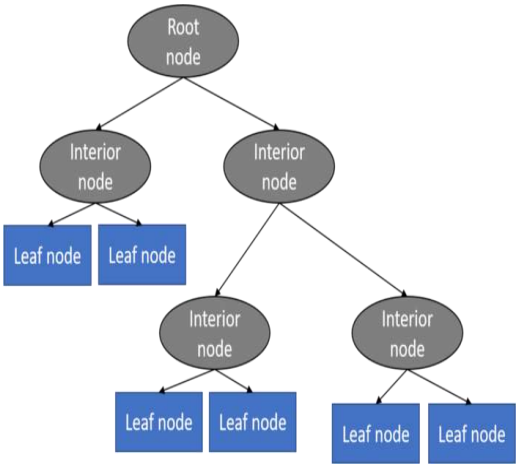
D. Machine Learning Approach In Stock MarketPrediction :

The vast majority of stockbrokers used specialised, fundamental, or time series analysis to make their predictions. Overall, these methodologies could not be totally trusted, necessitating the development of a solid system for financial exchange forecast. The methodology selected to use machine learning and AI, as well as a supervised classifier, to achieve the most accurate outcome[2].Prediction stock price used parse records to compute the expected, supply it to the user, and perform operations such as purchasing and promoting stocks autonomously through the use of the automation concept[3]. The Decision Tree Regression Algorithm and Linear Regressionwere employed..

E. DATASET

The historical data for the StarBucks company has been collected from Yahoo Finance [4]. The dataset includes 1 year data from 15/12/2020 to 15/12/2021 .The data contains information about the stock such as High, Low, Open, Close, Adjacent close and Volume. Only the day-wise closing price of the stock has been extracted. we'll be utilizing scikit-learn, csv, NumPy and matplotlib bundles to actualize and picture direct relapse toward the mean.

Open	High	Low	Close
97.24	97.92	96.86	97.76
97.42	97.54	96.95	97.01
97.62	97.86	96.88	96.88
97.51	98.72	98.3	98.3
98.5	98.54	97.94	98.2



ISSN 2394 - 7780

III. RESEARCH METHODOLOGY

A. REGRESSION

Regression is a method of forecasting that models the connection between a dependent variable and independent variables. It is essentially a mathematical approach to determining the relationship between variables. In machine learning, this is frequently used to predict the outcome of an incident based on the relationship between variables acquired from the data-set. One type of regression used in Machine Learning is statistical regression.

B. Linear Regression Algorithm

Linear Regression is a supervised learning-based machine learning technique. It carries out a regression analysis. On the basis of independent variables, regression models a goal prediction value. It is mostly used to determine the link between variables and predicting. Different regression models differ in terms of the type of relationship they evaluate between dependent and independent variables, as well as the amount of independent variables they utilise.

For visualisation, we shall use Stock Charts. Stock charts are one of the most important financial graphs because they allow investors to follow markets, calculate gains and losses, and make purchasing and selling choices. While a number of graphs are used to show market fluctuations, the most popular is most likely the simple line graph transformed into a histogram. The lines simply show changes in the value of a certain stock or the whole market over time. By converting the line graph into a stacked area chart or simply utilising numerous lines of varying colours, multiple stocks can be tracked and compared at the same time.

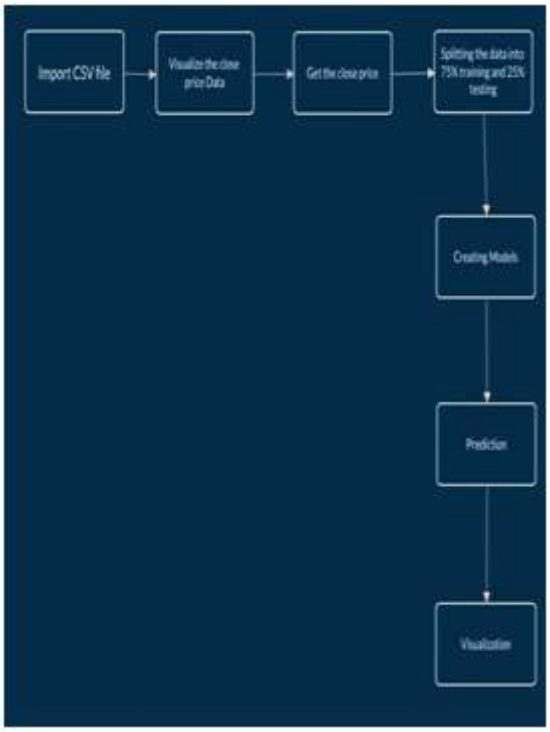
C. Decision Tree Algorithm

It's a tree-structured classifier with three different kinds of nodes. The Root Node is the starting node that represents the full sample and can be subsequently subdivided into nodes. The features of a data collection are represented by the interior nodes, while the decision rules are represented by the branches. Finally, the outcome is represented by the Leaf Nodes. This method is quite beneficial for dealing with decision-making issues.

A specific data point is traversed through the entire tree by answering True/False questions until it reaches the leaf node. The final prediction is the average of the dependent variable's value in that specific leaf node. The Tree is able to forecast a suitable value for the data point after several rounds.

Decision trees have the advantage of being simple to grasp, requiring minimal data cleaning, non-linearity having no effect on model performance, and having a small number of hyper-parameters to modify.

We will walk through the implementation of Decision Tree Regression in this article, in which we will predict the price of a company's stock for "n" days.



ISSN 2394 - 7780

V. EXPERIMENTAL RESULTS

The csv file contains the raw data based on which we are going to publish our findings. There are seven columns or seven attributes that describe the rise and fall in stock prices. Some of these attributes are (1) HIGH, which describes thehighest value the stock had in previous year. (2)LOW, is quite the contrary to HIGH and resembles the lowest value the stock had in previous year (3) OPEN is the value of the stockat the very beginning of the trading day, and (4)'CLOSE' stands for the price at which the stock is valued before the trading day closes. There areother attributes such as 'DATE', 'OPEN','HIGH','LOW','CLOSE' and VALUE, but the above mentioned four play a very crucial role in our findings attributes.

	A	B	C	D	E	F	G
1	Date	Open	High	Low	Close	Adj Close	Volume
2	19-11-2020	97.24	97.92	96.86	97.76	96.17024	4252300
3	20-11-2020	97.42	97.54	96.95	97.01	95.43244	4609700
4	23-11-2020	97.62	97.86	96.42	96.88	95.30454	4803200
5	24-11-2020	97.51	98.72	97.4	98.3	96.70147	6319900
6	25-11-2020	98.5	98.54	97.94	98.2	96.60308	4027100
7	27-11-2020	98.48	98.98	98.28	98.66	97.05561	2169700
8	30-11-2020	98.2	98.29	96.96	98.02	96.42601	5197300
9	01-12-2020	99	99.26	98.25	98.82	97.21301	4970000
10	02-12-2020	98.51	99.04	98.21	98.91	97.30155	3377900
11	03-12-2020	99.02	101	98.97	100.11	98.48203	6264100
12	04-12-2020	101.35	102.94	101.07	102.28	100.61675	6952700
13	07-12-2020	102.01	102.22	100.69	101.41	99.76089	4514800
14	08-12-2020	100.37	101.57	100.01	101.21	99.56413	3911300
15	09-12-2020	101.94	102.21	100.1	100.4	98.76731	6629900
16	10-12-2020	103.51	106.09	102.75	105.39	103.67617	12939200
17	11-12-2020	104.41	104.78	102.33	103	101.32504	6262600
18	14-12-2020	103.83	104.71	103.25	103.32	101.63982	5156200
19	15-12-2020	104.24	104.86	103.78	104.18	102.48584	5195200
20	16-12-2020	104.1	104.8	102.72	103.27	101.59064	6409300
21	17-12-2020	103.55	104.04	102.61	103.21	101.53162	4535700
22	18-12-2020	103.33	104.11	102.95	103.28	101.60048	10215000
23	21-12-2020	101.22	103.15	100.02	102.94	101.26601	7175900
24	22-12-2020	102.12	103.17	101.89	102.41	100.74463	4382700
25	23-12-2020	102.29	102.69	101.97	102.06	100.40031	3817300

This is a pictorial representation of the data present in our csv file. This particular file contains 254 such records. There are more than ten different trading codes available in the dataset and some of the records do not have relevant information that can help us train the machine, so the logical step is to process the rawdata. Thus we obtain a more refined datasetwhich can now be used to train the machine.

	Date	Open	High	Low	Close	Adj Close	Volume
0	2020-11-19	97.239998	97.919998	96.860001	97.760002	96.170242	4252300
1	2020-11-20	97.419998	97.540001	96.949997	97.010002	95.432442	4609700
2	2020-11-23	97.620003	97.860001	96.419998	96.879997	95.304543	4803200
3	2020-11-24	97.510002	98.720001	97.400002	98.300003	96.701469	6319900
4	2020-11-25	98.500000	98.540001	97.940002	98.199997	96.603081	4027100

This is the result of using the head(). Since we are using the pandas library to analyse the data, it returns the first five rows. Here five is the default value of the number of rows it returns unless stated otherwise.

	Close
0	97.760002
1	97.010002
2	96.879997
3	98.300003
4	98.199997

we want only "Close" column , so we had printedall Close values as follows with index numbers.

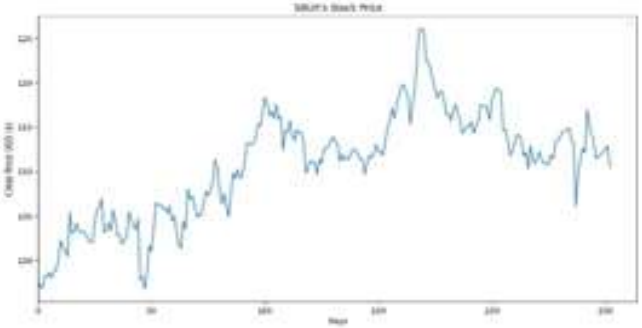


Fig 4: Stock Price Prediction Model

ISSN 2394 - 7780

This is a plot generated on basis of original stockprices using the "matplotlib.pyplot" library. Theplot is of the attributes "Days" vs "Close Price USD ($)". This is to show the trend of closing price of stock as time varies over a span of one year.

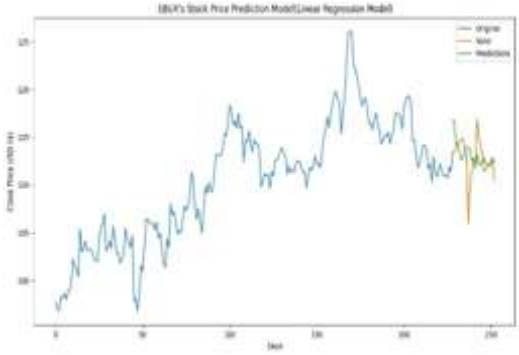


Fig 5 Stock Price Prediction Model(Linear Regression)

This is a plot generated on basis of predicted values of stock using Linear Regression from using the "matplotlib.pyplot" library. The plot is of the attributes "Days" vs "Close Price USD ($)". This is to show the trend of closing price ofstock as time varies over a span of one year.

Fig 5 Stock Price Prediction Model(Decision Tree Regression)

This is a plot generated on basis of predicted values of stock using Decision Tree Regression from using the "matplotlib.pyplot" library. The plot is of the attributes "Close Price USD ($)" vs "Days". This is to show the trend of closing priceof stock as time varies over a span of one year.

VI. CONCLUSION

The Machine learning has several wonderful applications, and it is still a very popular technique that is heavily reliant on data, even though it has grown in the future into a neural network and deep learning. Determining stock market returns is a tough challenge since stock values are constantly changing and are based on various criteria that produce complex patterns. The historical dataset accessible on the company's website includes only a few variables such as high, low, open, close, adjacent close value of stock prices, volume of shares traded, and so on, which are insufficient.

The purpose of this paper is to anticipate stock market prices using machine learning. There are other approaches to accomplish stock price prediction, however only two regression algorithms were used. The primary goal of this work is to employ a machine learning algorithm. For stock price prediction, Linear Regression and Decision Tree Regression have been used. The results show that Linear Regression has higher accuracy for bothsmall and large datasets. Decision Tree Regression, on either side, conveys the weak prediction price considering the size of the dataset.

VII. FUTURE ENHANCEMENT

Future scope of this project will involve adding more parameters and factors like the financial ratios, multiple instances, etc. The more the parameters are taken into account more will be the accuracy. The algorithms can also be appliedfor analyzing the contents of public comments and thus determine patterns/relationshipsbetween the customer and the corporateemployee. The use of traditional algorithms and data mining techniques can also help predict thecorporation's performance structure as a whole.

ISSN 2394 - 7780

REFERENCES

[1] Xi Zhang1, Siyu Qu1, Jieyun Huang1, Binxing Fang1, Philip Yu2, "Stock Market Prediction via Multi-Source Multiple Instance Learning." IEEE 2018.

[2] Loke.K.S. "Impact Of Financial Ratios And Technical Analysis On Stock Price Prediction Using Random Forests", IEEE, 2017.

[3] Raut Sushrut Deepak, Shinde Isha Uday, Dr. Malathi, "Machine Learning Approach in Stock Market Prediction", IJPAM 2017.

[4] Yahoo Finance - Business Finance Stock Market News, <https://in.finance.yahoo.com/> [Accessed on December ,2021]

[5] Belfast Telegraph (2020), „Police Warn Businesses of Cyber Crime „Surge" During Lockdown". Available at: https://www.belfasttelegraph.co.uk/news/northern -ireland/policewarn-businessesof-cyber-crime- surge-during-lockdown-39151612.html [Accessed22 August 2020].

[6] Krombholz K., Hobel H., Huber M., & Weippl E. Advanced social engineering attacks // Journal of Information Security and applications,2015, vol. 22, pp. 113-122.

[7] N. K. Dixit and L. Mishra, "THE NEW AGE OF COMPUTER VIRUS AND THEIR DETECTION," International Journal of Network Security & Its Applications (IJNSA), vol. 4, no. 3, pp. 79 - 96, 2012.

[8] Gharehchopogh, F. S., &Khalifehlou, Z. A. (2012). A New Approach in Software Cost Estimation Using Regression Based Classifier. AWERProcedia Information Technology and Computer Science, Vol: 2, pp. 252-256.

[9] Maknickas, A. and Maknickiene, N. (2019), "Support system for trading in exchange market by distributional forecasting model", Informatica, Vol. 30 No. 1, pp. 73-90.

[10] Yan, Robert, and Charles Ling. 2007. "Machine Learning for Stock Selection." Industrial and Government Track Short Paper Collection 1038- 1042.

[11] Nikfarjam, A., Emadzadeh, E., &Muthaiyah, S. (2010). Text mining approaches for stock market prediction. In Computer and Automation Engineering (ICCAE), 2010 the 2nd International Conference on (Vol. 4, pp. 256- 260). IEEE

[12] Pesaran, M. H., &Timmermann, A. (1994). Forecasting stock returns an examination of stock market trading in the presence of transaction costs. Journal of Forecasting, 13(4), 335-367.

[13] VivekKanade, BhausahebDevikar, SayaliPhadatare, PranaliMunde, ShubhangiSonone. "Stock Market Prediction: Using Historical Data Analysis", IJARCSSE 2017.

[14] Chong, E.; Han, C.; Park, F.C. Deep learning networks for stock market analysis and prediction: Methodology, data representations, and case studies. Expert Syst. Appl. 2017, 83, 187–205

[15] Zhu, Y.; Yang, H.; Jiang, J.; Huang, Q. An adaptive box-normalization stock index trading strategy based on reinforcement learning. In Proceedings of the International Conference on Neural Information Processing, Siem Reap, Cambodia, 13–16 December 2018; Springer: Berlin/Heidelberg, Germany, 2018.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

BLOCKCHAIN: ARCHITECTURE AND SWOT ANALYSIS OF APPLICATION

[1]Smruti Nanavaty and [2]Rajvi Mehta
[1]Assistant Professor and [2]M.Sc.IT Student, Usha Pravin Gandhi College of Arts, Science and Commerce

ABSTRACT
The concept of the Internet of Things (IoT), everyday devices are becoming smart, and off-line. This vision is becoming a reality, thanks to the advancement in technology, there are still many challenges that need to be addressed, particularly in the areas of safety and security, such as the reliability of the data. Given the expected development of the Internet of Things in the years to come, it is necessary to ensure that the credibility of such a large incoming source of information. Blockchain has become a key technology that will change the way in which the information is to be exchanged. Building trust in distributed environments without the need for the government there is a technological development that can be a lot of change in the industry, including IoT. With innovative technologies such as big data and cloud computing services will be used by the IoT to overcome its limitations, since its inception, and we think blockchain will be one of them. This article focuses on the Blockchain Architecture and the SWOT (Strength, Weakness, Opportunity, and Threats) Analysis of Blockchain Applications.

I. INTRODUCTION
The Internet of Things (IoT) system is a complex area where many organizations and devices meet. The current system relies on reliability, internally a model that supports interaction due to the diversity of multiple cloud configurations and devices. And the security system for the current infrastructure is limited only to anonymous security leaving many entry/exit points at risk IoT applications. Given the involvement of many stakeholders in IoT, it is important to build trust, prevention, low- level and collaborative systems to create IoT systems that work. Therefore, the need to exploit the blockchain's nature Features are evident in the future of IoT infrastructure.

II. PURPOSE OR OBJECTIVES
This Research paper focuses on the Blockchain architecture, the distribution is working in a very complex manner, with each block that distributes the data across the network. A database of all the systems that are configured with the same details, rules, and terms, so it is referred to as a shared condition. The operation of this architecture to work on three main factors: decentralization, liability, and protection. Also, this research includes the S.W.O.T Analysis of blockchain applications with all security aspects. The research is based on secondary data, collected through various books, business magazines, journals, newspapers, research, online research, websites, etc.,

III. LITERATURE REVIEW
[1] Through cloud-based development, blockchain technology has provided a mechanism to digitize all of our actions. Blockchain is rapidly evolving to be the next disruptive innovation for safe connectivity. It is a complicated technology that encapsulated an inventive and powerful ambition to build a comprehensive answer to internet security. [2] The purpose of this article is to examine the positive and negative aspects that arise when developing Industry 4.0 apps employing blockchain and smart contracts This study also provides a comprehensive overview of the most relevant blockchain-based applications for Industry technologies. As a result, its goal is to give a comprehensive roadmap for future industry developers to determine how blockchain might improve the next generation of cyber-secure industrial systems. [3] We take a close look at current blockchain applications in Industry 4.0 and IIoT settings. We present current research trends in each of the major industrial sectors, as well as successful commercial blockchain implementations in these sectors. We also go into the various challenges that each industry faces when it comes to implementing blockchain. We also look at new application possibilities for blockchain technology in Industry 4.0, as well as some of the existing challenges. We believe that our findings pave the way for more empowering and accessible research in this area, as well as aiding decision- makers in the adoption and investment of blockchain in the Industry 4.0 and IIoT fields. Malik et al. [4] presented a permissions blockchain system that employed a three-tier architecture to assure data availability to consumers and was governed by a food supply chain consortium that included government and regulatory organisations.[5] Furthermore, Ye et al. verified that supply chain data is visible, traceable, and irreversible, as well as considering classified data encryption security, allowing businesses to share data without revealing sensitive information [6] Traditional supply chain processes are rapidly being infiltrated by blockchain technology. Maersk, for example, formed a cross-professional chain alliance by bringing together joint insurance institutions, blockchain businesses, and other stakeholders to build the world's

ISSN 2394 - 7780

first blockchain platform for maritime insurance. [7] The blockchain, which is at the heart of Bitcoin, has recently gotten a lot of attention. Blockchain acts as an immutable ledger that enables decentralized transactions. Manner. Blockchain-based applications are aplenty, spanning a wide range of topics. Several industries, including financial services and reputation management, as well as the Internet of Things (IoT). However, a number of difficulties with blockchain technology, including as scalability and security concerns, must be addressed.

IV. Blockchain Architecture

Blockchain, like a traditional public ledger, is made up of blocks that each include a full list of transaction data.A block has just one parent block if the block header contains a preceding block hash. The hashes of uncle blocks (children of the block's ancestors) would also be preserved in the Ethereum network. In a blockchain with no parent blocks, the genesis block is the first block. The internals of blockchain are then thoroughly discussed.

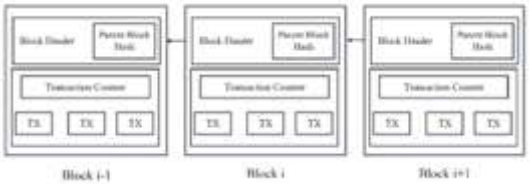


Fig. 1: An example of blockchain

The Blockchain Architecture is made up of essential components. They are as follows:

Node: In the blockchain architecture, a node is a user/computer. Each node maintains its own copy of the blockchain ledger.

Transaction: A transaction, which refers to the records and information stored in the block, is the smallest building block of a blockchain system.

Block: A block is a data structure that stores/records a collection of transactions before being shared (distributed) among all network nodes.

Chain: A chain is a collection of blocks that are arranged in a specific order.

Miners: Miners are the individual nodes that validate blocks before adding them to the blockchain structure. Consensus Algorithm: When performing blockchain activities, it is a set of rules and processes that must be followed to the letter.

a) BLOCK

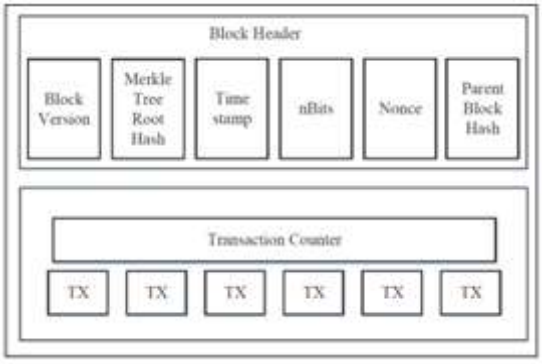


Fig. 2: Block Structure

(i) Block version: specifies which set of block validation criteria should be applied.

(ii) Merkle tree root hash: the sum of all transactions in the block's hash value.

(iii) Timestamp: from January 1, 1970, the current time has been expressed in seconds in universal time.

(iv) NBits: a valid block hash's goal threshold.

(v) NOnce: a 4-byte field that starts at 0 and rises with each hash computation.

(vi) Parent Block: A 256-bit hash value that points to the previous block is called the parent block hash. The block body is made up of a transaction counter and transactions. The block size and transaction size define the maximum number of transactions that may be contained in a block. Blockchain uses an asymmetric cryptographic method to verify transaction authenticity. A digital signature based on asymmetric cryptography is employed in an untrustworthy environment.

ISSN 2394 - 7780

b) DIGITAL SIGNATURE

Each user has a pair of private and public keys. A private key is used to sign the transactions, which must be kept secret. The digitally signed transactions are broadcast throughout the network. The signing phase and the verification phase are the two steps of a typical digital signature. User Alice, for example, wishes to send a message to user Bob. (1) During the signature step, Alice encrypts her data with her private key and sends both the encrypted and original data to Bob. (2) In the verification phase, Bob validates the value using Alice's public key. Bob could simply inspect the data to see whether it had been tampered with in this way.

c) Characteristics of Blockchain Architecture

Data encryption: The blockchain architecture ensures that all transactions have the highest level of trust, validation, and verification for all parties.

Tamper-proof: No record can be tampered with because of the transparency.

Traceable to source of origin: Any transaction may be easily traced back to its source of origin because every step of the process is meticulously tracked within the system.

Anonymity: Each node or user has a self- generated address that protects the participant's true identity in the blockchain system.

Transparency: Blockchain eliminates any opportunities or dangers of corruption of the architecture and weakening the highly influential computation by the systems involved due to its pure transparency and see-through procedures.

d) About Blockchain systems

In a public blockchain, all records are visible to the public, and anybody may participate in the consensus process. In contrast, a consortium blockchain allows just a small number of pre-selected nodes to participate in the consensus process. Only nodes from a single organisation can participate in a private blockchain.

The firm would be allowed to take part in the consensus-building process. A private blockchain is classified as a centralised network since it is entirely controlled by one firm. The consortium blockchain developed by multiple companies is partially decentralised since only a small number of nodes will be chosen to decide the consensus.

V. Blockchain Application and its S.W.O.T Analysis

1. Cryptocurrency

The financial sector, particularly in the sphere of One of the most active areas of blockchain is cryptocurrency. Since the first carrier bitcoin was added to the blockchain, a variety other cryptocurrencies have emerged. Bitcoin's anonymity, verifiability, decentralisation, and consensus mechanisms have made it a popular cryptocurrency its value has skyrocketed to an incredible $6,300 per BTC. Simultaneously, new coins with enhanced characteristics have arisen, forming the current thriving cryptocurrency market. Ethereum One of these was, which released a public blockchain platform in 2015 on which smart contracts could be deployed. Blockchain technology is currently being utilised in a wider range of commercial applications, such as contract processing, thanks to the emergence of contracts.

S.W.O.T Analysis of using Blockchain for Cryptocurrency	
Strength	**It is impossible to track or steal cryptocurrency.** Between the sender and the receiver, Bitcoin uses a blockchain (a peer-to-peer) network. There are only these two parties engaged. It differs from any other mode of currency transfer that includes a third party, such as a bank. Bitcoin transactions do not allow for the use of an intermediary. Bitcoin is a tax-free money since that pesky third party does not exist. Bitcoin is not regulated or controlled by the government.
Weakness	**Slow transactions and a loss of accessibility are crippling.** Transactions on Bitcoin aren't as quick as they were a few years ago. One of the disadvantages of Blockchain is that the more individuals that utilize it, the slower transactions become. The blocks essentially grow in size as the system is used more. Making the entire process clumsy and time-consuming.

ISSN 2394 - 7780

Opportunities	**Data safety from being breached** Brands like Facebook and Wells Fargo are notorious for data breaches involving user information. The blockchain is a fantastic technology with a lot of potential. Data such as criminal histories, birth certificates, and public records may be kept secret using the blocks. It has the potential to pave the road for unbreakable encryption. For data protection, that is something that the general public is trending toward.
Threat	**The anonymity in the midst of governments and banks** Anonymous purchasing can be hazardous in the wrong hands. Criminals will be drawn to the transaction since it is untraceable. Because the more people who accept Bitcoin, the more likely it will be utilized for criminal purposes. After all, it will be an issue for the government or law enforcement.

2. Internet of Things

All interactions are recorded on the blockchain, which allows for an immutable record of transactions. This technique ensures that all chosen interactions are traceable because their data can be searched in the blockchain, and it also increases IoT device autonomy. Slock is an example of an IoT application that aims to trade or rent. It may be able to use this strategy to supply

their services. Nonetheless, documenting all interactions in blockchain would necessitate an increase in bandwidth and data, which is one of blockchain's well- known problems. All IoT data linked with these transactions, on the other hand, should be recorded in blockchain.

S.W.O.T Analysis of using Blockchain for Internet of Things (IOT)	
Strength	**No Centralized Network:** The decentralized network of Blockchain is the most essential feature that makes it a good IoT security solution. **Extendable Database**: Because of the large number of nodes in a network, a large amount of data may be stored.
Weakness	**Compromise of Data is a Possibility:** Because it is not completely impenetrable, blockchain is not a perfect security solution. **Massive Processing Power is Required:** It is critical that the nodes have enough processing power and time to design and run encryption algorithms.
Opportunities	**Secure Communication amongst IoT Devices:** The Internet of Things can benefit from secure communication in an IoT-powered network by leveraging the capabilities of Blockchain technology. After all, Blockchain technology would treat communication as a transaction, ensuring its integrity in the same way that it would a transaction.
Threat	**Tax Issues**: While Blockchain has many advantages in the IoT environment, fully implementing the technology will be difficult and time-consuming. For example, there are legal concerns that must be addressed prior to the technology's cross-border adoption.

3. Healthcare

As new business cases develop, Blockchains are also attracting interest from the healthcare industry. Disintermediation, transparency, auditability, industry collaboration, and new business models are all examples of new business models are all features of the blockchain that healthcare businesses are interested in. The dispersion of medical records induced by transfers between medical facilities has become a major issue.

.	**S.W.O.T Analysis of using Blockchain for Healthcare**
Strength	**Efficiency in terms of costs** Rapid access to medical information Distribute information that is temper-proof
Weakness	**System providers have a smaller number of software options.** There isn't much scalability. A lack of large-scale storage capacity
Opportunities	**Lowering the risk of fraud and improving the medical supply chain** Beneficiaries have more control over their personal information. In the healthcare industry, there is a lot of potential for entrepreneurs and forged partnerships.

ISSN 2394 - 7780

	Data anonymity will aid medical research.
Threat	**Technology's hesitant societal adoption** There is no standardization To utilize blockchain for sensitive data, there are cultural and trust concerns.

4. BANKING

Traditional securities are giving way to high-tech securities in the banking business. Industry has begun to experiment with blockchain by recreating current asset transactions on the blockchain. This does, however, allow some room for the blockchain solution to succeed. The main advantages of Blockchain technology in the banking business are that it enhances efficiency, security, and transparency, which offers immutable records, enables

The blockchain offers the possibility of creating a platform for secure recording. Integrating highly fragmented healthcare information with the blockchain can provide a means to track personal health records. Access to medical records, on the other hand, raises ethical concerns. laying the groundwork for a high-integrity tracking capability is a big difficulty with this type of application

for quicker transaction times, and reduces costs by removing the requirement for third-party involvement. The unchangeable transaction history is one of the most important characteristics of blockchain.

Any purchases that have already been made are final and cannot be reversed. Many crimes against financial institutions will be reduced as a result of this. Contracting parties agree to deal with one another.

S.W.O.T Analysis of using Blockchain for Banking	
Strength	Efficiency in Operations Makes it easy to share information about specific products or trades. There are no more documents to be forwarded along. Everything can now be recorded on the blockchain. Encryption and data storage that is tamper-proof Removes the central authority that has complete control over the data.
Weakness	Business rules change on a regular basis, but blockchain does not. The majority of blockchains aren't modular. It is difficult to replace an obsolete encryption module. What if business regulations change and we need to export data to a new blockchain with updated data models? A blockchain does not provide an immediate escape plan out of the box. Existing approaches to regulatory compliance may be in contradiction.
Opportunities	Provides a Big Data and analytic research platform. Returns control to the user; for example, rather than Google and Facebook using your data, you can choose who has access to it. The blockchain will store all of these permits. As the world becomes more digital, more individuals will accept blockchain as a reality in their daily lives.
Threat	The problem of scalability is that there are too many transactions (overload), despite the fact that there are various solutions available. Mining pools and big mining farms are examples of unwelcome centralization. Quantum computers have the power to decode data (in the future). In a fast-changing environment, there is a lot of hype. Mining and hacking attacks are always a possibility.

VI. CONCLUSION

With blockchain promising solutions, this is about to change. for dealing with both concerns about security and integrity In this article, we reviewed the most recent studies on the usage of blockchain technology in a number of industries. We also investigated a number of commercial implementations of To provide an abstract, blockchain will be used in Industry 4.0 and IIoT. In practise, this is a metric of adoption. Blockchains have sparked a lot of attention around the world. In this paper, we examine the use of blockchain in a variety of fields, including cryptocurrencies, healthcare, Banking and Internet of things uses are all possibilities.

Our research article presents a for anyone interested in blockchains, this is a timely summary. Furthermore, the discussion will encourage blockchain uses in a wider range of fields. We also talked about the obstacles that each of these industries is facing when it comes to utilising blockchain. While several blockchain initiatives

ISSN 2394 - 7780

have developed in recent years, blockchain research is still in its early stages. In order for blockchain to be totally useful and adaptable, more industry-oriented research needs be done to solve some of these concerns, including personal data. Security, block data scalability, data confidentiality and privacy of participating organisations, blockchain integration and adoption costs, and regulatory compliance are all factors to consider.

VII. REFERENCE

[1] "Blockchain Technology for the Advancement of the Future" by Quoc Khanh Nguyen, Quang Vang Dang

[2] "Application of Blockchain to the Next Generation of Cybersecure Industry 4.0 Smart Factories" by Tiago M. Fernández- Caramés1, And Paula Fraga-Lamas1

[3] "Blockchain Applications for Industry 4.0 and Industrial IoT: A Review"

[4] S. Malik, S. S. Kanhere, and R. Jurdak, ''ProductChain: Scalable blockchain framework to support provenance in supply chains,'' in Proc. IEEE 17th Int. Symp. Netw. Comput. Appl. (NCA), Cambridge, MA, USA, Nov. 2018, pp. 1–10

[5] X. Ye, Q. RTShao, and R. Xiao, ''A supply chain prototype system based on blockchain, smart contract and Internet of Things,'' Sci. Technol. Rev., vol. 35, no. 23, pp. 62–69, 2017.

[6] P. Baipai. (2017). How IBM and Maersk Will Use the Blockchain to Change the Shipping Industry. [Online]. Available: http://www.nasdaq. com/article/how-ibm- and-maersk-will-use-the-blockchain-to- change-theshipping-industry-cm756797

[11] J. Nation. An Ov\erview of Blockchain Technology: Architecture, Consensus, and Future Trends Zibin Zheng1 (2017)

[7] Blockchain in Industries: A Survey Jameela Al-Jaroodi1 , Member, IEEE, Nader Mohamed2 , Member, IEEE 1Department of Engineering, Robert Morris University, Moon, PA 15108 USA 2Middleware Technologies Laboratory, Pittsburgh, PA 15057 USA

[8] A Survey of the Blockchain Applications in Different Domains

[9] A Next-Generation Smart Contract and Decentralized the Application Platform https://github.com/ethereum/wiki/wiki/Whi te-Paper

[10] Blockchain Whitepaper https://static.huaweicloud.com/upload/files/ pdf/20180411/20180411144924_27164.pdf

[11] Lin, I. C., and Liao, T. C. (2017). Survey of Blockchain Security Issues and Challenges. IJ Network Security, 19(5),653- 659.

[12] Blockchain: Opportunities for Health Care https://www.healthit.gov/sites/default/files/ 4-37 hhs_blockchain_challenge_deloitte_consult ing_llp.pdf

[13] Blockchain for Healthcare https://www.ehidc.org/sites/default/files/res ources/files/blockchain-for-healthcare- 341.pdf

[14] Will Blockchain Transform Healthcare? https://www.forbes.com/sites/ciocentral/2018/08/05/willblockchain-transform- healthcare/#2e63a56c553d

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

INCREASING RATE OF CYBER ATTACKS IN "COVID-19 PANDEMIC": A SYSTEMATIC REVIEW

[1]Smruti Nanavaty and [2]Pratik Shirsat
[1]Assistant Professor, Department of Information Technology, Technology, Usha Pravin Gandhi College of Arts, Science and Commerce, Science and Commerce
[2]M.Sc. (I.T.) Student, Department of Information, Usha Pravin Gandhi College of Arts,

ABSTRACT
In the past decade and during the covid-19 pandemic, it was been observed through websites and news that there was a huge rise in cybercrimes. Criminals are taking undue advantage of pandemic and are attacking the network systems, computer systems of individual which has given rise to cyber-attacks. A widespread of coronavirus across the globe has been a boon to all the attackers and through this, it has been easy for them to access the vital information of the users, individuals and organizations. Many malicious attacks have increased and many organizations have faced tremendous financial loss. This review focuses on how covid-19 pandemic has increased the rate of cyber-attacks and cybercrimes across the globe and why there is a need to establish enterprise risk management plan, strategies and how the value and importance of cyber insurance in covid-19 pandemic has increased.

Keywords: Cyber-attacks, cybercrimes, risk management, pandemic, covid-19, cyber insurance

I. INTRODUCTION
The covid-19 pandemic has brought the significant changes in the world and in everybody's personal and professional life. Due to the pandemic, it is clearly evident, that the high number of cyber-attacks has majorly disrupted all the services and sectors, thereby affecting all the business organizations and firms all over the world. We as humans are increasing our expectations from machines to be smarter, smaller, portable, faster, and efficient and make our day to day work easier.

Covid-19 has drastically changed whole scenario for the working of organizations, schools, hygiene habits etc. It was seen since last march, 2020 in India that all the organizations, schools, colleges, offices are providing the service of work from home (WFH), schools and colleges are using online education approach. The major rise in cyber-crimes and cyber-attacks is because of the work from home (WFH) services. All the employees are working from home today during this crisis. Because of work from home services, cyber-attacks have resulted in damage of financial loss, service disruption, fetching the vital credentials etc. All the data was safe and was on secure VPN connection before the pre-corona period but as organizations and all the firms have moved away from their physical organization space there has been a huge rise in cyber-attacks. This has affected all the sectors such as IT, HR, Finance, Marketing, Supply change management and so on. Attackers and criminals are using new techniques as it is easy for them to fetch the data through Wi-Fi, hotspot network lines which are used by everyone while working on the remote session.

Many organizations have been victims of cyber-attacks in pandemic. Some of them have recovered themselves from the attack whereas some have planned and implemented a risk assessment and mitigation strategy on remote access.

II. LITERATURE REVIEW
The following is the literature review on rising cybercrimes in last decade and in pandemic period:
Cassim F. [1] had mentioned in the year 2009 in the article formulating specialized legislation to address the growing spectre of cyber crime: A comparative study author looks at the cyber legislation formulated to address cybercrime in the United States of America, The United Kingdom, Australia, India, The gulf Countries and South Africa. Susheel B and Durgesh P [2] described in their paper about the Study of Indian Banks Websites for Cyber Crime Safety Mechanism that security plays a vital role in implementation of technology specially in banking sector. Ambika Choudhary, [3] a technical journalist had mentioned in her blog the top ransomware attacks of cybersecurity in the year 2020 that shook the internet. The following are the majorly known ransomware attacks:

- Cognizant Ransomware Attack
- University of California San Francisco (UCSF) Ransomware Attack
- Canon Attack

ISSN 2394 - 7780

- Carnival Corporation Suffers Ransomware Attack
- Magellan Health Ransomware Attack
- Communications & Power Industries (CPI) Ransomware Attack
- Baltimore County Public Schools Attack

Victim Firm	Month/ Date	Data Loss/Data Breach	Cost for Remediation
Cognizant	April	Network Connection line	$50 to $70 Million
UCSF	June	Attack occurred in a limited part of the UCSF School of Medicine's IT environment.	$1.14 Million
Canon Attack	June	Stole the data from the company servers.	---
Carnival Corporation	August	Unauthorised third party gained access to certain personal information.	Approx. $2.10 Million
Magellan Health	April	Data breach, Hacker got access to health plan's servers.	$1.7 Million
Communications & Power Industries (CPI)	January	A domain admin clicked on a malicious link while they were logged in that immediately triggered the file-encrypting malware	Ransom of $500,000 was paid.
Baltimore County Public Schools Attack	November	All pupils learning remotely because of the pandemic could suddenly no longer access lessons due to attack.	---

Fig. 1. Ransomware attacks in year 2020

Apart from the ransomware attacks, many organizations have faced a phishing attack and one of the famous and majorly known phishing attack was of NDTV journalist Nidhi Razdan [4] who had tweeted on 15th January, 2021 describing for being the major target of a very serious phishing attack and had learnt that the offer she had got from Harvard University of Massachusetts was fake and an attacker had taken advantage of covid-19 pandemic. It was also stated that by Nidhi Razdan that, "After hearing from the university, I have learnt that I have been victim of a sophisticated and coordinated phishing attack. I did not receive an offer by Harvard University to join their faculty as an associate professor of journalism." As mentioned in one of the articles of outlook magazine[14] it was said that India's cyber security agency, CERT-In warned that the cyber attackers were targeting the banking customers using a phishing attack in order to get track of the sensitive information such as mobile number, OTP, internet banking credentials etc. A unique web application, ngrok platform was used for carrying out the malicious activities. Users were attacked by receiving a SMS containing the malicious link which ended with ngrok.io/xxxbank. After clicking on this link, the attacker generated an OTP which user entered on his device thus by attacker gaining the access to user's device and performing fraudulent activities.

III. METHODOLGY

As observed through newspaper and websites, covid-19 pandemic has forced everyone to adopt the technique of working from home and thereby shifting from office environment and secured LAN connection network to home Wi-Fi, hotspot etc. This has given a massive increase in number of cyber security crimes. The research methodology indicates about the increase in rise of cyber-attacks and how it has affected the individuals across the globe. [5] As covid-19 virus is increasing day by day, attackers are using different malwares and new techniques to steal the credentials of the customers. A coronavirus tracker website or links are created by hackers to spread the malware and also phishing emails are being sent to people related to covid-19 information.

Cybercrimes are constantly evolving and increasing challenges for all the business organizations today. According to research, some of the cybercriminals are posing as government entities providing financial aid those who are in need. Such type of a phishing attacks has increased in this crisis period. A typical phishing attack begins with inappropriate mails and once the user access or open the mail, the whole data gets transferred on criminal's platform. The confused or not so educated users about cyber-crimes may click on the email link and becomes a victim of phishing attack. It was also observed that most of the victims of ransomware attacks spend a lot of money for investigating the attack. The main reasons for the rise of cybercrimes in covid-19 pandemic are as follows:

ISSN 2394 - 7780

- Shared connection network (Home WIFI)
- Portable device such as dongle which may not have secured connection line.
- Using the mobile hotspot. (Many times it is observed we get a message while connecting to hotspot like. "Network provider might get to know the content you share")
- Hostel or People living in PG's might share a same network connection who works for different organization.
- No proper LAN network connection which is established in office areas.

Cybercrimes will also rise in second half of 2021 as even today, WFH, Online schooling techniques are used because of massive rise in corona virus. In order to secure the data and information from cyber criminals the following measures are to be adopted by everyone which are as follows:

- Avoid clicking on unknown links appearing in the inbox section
- Do not use any social media applications on laptops and completely avoid usage of it on organization's laptop.
- Firms/Organizations should carry out the Information Security awareness Training every quarter to inform all the employees about the cybercrimes.
- Do not share or submit any password through google forms.

[6]World Health Organization (WHO) is aware of suspicious emails and phishing attacks and attackers/hackers are taking advantage of covid-19 pandemic. The emails contain the details related to WHO and appear to be from the same organization. The email ask to give the sensitive credentials of the users, click a malicious email link which can fetch the passwords and username when a link is clicked or open an attachment which is attached to that particular phishing email.

Cyber criminals use the phishing attack technique to control the people's devices and access their data and vital information, which could lead to draining of bank balance in a bank account said by an official spokesperson of the [11]Punjab Bureau of Investigation.

Disruption of services in the organization has led to a negative impact on the companies due to attacks which have occurred in pandemic. [8]AZORult was used to download ransomware programs All the attackers are taking the coronavirus pandemic as an advantage to attack the personal data from health related sections to global shipping industry, economic as well as secondary section.

A. Risk Management Framework

Risk management is a process which helps an organization to identify risk, categorize the risk as per risk matrix (severity and impact) and helps to mitigate the risk. As attackers have taken huge advantage of covid-19 pandemic, establishing an enterprise risk management plan and implementing it has equally become an utmost need for all the organization. It is a vital process for all the business as it increases the revenue and prevents financial losses which are occurred by risk. Implementing an effective risk management plan in firm improves and increases the success rate and opportunities and minimizes cyber threats.

The requirement of risk management is necessary as it will help to secure the information and data and if it gets in contact with any threat then it will not affect the organization's data. The risk-informed strategy refers to the treatment of risk avoidance, reduction, transfer and retention using risk assessments in an absolute or relative way. [10]A risk can be defined in the following was:

- An occurrence of an unfortunate situation.
- Exposure of some part of organization's data to the hacker
- Identification of a negative event through phishing emails.
- The repercussions of an unwanted activity and associated uncertainties with it.
- Probability of an activity which is unavoidable.
- Identify suspicious pop-ups and reporting a slower than normal network.
- Noting of unusual password activity which mentions that your password has been changed.
- Encryption of files by blocking the access of the owner.
- Owner's contacts letting the owner know that they have received strange messages and emails from the owner.

ISSN 2394 - 7780

Fig. 2. Risk Management Framework

A risk management framework consists of 5 major factors which are as follows:

1. **Identification of Risk:** To identify potential risk and is one of the critical steps in risk management framework.
2. **Risk Assessment:** Assessing the risk with its severity level and impact it has made to an asset of a firm.
3. **Response plan / Strategy:** Planning and creating a cyber strategy or plan to deal with the unwanted risk.
4. **Risk Mitigation:** Implementation of risk management plan helps an organization to mitigate the risk in a correct way.
5. **Monitoring Performance:** To monitor the performance of an organization post cyber-attack incident.

Risk analysis is defined as a function of impact and likelihood. These are the factors which help an organization to measure and analysis the impact risk has made and what is the severity of the risk. Documentation of a risk matrix is necessary in order to mitigate the risk and monitor the performance of the systems after mitigation process. It helps to respond to the questions like how soon can we mitigate the risk, how much time is taken to recover from the damage, how much downtime is tolerated by the system etc. A risk control matrix has different type of scales which are useful for the interpretation of the risk. It is also useful in prioritizing the risk such as high, medium or low.

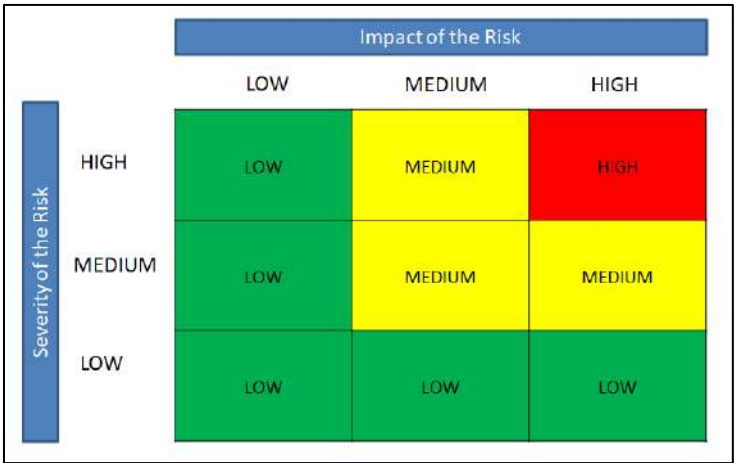


Fig. 3. Risk Control Matrix (3x3 matrix)

Risk environment in every organization is different and unique which completely depends on various factors such as type of the business, size, policies and procedures and rules and regulations. Risk environment of an organization can be better understood with the help of risk control matrix. Each and every organization should build and document a risk management policy and procedure which can help an organization to understand the risk in a better way.[7]

ISSN 2394 - 7780

Problem Statement of Risk	Risk Rating	Observation and control description	Control Decision
Unauthorized access to the emails and data	High	Define access rights and access control matrix & policy	Preventive

Fig. 4. Risk matrix observation table

Defining and documenting a cyber crisis management plan (CCMP) also helps an organization to mitigate the risk in an efficient and effective manner.

[12] A cyber crisis management plan clearly defines the roles and responsibilities of all the employees in an organization. When a cyber incident occurs, communication team plays a pivotal role during the crisis along with CEO of the organization. An email has to be send by the IT service team/IT Service HelpDesk to all the employees of the company spreading awareness about the occurrence of cyber incident.

B. Need for Cyber Insurance

Cyber insurance or cyber liability insurance is created to help organizations to reduce the financial risk which is associated with the online business and transactions. Financial loss caused by ransomware, distributed denial-of-service (DDoS)[13], malware attacks can be retrieved with the help of insurance. Cyber liability coverage helps with cost associated with remediation and communication, costs incurred due to incident, costs associated with lawsuits, legal fines and cost related to response and recovery process. It is considered as the cyber-risk mitigation measure and is becoming popular amidst the pandemic. The rise of coronavirus identified the need that enterprises must increase corporate resilience and help ensure community well-being by embracing virtual collaboration tools and practices.

Business organization that create, store and manage electronic data online, such as users credentials, customer contacts, passwords, customer sales, PII and credit card numbers have utmost need of cyber insurance as loss. Industries such as finance and healthcare needs cyber insurance as those are most commonly target by cyber criminals and hackers. If the organization is bigger in size then there are high chances of risk and it stands to reason that there is requirement of wider scope of coverage (larger the size of an organization, the bigger the risk). Therefore, having cyber insurance in place is extremely vital as it helps to recover the loss caused by cybercrimes.

IV. CONCLUSION

As cybercrimes are increasing day by day, the systematic study shows that people across the globe are unaware about the rising cyber-attacks and the consequences they have to face in case of a breach. It is necessary to educate all departments of the company about rising cybercrimes and how to prevent it by conducting cyber crisis management seminars and plans. It is vital to develop a cyber risk framework in order to mitigate the risk. People should avoid clicking on malicious links as it can directly access the credentials and other vital information of the user. Further study will focus on reviewing the techniques to improve the robustness of the system and a systematic review to bolster the cybersecurity framework.

REFERENCES

[1] Cassim F, "Literature Review on Cyber Security" pp. 2, 2009,

[2] Susheel B, Durgesh P, "Literature Review" Cyber security, pp. 4, 2011

[3] Ambika Choudhary, "Top Ransomware Attacks of 2020 that shook the internet", Ransomware attacks in the covid-19 pandemic, 2020,

[4] Nidhi Razdan, "Victim of Phishing Attack", No offer from Harvard University, 2353118, Ndtv.com website, India-news, January 16, 2021.

[5] thesslstore.com, " coronavirus-scams-phishing-websites-emails-target-unsuspecting-user" march 02, 2021.

[6] who.int, “Beware of criminals pretending to be WHO”, World Health Organization, About WHO, cyber criminals.

[7] MASD & Co. https: //taxguru.in/chartered-accountant/risk-control-matrix.html, 27th May, 2020

[8] techrepublic.com, “global shipping industry attacked by coronavirus themed malware”.

[9] Ciscoeconomictimes,“Covid-19 related Phishing attack up by 667 report”, 74839322.

[10] Terje Aven, “Risk assessment and risk management: Review of recent advances on their foundation” university of Stavanger, ullandhaug, 4036 stravanger, Norway.

ISSN 2394 - 7780

[11] hindustantimes, "Beware of sms claiming to provide free covid relief package', July 25 2020, Chandigarh.

[12] Playbook_CyberCrisisCommn_220719, Reserve Bank Information Technology Private Limited (ReBIT)

[13] Cisco.com, "www.cisco.com", security solutions, cyber-insurance.

[14] Outlookindia.com, "a new phishing attack: cyber security agency warns banking customers in India", August 12, 2021.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

DATA ANALYSIS FOR THE PREDICTION OF HEART DISEASES

Neelam Naik and Srushti Shah
Usha Pravin Gandhi College of Arts, Science and Commerce, Vile Parle, Mumbai, University of Mumbai, India

ABSTRACT

The leading cause of death in the world is heart disease. Many advanced technologies are used to treat heart disease. The most common problem in medical centres is that many medical staff members don't have equal knowledge and expertise to treat their patients, so they deduce their own decisions, which result in poor outcomes and sometimes fatalities. It is now possible to use machine learning algorithms and data mining techniques to automatically diagnose heart disease in hospitals, which plays a crucial role in overcoming these problems. Various health parameters can be analysed to predict heart disease. Different algorithms are available for predicting heart disease. In the present study decision tree algorithm is used for predicting the possibility of heart disease based on patient's health condition.

Keywords: Heart disease, prediction, Decision tree, Naive Bayes classifier, data analysis.

I. INTRODUCTION

Cardiovascular disease (CVDs) consists of a group of heart and blood vessel disorders that is responsible for more than half of all worldwide deaths. CVDs are one of the leading causes of death worldwide, accounting for approximately 32% of global deaths in 2019, according to the World Health Organization. The major risk factors associated with multiple vascular phenotypes are typically behavioural in nature, such as sedentary lifestyle, tobacco use, unhealthy diet, and alcohol abuse. However, multiple vascular phenotypes may also be strongly related to a genetic background. Aiming to identify, diagnose, and utilize genetic biomarkers for the treatment of cardiac patients is precision cardiology. As non-invasive methods of genetic testing become more prevalent and the outlook toward genetic testing becomes more positive, non-oncology application areas are slowly becoming more of a focus. These are mainly infectious diseases and cardiovascular diseases.

A group of heart and blood vessel disorders known as cardiovascular diseases (CVDs) contribute to most deaths worldwide. According to WHO these chronic diseases include:

- Coronary Heart Disease: A disease of the blood vessels in the heart.
- Cerebrovascular Disease: A blood vessel disorder that affects the brain.
- Peripheral Arterial Disease: A disease affecting the arms and legs.
- Rheumatic Heart Disease: damage to the heart valves and the heart muscles caused by streptococcal bacteria.
- Congenital Heart Disease: Birth defects affecting the normal development and working of the heart.
- Deep vein thrombosis and pulmonary embolism: The formation of blood clots in leg veins that can spread to the heart and lungs.

The leading cause of death worldwide is heart disease. More people die from cardiovascular disease per year than from any other cause, with 12 million people dying every year from heart disease. One person dies from heart disease every 34 seconds in the United States.

Heart attacks are tragic events resulting from a blockage in the flow of blood to the heart or brain. Blood pressure, glucose, and lipid levels are elevated in people at risk of heart disease. The basic health facilities can measure these factors at home with relative ease.

Heart disease is categorized into three types: Cardiomyopathy, Cardiovascular disease, and coronary heart disease. A "heart disease" refers to any condition that affects the heart and blood vessels, and how fluids are released into and circulate in the body. Cardiovascular diseases (CVD) cause a wide range of illnesses and disabilities. It is one of the most important and difficult aspects of medical diagnosis.

The automation of this task is very helpful in medical diagnosis, which is considered as a critical, but difficult, task. Unfortunately, not all physicians are experts in all areas of medicine, and there are some places where there are few resources. A data mining program can help uncover hidden patterns and knowledge that can aid in making good decisions. Health professionals use this information to make accurate decisions and provide quality care to the public. It is also important that health care organizations provide assistance to professionals

ISSN 2394 - 7780

who are lacking in knowledge and skills. Current methods have a major limitation in their ability to draw accurate conclusions. For the prediction of heart disease based on some health parameters, we employ different data mining techniques and machine learning algorithms, including Naive Bayes, k Nearest Neighbour (KNN), Decision Tree, Artificial Neural Network (ANN), and Random Forest.

Based on a study from 2016, humans generated data in excess of ten exabytes, or 5x1018 bytes from various sources in 2016 (Lyman and Varian 2003). An analysis of exploratory data (EDA) is a technique that reveals hidden structure within a dataset, enhances insight into a dataset, identifies anomalies, and builds parsimonious models to test the underlying assumptions. There are two types of exploratory data analysis (EDA): Graphical or non-graphical and Univariate or multivariate. Univariate data consider one column of data at a time while multivariate methods consider more than two variables. There are two types of diagnostic methods for diseases, invasive and non-invasive

Heart disease is the biggest challenge in the medical industry, and it is based on factors like physical examinations, symptoms, and signs of the patient. A number of factors can influence the risk of heart disease, including cholesterol levels in the body, smoking habits, obesity, family history of diseases, high blood pressure and workplace conditions. In predicting heart disease, machine learning algorithms are vital and accurate. Advances in technology have enabled machine language to be combined with big data tools to manage unstructured and exponentially growing data sets.

Numerous studies have been carried out and a variety of machine learning models have been used in order to predict and classify heart disease. Automatic classification can be used to identify high-risk and low-risk patients with congestive heart failure. Heart disease can be fatal and should not be taken lightly. Heart disease is more likely to affect men than women. The risk of suffering a heart attack is about twice as high for men as for women throughout life. The higher risk of heart disease persisted even after controlling for traditional cardiovascular risk factors such as high cholesterol, high blood pressure, diabetes, or body mass index. It is considered a benchmark dataset when one is working on heart disease prediction because it contains some important parameters, such as dates from 1998.

II. LITERATURE REVIEW

Ordonez [1] proposes that heart disease can be predicted from simple patient attributes. They have used a system that includes 13 simple attributes such as sex, blood pressure, cholesterol, etc., to predict whether a specific patient will suffer from heart disease. Research dataset has been expanded with two more attributes: fat content and smoking behaviour. Predictions are made by using data mining classification algorithms, such as Decision Trees, Naive Bayes, and Neural Networks. A Heart disease database is used to analyse the results.

In Paper [2] Researchers used data from Kaggle to apply knowledge discovery processes on prediction of stroke patients based on Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs), giving accuracy of 81.82% and 80.38% for ANNs and SVMs, respectively, for training datasets and 85.9% and 84.26% for Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs), respectively, for test datasets.

The multilayer perceptron with back-propagation technique was used by Shantakumar et al. [3]to create an effective and intelligent heart attack prediction system. As a result, the MAFIA algorithm is used to identify the incidence patterns of heart disease based on the extracted data.

Yilmaz and colleagues [4] have developed a method of determining the patient's condition by using least squares support vector machines (LS-SVMs) and binary decision trees.

Frawley et al. [5] investigated the probability of survival following coronary heart disease (CHD), which is one of the most difficult research challenges facing medical society. To determine the impartial estimate of the three prediction models for performance comparison, they also used cross-validation techniques with 10-fold repeatability.

This study found that artificial neural network (ANN) showed the best accuracy of 84.25 % as compared to other models. Despite the fact that other models had higher accuracy than ANN, this model with lower accuracy was chosen as a final model so that the accuracy of the model can be assessed without compromising the [6] accuracy of the prediction.

Naive Bayes algorithm is a highly accurate algorithm that can predict heart disease and outperforms hidden naive Bayes in terms of [7] accuracy.

According to Lee, et al.[8,] a novel methodology for expanding, studying, and analysing the multi-parametric and linear features of Heart Rate Variability diagnosing cardiovascular disease has been developed. The

ISSN 2394 - 7780

researchers have experimented with linear and non-linear features to estimate various type of classifiers, e.g., Bayesian classifiers, CMAR, C4.5, and SVM. Their experiments showed that SVMs were more effective than other classifiers.

We developed a machine learning algorithm to predict cardiovascular disease through Random Forest, Decision Tree SVM (support vector machine), and KNN, while Random Forest was found to have the highest [9] accuracy of 85%.

Noh, et al. [10] developed a method of classifying features based on an associative classifier constructed with the efficient FPgrowth method. In our pattern-generating process, the volume of patterns may be diverse and extremely large, so they proposed a rule to measure the cohesion and, in turn, allow a tough choice of pruning patterns.

In paper [11], we use data from the UCI repository to compare the performance of Naive Bayes, KNN, Decision Tree, and ANN machine learning algorithms. ANN had the highest accuracy of 85.3%. Meanwhile, Naive Bayes, KNN, and Decision Tree gave an accuracy of almost 78% each.

According to Parthiban et co. [12], the proposed Coactive Neuro-Fuzzy Inference System (CANFIS) can be used to identify and predict heart disease. They use genetic algorithms with fuzzy logic to diagnose occurrences of diseases based on the collective nature of neural networks. We evaluated the performance of the proposed CANFIS model on the basis of classification accuracy and training performance. In conclusion, they demonstrate that the proposed CANFIS model is highly effective in predicting heart disease.

Yanwei, et. al [13] developed a classification method using multi parametric features obtained from HRV (Heart Rate Variability) of an ECG and the data is pre-processed before a heart disease prediction model is built to classify a patient's risk of heart disease.

Singh et al. [14] Using two clustering algorithms, constructed a clustering matrix combining a partitioning algorithm (K-Means) and a hierarchical clustering algorithm (agglomerative). With large data sets, the K-means algorithm is more effective and scalable, and it converges quickly. In hierarchical clustering, smaller clusters are frequently merged into a larger cluster, or a larger cluster is divided between smaller clusters to form a hierarchy of clusters. Based on accuracy and running time, they calculated the performance of K-means and hierarchical clustering algorithms using WEKA data mining tool.

Researchers developed a computational model by Guru et al. [15] to identify five major heart diseases using a three-layer multilayer perceptron model. A back propagation algorithm is used with an adaptive learning rate, a momentum term, and forgetting mechanics to train the proposed decision support system.

The data used in Paper [16] comes from the UCI data repository. Propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perception using WEKA software. Based on performance from different factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieves optimum performance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms is still not satisfactory.

The paper [17] uses WEKA to evaluate the performance of several machine learning algorithms. ANN and PCA were combined to increase the speed of the analysis. Before application of PCA, it shows 94.5% accuracy, but after application of PCA, it shows 97.7% accuracy. There is a huge difference.

A study by Duff, et al. [18] involved 533 patients who had suffered cardiac arrests and their results were integrated into analyses of heart disease probabilities. Analyses were performed using a variety of Bayesian networks, as well as classical statistical analysis.

Using several data mining techniques such as Decision Trees, Naive Bayes, and Neural Networks, Palaniappan, et al. [19] conducted a study and developed a model called Intelligent Heart Disease Prediction System (IHDPS).

III. RESEARCH METHODOLOGY

To predict the emergence of heart disease in order to detect it at an early stage in a short period of time is the main purpose of the proposed method. To model heart disease from certain health parameters, we used different data mining techniques and machine learning algorithms, such as Naive Bayes, K Nearest Neighbour (KNN), K- Means Clustering, Decision Trees, Artificial Neural Networks (ANN), and Random Forests.

ISSN 2394 - 7780

Based on publicly available heart disease data, the analysis is conducted. Twenty-nine records are contained in the dataset, with 8 attributes including age, type of chest pain, blood pressure, blood glucose level, ECG in rest, and heart rate.

A raw dataset of 303 Cleveland heart disease patients with 76 features is used in defining the proposed algorithm. Data inconsistencies are eliminated as part of the pre-processing method to eliminate errors. 209 samples are analysed for seven independent factors including age, chest pain type, blood pressure, blood glucose level, resting ECG, resting heart rate, four types of chest pain, and the habit of exercising. As coronary fatty streaks start to appear in the adolescent stage, age is the most important risk factor for heart disease. Therefore, only data for males is considered here since males have a higher risk of coronary diseases than females. Angina results from oxygen-poor blood not reaching the heart muscles sufficiently. Heart disease is closely related to high blood pressure because it damages arteries. Diabetes can significantly increase the risk of heart disease when high blood pressure is present. High blood pressure and high heart rate can increase the risk even more. Coronary disease risk directly relates to heart rate. Feeling gripped and tight is a symptom of heart disease, often found on the chest but also spreading to the shoulders and abdomen. As a rule, there are four different types of anginas: atypical angina, typical angina, asymptomatic angina, and nonanginal pain.

DECISIONTREES

One of the supervised learning algorithms, a decision tree is an intuitive and easy-to-understand classification algorithm. A decision tree can deal with categorical as well as numerical data. The tree structure of a decision tree has nodes, branches, and leaves, in which each branch represents one attribute value, each internal node specifies a specific test, and leaves indicate the class where the outcome is predicted. Using the predictive attribute and the given rules, the classification starts at the root node and moves to the leaf nodes. Classification and Regression Tree (CART), ID3, C4.5, J48, and CHAID are the most commonly used decision tree algorithms for disease prediction. The CART algorithm is an important decision tree algorithm that lies at the foundation of machine learning. Moreover, it is also the basis for other powerful machine learning algorithms like bagged decision trees, random forest, and boosted decision trees.

Naive Bayes classifier

A classifier based on Bayes' theorem is called naive bayes. According to the Naive Bayesian classifier theorem, particular features within a class exist independently of other features. Therefore, it can predict heart disease with high accuracy. Data sets are classified using conditional probability using the Naive Bayes method, which uses conditional probability to compute posterior probability. It follows this equation.

$(C|X)$=PX|C∗P(C)/P(X)

Where X is the instance to be predicted, and C is the class value for instance. The above-given formula or equation helps to determine the class in which feature expected to categorize.

IV. ANALYSIS OF DATA

Cleveland heart disease dataset (Source: https://www.kaggle.com/ronitf/heart-disease-uci) is used for the present study. The various features available in this dataset are described below.

Age -age in years

Sex- (1 = male; 0 = female)

Cp - chest pain type

Trestbps-resting blood pressure (in mm Hg on admission to the hospital)

Chol - serum cholestoral in mg/dl

Fbs - (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)

Restecg - resting electrocardiographic results

Thalach - maximum heart rate achieved

Exang- exercise induced angina (1 = yes; 0 = no)

Oldpeak- ST depression induced by exercise relative to rest

Slope- the slope of the peak exercise ST segment

Ca - number of major vessels (0-3) colored by flourosopy

ISSN 2394 - 7780

Thal - 3 = normal; 6 = fixed defect; 7 = reversable defect

Target - 1 or 0

Confusion matrix for CART algorithm is shown in table 1.

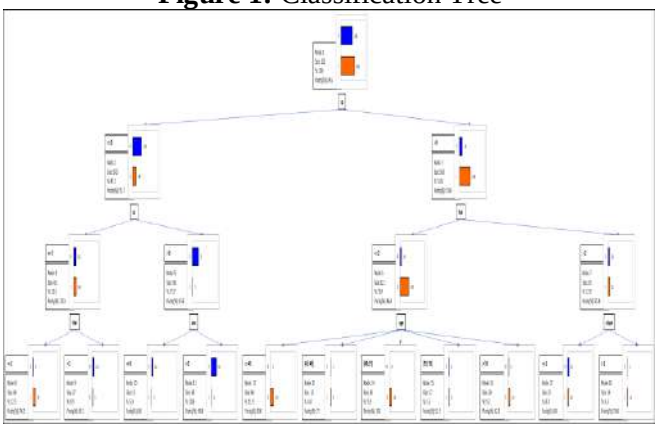
The total correct prediction is 84.16%. In this study authors have used data analysis techniques in healthcare to treat heart disease. The purpose of the experiment is to apply classification algorithm that gives accurate prediction of heart disease. From the analysis done we conclude that Classification and Regression Tree algorithm gives 84.16% of accuracy while predicting heart diseases.

Table 1: Confusion matrix for CART algorithm

from \ to	0	1	Total	% correct
0	110	28	138	79.71%
1	20	145	165	87.88%
Total	130	173	303	84.16%

The classification tree generated is shown in figure 1.

Figure 1: Classification Tree



V. FUTURE ENHANCEMENT

In order to predict heart disease before it causes to help physicians, data analysis is used to identify key patterns and features from the medical records of the patient. Other objectives include increasing the number of features in the data set and enhancing the accuracy of the prediction models.

VI. CONCLUSION

The leading cause of disability and premature death is heart stroke and vascular disease. Using risk factors alone to predict heart disease is difficult. Recognizing heart disease is primarily based on chest pain. This research work examines algorithms used in data analysis to analyse and predict heart disease. In this paper, four types of chest pain are used to predict heart diseases using major factors. Supervised machine learning algorithms like CART decision tree algorithm is used in this study, which provides 84.16% of prediction accuracy.

VII. REFERENCES

[1] Carlos Ordonez, “Improving Heart Disease Prediction using Constrained Association Rules”, Technical Seminar Presentation, University of Tokyo, 2004.

[2] C. Beyene, P. Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, https://www.researchgate.net/publication/323277772_Survey_on_prediction_and_analysis_the_occurrence_of_heart_disease_using_data_mining_techniques,118(8):165-173 · January 2018

[3] Ersen Yilmaz and Caglar Kilikcier, “Determination of Patient State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree”, Master Thesis, Department of Electrical Electronic Engineering, Uludag University, 2013.

[4] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, “Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method”, Studies in Health Technology and Informatics, Vol. 107, No. 2, pp. 1256-1259, 2004.

ISSN 2394 - 7780

[5] Heon Gyu Lee, Ki Yong Noh and Keun Ho Ryu, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining, pp. 56-66, 2007.

[6] Hossam Meshref, "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach", https://www.researchgate.net/publication/338428682_Cardiovascular_Disease_Diagnosis_A_Machine_Learning_Interpretation_Approach, January 2019.

[7] Jabbar Akhil, Shirina Samreen, "heart disease prediction system based on hidden naïve Bayes classifier", https://www.researchgate.net/publication/309735105_Heart_disease_prediction_system_based_on_hidden_naive_Bayes_classifier, October 2016.

[8] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Intelligent Computing in Signal Processingand Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[9] Komal Kumar Napa, G. Sarika Sindhu, D. Krishna Prashanthi, A. Shaeen Sulthana, "Analysis and Prediction of Cardiovascular Disease using Machine Learning Classifiers", https://www.researchgate.net/publication/340885231_Analysis_and_Prediction_of_Cardio_Vascular_Disease_using_Machine_Learning_Classifiers, April 2020.

[10] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[11] Muhammad Usama Riaz, Shahid Mehmood Awan, Abdul Ghaffar Khan, "Prediction Of Heart Disease Using Artificial Neural Network", Https://Www.Researchgate.Net/Publication/328630348_Prediction_Of_Heart_Disease_Using_Artificial_Neural_Network. October 2018

[12] Niti Guru, Anil Dahiya and Navin Rajpal, "Decision Support System for Heart Disease Diagnosis using Neural Network", Delhi Business Review, Vol. 8, No. 1, pp. 1-6, 2007.

[13] Nidhi Singh and Divakar Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", Ph. D Dissertation, Department of Computer Science and Engineering, Barkatullah University Institute of Technology, 2012.

[14] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008.

[15] Shantakumar B. Patil and Y.S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol. 31, No. 4, pp. 642-656, 2009.

[16] Ujma Ansari, Jyoti Soni, Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", 258493784_Predictive_Data_Mining_for_Medical_Diagnosis_An_Overview_of_Heart_Disease_Prediction. March 2011Data Mining in Healthcare for Heart Diseases

[17] Umair Shafique, Irfan Ul Mustafa, Haseeb Qaiser, Fiaz Majeed, "Data Mining in Healthcare for Heart Diseases", https://www.researchgate.net/publication/274718934_Data_Mining_in_Healthcare_for_Heart_Diseases. March 2015.

[18] W.J. Frawley and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview", AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.

[19] X. Yanwei et al., "Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease", Proceedings of International Conference on Convergence Information Technology, pp. 868-872, 2007.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

COMPARATIVE STUDY ON VIRTUAL PERSONAL ASSISTANTS

[1]Smruti Nanavaty and [2]Qureshi Sumaiya Mohd. Ishaq
[1]Assistant Professor and [2]MSc.IT Student, Department of Information Technology, Usha Pravin College of Arts, Science and Commerce

ABSTRACT

In this new age of technology and innovation, AI and machine learning have made our lives much simpler. The benefits of these technologies can be found in a wide range of fields, including education, industries, e-commerce, etc., but communication is perhaps the most important. In this paper, we explore different types of voice assistants that are used in ordinary life. Virtual Personal Assistants (VPA) are one of the most successful products of Artificial Intelligence, since they allow humans to outsource their work to machines. As Modern Mobile Technology is currently gaining fame for the User Experience, as it is very easy to access applications and services no matter where you are in the world. Today, we are accustomed to the use of smartphones and other types of computers to perform the same functions using the internet. Here we compare some of the leading virtual personal assistants, including Siri, Google Assistant, Alexa, and Cortana. During testing, these applications were evaluated for their features of input and output accuracy and their ability to understand commands and questions outside their embedded structure. This sample gives as the idea that which assistant is in more use and how many people are satisfied with it. Due to the heterogeneity in the available personal assistants, the user may have difficulty selecting one. but result show high inclination towards siri and google assistant as the most used VPA and user satisfaction.

Keywords: *Artificial Intelligence, Natural Language Processing, Digital assistant, Microsoft, Cortana, Apple, Siri, Amazon Alexa, Google Assistant, Speech Recognition, Robotics, Internet of thing, Machine Learning*.

I. INTRODUCTION

Modern Mobile Technology has become extremely popular due to its exceptional User Experience, since it is extremely easy to access applications and services no matter where you are. Today, we are accustomed to the use of smartphones and other types of computers to perform the same functions using the internet. All thanks to Internet of thing (IoT), Cloud Computing and Artificial Intelligence.

The widespread use of smartphones has caused numerous voice assistants to emerge, such as Google's Assistant, Apple's Siri, Microsoft's Cortana and Amazon's Alexa. In order to provide services to the users, voice assistants use technologies such as Natural Language Processing (NLP),voice recognition, speech synthesis. Virtual Personal Assistants (VPA) are one of the most successful products of Artificial Intelligence, since they allow humans to outsource their work to machines. The functionalities of the personal assistant will vary, however, depending on how it's implemented and who's involved. Now, voice assistants can also be integrated into smart speakers, devices that have a microphone and speaker that let them communicate with users. During the study and survey, our main goal was to figure out what the true potential of this personal assistant is.

II. LITERATURE REVIEW:

A. Overview on Artificial Intelligence:

Since artificial intelligence emerged as an academic field in 1956, it has experienced several periods of disappointment, optimism, and funding loss, which were soon followed by new approaches, successes, and renewed funding. Several different approaches to AI have been tried and abandoned over the years, including modeling the brain, simulating human problem solving, and organizing large databases of knowledge. Machine learning has been dominated by highly mathematical statistical algorithms since the early 21st century. As a result, it has successfully solved an increasing number of challenging problems in both industry and academia. There are several subfields of AI research, according to the goals and tools they use. In most cases, AI is used to address problems like reasoning, representation of knowledge, decision making, strategy, learning, natural language processing, vision and the capability to move and control objects. General intelligence (having the skills to handle arbitrary problems) is the major objective of this field.

Artificial intelligence researchers have developed a range of problem-solving techniques designed to address these issues, such as artificial neural networks, formal logic, strategies utilizing probability, economics and statistics. Other fields that AI draws from include psychology, literature, philosophy, and computer science.

The following is a list of four possible purposes or definitions of artificial intelligence (AI), which classifies computer systems based on rationality and thinking vs. acting:

ISSN 2394 - 7780

Human Approach:

- Software capable of thinking like humans
- System that behaves like a human

Ideal Approach:

- Systems capable of rational thought
- Systems that act rationally

B. MACHINE LEARNING

The field of Machine Learning and Data Science has become increasingly popular. Computers can now solve real-world problems through machine learning and data science. In turn, complex mathematics creates ML algorithms that eventually produce ML systems. In this field, algorithms and data are used to mimic how humans learn, gradually improving their accuracy over time. There are many applications of machine learning, such as medical diagnosis, email filtering, voice recognition, and computer vision, when conventional algorithms are difficult or not feasible to develop. Methods, theory and applications related to mathematical optimization are relevant to the field of machine learning. In some instances, machine learning simulates the functioning of a biological brain by using data and neural networks.It is also known as predictive analytics when applied to business problems.

C. ROBOTICS

AI is a branch of engineering, combining electrical and mechanical engineering, computer science, and robotics to design, build, and deploy robots. Various aspects of robotics:

- Robots are mechanically constructed, shaped, or fashioned to accomplish a particular task.
- The machines are powered and controlled by electrical components.
- Robots are controlled by computer programs that determine what, when, and how they do things.

D. Natural Processing Language

Natural language processing (NLP) is the study of how computers interact with human language, artificial intelligence and computer science. In particular, it focuses on programming computers to study and analyze large amounts of natural language data. To achieve this goal, the computer must be able to "understand" the content of a document, including the local nuances of the language within it. In addition to correctly locating and extracting information from documents, the technology can also categorize and organize them within themselves. NLP applications that are straightforward include retrieving information, answering questions, and translating text. In symbolic AI, the semantics of sentences is translated into logic by formal syntax. A combination of intractable nature of logic as well as breadth of common sense knowledge prevented this from producing useful applications. Co-occurrence frequencies are now used as statistical techniques. (the frequency with which words appear together), keyword spotting (finding a particular word in text), and deep learning based on transformers (which can find patterns in text).

There are three parts to NLP:

- Speech Recognition: The process of translating spoken language into text.
- Natural Language Understanding (NLU): A computer's ability to comprehend human speech.
- Natural Language Generation (NLG): The process of producing natural language by a computer.

E. Digital assistant

The intelligent virtual assistant (IVA) or intelligent personal assistant (IPA) is a software agent that provides services or performs tasks for an individual based on commands or questions. Virtual assistants generally or specifically accessible through online chat are sometimes referred to as "chatbots.".

Virtual assistants can interpret voice commands and respond by synthesizing a voice. Voice Assistants allow users to access information related to their home automation devices, control media playback, and manage basic tasks such as calendars, to-do lists, and email.

Virtual or Digital Assistant:

- Google assistant
- Cortana
- Siri
- Alexa

ISSN 2394 - 7780

TABLE I. Overview table on different VPA

VPA	Relaesing date	Developer	Website	OS	Platform	Lang. Support
Siri	14 October 2011	Apple	Www.apple. com/io s/siri	Ios 5 onward	Iphone, mac, apple watch, ipad and more apple product	21
Google assistant	18 May 2016	Google	Assistant.goo gle.com	Android	Android device like smartphone. Headphone, speaker and tv	12
Cortana	2 April 2014	Microsoft	Microsoft.co m/enus/wind ows/cortana	Windows	Amazon alexa, windows 8 and onward, xbox ,skype , cyanogen os	9
Alexa	6th Nov 2014	Amazon	https://alexa. amazon.com	Android,ios	Fire OS 5.0 or later, iOS 11.0 or later Android 4.4 or later	8

a. GOOGLE ASSISTANT

Initial date: 18 May 2016 by Google. With artificial intelligence at its core, Google Assistant runs across iOS and Android. Users' data is not stored by Google Assistant without their permission. If this feature is desired, the user can check a box under Voice & Audio Activity (VAA). The Google Assistant receives audio files from the cloud when VAA is enabled, and these audio files are used to enhance its performance. A few of the features are:

- A platform that allows third-party device makers to implement their own "Action on Google" commands
- Interacting via text and learning more languages
- In order to enhance location-specific queries, users can enter specific geographic locations on their devices

b. CORTANA

Microsoft's personal productivity assistant, Cortana, can help you work more efficiently and focus on what matters. Microsoft's Cortana is a virtual assistant that utilizes the Bing search engine to provide users with information and reminders. According to its software platform and region, Cortana is currently available in languages such as English, German, Portuguese, Spanish, French, Chinese, Italian, and Japanese.

c. Apple's Siri

In addition to its role in Apple Inc.'s iOS, macOS, iPadOS, watchOS, and tvOS operating systems, Siri also features a natural-language user interface, Focus-tracking, gesture-based control and other features. By delegating requests to a set of Internet services, the virtual assistant provides answers, makes recommendations, and performs actions. Continually using the software, the software adapts to user preferences, language usage, and searches. This results in personalized results. An earlier version of Siri was developed by the SRI International Artificial Intelligence Center. It relies on Nuance Communications' speech recognition engine, while Siri uses advanced machine learning techniques.

d. Amazon's Alexa

Initial date: 6th November 2014 by Amazon Alexa, also known as Alex, is a virtual assistant technology developed by Amazon that is based largely on a speech synthesiser called Ivona, bought by the company in 2013. This technology initially appeared in the Amazon Echo smart speaker, Echo Studio, Echo Dot, and Amazon Tap speakers by Amazon Lab126. Voice interaction is available, music playback is available, to-do

ISSN 2394 - 7780

lists can be made, alarms can be set, podcasts can be listened to and news, traffic, and sports can be viewed in real time.

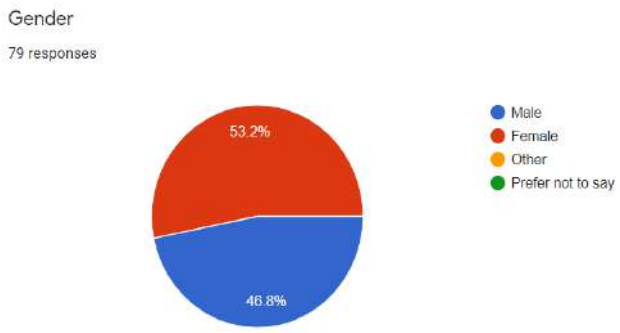
Table II. Feature table for different VPA

Sr.No	Feature	Siri	Google Assistant	Cortana	Alexa
1	Trigger	"Hey Siri"	"OK Google"	"Hey Cortana"	"Alexa"
2	Type Question	×	✓	✓	×
3	Set Alarm	✓	✓	✓	×
4	Access function within app	✓	Limited	Limited	✓
5	Can use on computer	✓	✓	✓	×
6	Makes call	✓	✓	×	×
7	Follow-up question	Limited	Works well	Limited	Limited
8	Geofancing	Limited	✓	✓	×
9	Search Engine	Google	Bing	Google	Bing
10	Third party app support	✓	✓	✓	✓
11	Conversational Question	✓	Limited	✓	✓

III. METHODOLGY

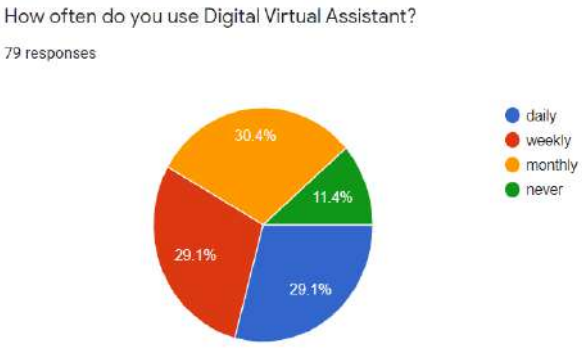
We recruited 79 users (42 Female, 37 Male) among them 59% of population are frequent users.

FIGURE I:

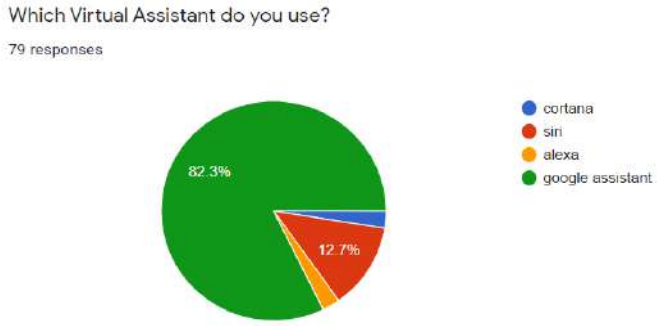


English was their first language for most of them. Virtual assistants were our focus. Some of our users addressed the fact that they weren't active users. They used different devices to create their experiences.

FIGURE II:



Through above given chart we can say that around 58% do use Assistant frequently or whenever they need. The below chart shows that in our sample Google assistant is the most used assistant in all.



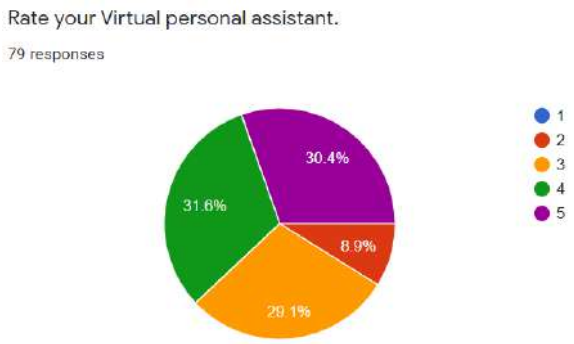
ISSN 2394 - 7780

Result shows 82% for Google assistant, 13% for Siri, 5% for both cortana and Alexa.

Mostly user is satisfied with their devices and assistant. And as per other questions in form for survey we say that all most all questions were correctly answered by assistant to user.

Chart for Final rating is like below:

FIGURE III:



the rating here is in manner of 1-5 i.e. low-high. It shows that around 62-70% is satisfied by digital personal assistant.

Further we tested the fault rate of consecutive VPA by asking following question:

1. **Which team won the soccer world cup of Italy 90?**

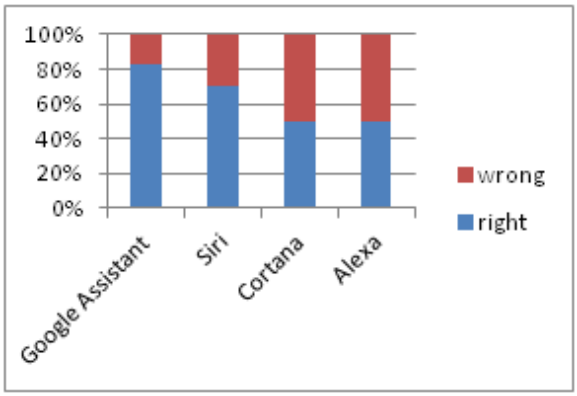
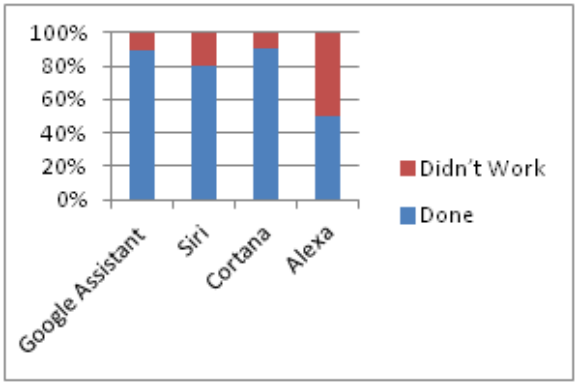


FIGURE IV:

The above chart shows the correctness percentage of each VPA.

2. Asked the user to Set the alarm for 11 o'clock through VPA:

FIGURE V:



Above graph shows the percentage of success in Task for Different VPA.

IV. CONCLUSION

The purpose of this study was to evaluate four VPA in order to determine which was the most effective by looking at the quality of their answers, features and user satisfaction. Among the personal assistants included in the study were Siri, Cortana, Alexa, and Google Assistant. Overall, 79 participants participated.

ISSN 2394 - 7780

The result described that siri and google did the best task. From the test result of sample we say that Google assistant get highest result with average of 82% and following up with 12.7%, 2.5% and 2.5% simultaneously for siri, cortana and alexa . Some conversation and research gives the idea that each age group uses voice assistants differently. There is great insight and some early evidence that voice assistants can be used for purposes other than home entertainment and basic tasks.

REFERENCES

[1] Dewi Agushinta R, Baby Lolita, Dwike Aprilia Setianti, Hardimen Wahyudi, I Putu Partadiyasa(2021) Comparative Study of Intelligent Personal Assistant. International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 6, June 2012

[2] Patil, Dr & Shewale, Atharva & Bhushan, Ekta & Fernandes, Alister & Khartadkar, Rucha. (2021). A Voice Based Assistant Using Google Dialogflow and Machine Learning. International Journal of Scientific Research in Science and Technology. 06-17. 10.32628/IJSRST218311.

[3] Dr. Sachin Upadhye.(2019) Comparative analysis of Virtual Personal Assistant(s).

[4] Tulshan, Amrita & Dhage, Sudhir. (2019). Survey on Virtual Assistant: Google Assistant, Siri, Cortana, Alexa: 4th International Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers. 10.1007/978-981-13-5758-9_17.

[5] Bohouta, Gamal & Këpuska, Veton. (2018). Next-Generation of Virtual Personal Assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home).

[6] Dekate, Abhay & Kulkarni, Chaitanya & Killedar, Rohan. (2016). Study of Voice Controlled Personal Assistant Device. International Journal of Computer Trends and Technology. 42. 42-46. 10.14445/22312803/IJCTT-V42P107.

[7] Imrie, Peter & Bednar, Peter. (2013). Virtual Personal Assistant. International Journal of Engineering and Innovative Technology (IJEIT).

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

NEED FOR ANALYTICS TO BRIDGE THE GAP BETWEEN BUSINESS AND INFORMATION TECHNOLOGY FOR EFFECTIVENESS

[1]Swapnali Lotlikar and [2]Prachi Shah
[1]Assistant Professor and [2]M.Sc. (I.T.) Student, Department of Information Technology, Usha Pravin Gandhi College of Arts, Science and Commerce

ABSTRACT
Over the decades, the misperception between Business and Information technology has been peculiar. This chasm between the two creates problems that can be inefficacious. The experts in the analytics domain advised, bridging this gap should be a high priority, as yet this problem is still existent today [8]. A key component for bridging this Business/IT chasm is an analytics-driven outlook for decision-making effectiveness. This research focuses on how analytics can help for bridging the gap between Business and IT chasm in an organization with the help of some renowned companies' reason for using analytics, also some aspects of business intelligence and big data with its importance on organizations business has been highlighted.

Keywords: Data Analytics, Business Analytics, Business Intelligence, Big Data

I. INTRODUCTION
It is foreseeable that there are Business and IT cultures in an organization. There's this thought among business people, that IT does not understand the business, they are unable to solve an issue, and they only care about technology [1]. In contrast, IT people might think, Business people don't have any technical knowledge, they are unable to comprehend the requirement of building and maintaining the system [8]. This is where analytics comes as assistance for effectiveness between the two. Jimmy Augustine, working in HP under the Application Performance Management department, said: "As an IT industry, we should force bridging the IT and Business gap. Why? Both should be like one, like peanut butter and jam. One hand should know what the other hand is thinking and vice versa." [8]

Today, for many organizations, Technology *is* Business, and that is why they have adopted analytics as to the crux. Analytics has emerged as an important aspect for a better decision-making process. Companies in the industry use analytics for providing value to users. Many organizations think analytics is a key realm in IT, giving the apposite support to help improve their business. An analyst, acting as a new communication intermediary, is expected to significantly improve communication coherence between business departments and IT, which was once highly cumbersome and ad-hoc. It allows business managers to understand the dynamics of their business, foresee market shifts, and also manage risks.

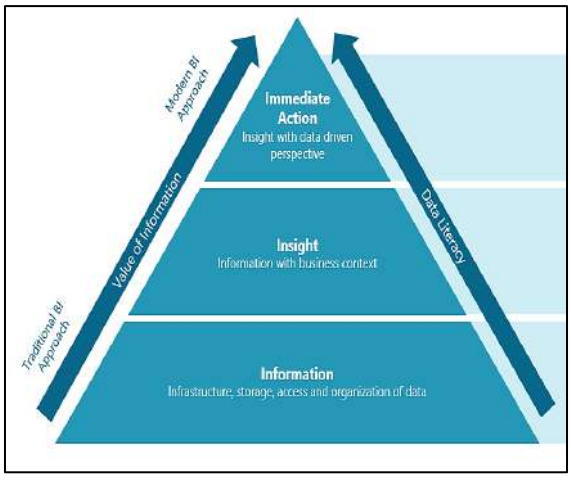


Fig. 1. Overview of Analytics

Most IT organizations have already started to make a move towards advanced analytics. Advanced analytics in general, not only augments value to business performance and propels profitability, but also offers affluence of possibilities when utilized effectively. Although, many organizations still struggle in maximizing these opportunities. To become analytically data-driven and generate data more accessible, organizations must work on alleviating these challenges to obtain the full power of analytics. This is where Business Analyst and Data Analyst excels in such tasks to alleviate such struggles. When proper data is provided, a business can augment its chances with accuracy. Data plays a vital role in successful decision-making effectiveness. Need for clean

ISSN 2394 - 7780

and modified datasets to uncover trends and insights that are subsequently used for profitable business decisions. Here, business analytics provides the leaders with the tools to transform their important operational, product, and consumer data into valuable insights that lead to agile decision-making effectiveness and business success. From enabling businesses to make consumer-oriented decisions in helping them address operational inefficiencies, analytics is radically changing its perception of the importance of data. The paper is divided into four different sections. The first section consists of analytics and their types, the second section mainly focuses on the thoughts and reviews of other researchers and writers, the third section provides an overview of advanced analytics and concludes with the approaches used in the paper and appropriate measures to take into consideration while mitigating the gap.

II. LITERATURE REVIEW

Hugh J. Watson, Professor of MIS, Terry College of Business at the University of Georgia, had mentioned in his paper in March 2018, to bridge this culture chasm, both the IT, as well as Business departments, should have an understanding between the two by establishing Business Intelligence centre in an organization. These BI centres consist of a group of specialized people from IT backgrounds. In 2004, he was interviewing for a Continental Airlines case study [1], he was unable to distinguish if the analysts were from the data warehousing unit or the business unit. That is why he emphasizes the importance of Business Analysts in an organization as they are prominent in communicating between the Business and Technology unit, to abstain from this cultural chasm.

Guangming Cao, Yanqing Duan, and Gendao Li stated in their paper that Business Analytics should be linked to decision-making effectiveness. The authors implied that organizations require Business Analytics for making faster and better decisions. The convergence of big data analytics, evolvement in IT, and Business Analytics have introduced decision-making at a thoroughly new level that is data-driven, allowing Business and IT managers to see the invisible past [2].

Sonia Johnson, Head of Marketing, Inside Info, mentioned in an article, as a business is growing more into the data-driven culture, IT organizations should start investing in new and advanced data software and tools which has strong ingesting and integration capabilities. By taking advantage of these applications and aligning the IT team, Business, and Data leaders together it has been inferred that Data analytics should become a part of a company's culture [12].

Now, while we look at big data analytics, business intelligence, and cloud, a lot of companies have started using analytics as their core performer to save company money from redundant use, maintain customer relationships, gain better insights and decisions, and reach the top of their industry business. In this decade, there is a canard that Big Data is Big Business, which has turned out to be true.

In his article, Eleanor O'Neill, has bogged down into a few of the top companies using data and analytics to procure a cut-throat edge [13]. which is encapsulated in the following table:

Firm	Type of Industry	Usage
Amazon	E-commerce	-For recommendations (collaborative filtering engine), this has helped the company to earn 35% of its annual sales. -Uses predictive analytics, to bolster customer satisfaction and build a loyal relationship. -Uses dynamic pricing, for optimization of product prices.
Netflix	Entertain-ment	-For recommending TV shows and movies to the subscribers according to their preferences. -For optimization of quality and firmness of the video -To create a personalization account for each customer
Starbucks	Retail Coffee	-Collects data to recommend products to loyal customers. -To create Marketing Campaigns and menus for attracting customers. -Uses data like incomes, traffic patterns, and population to analyse and target locations for new stores.
American Express	Banking, Financial services	-Identifying extortion, and bringing traders and clients closer together. -To distinguish more false transactions and save millions. -To customize offers to hold clients and utilize this information to keep up with associations with clients

ISSN 2394 - 7780

T-Mobile	TeleComm-unication	- It combines transactions of customers and data interaction to predict fluctuations from the customer's side -For advertising Campaigns

Fig. 2. Use of Analytics in MNCs

In his paper in [7], the author had worked with a variety of companies in six data-rich industries, they found that fully utilizing the data and analytics requires three mutually encouraging capabilities. Firstly, organizations must be prepared to identify, combine, and manage different sources of insights and knowledge. Secondly, they need the capability to develop advanced analytics models for the prediction and optimization of the outcomes. Thirdly, and most important, management must gather the muscle to rework the enterprise in order that the information and models actually capitulate better decisions. Two crucial features bolster those activities: a pellucid strategy for a way to utilize data and analytics to participate, and deployment of the pertinent technology architecture and capabilities.
In an article of PwC, the authors mentioned The New IT platform, to bridge the gap between Business and IT, as the gap between the two is growing wider. A survey was conducted of nearly 1500 technology and business executives and only 54% of the respondents agreed that business and IT share a mutual understanding of a corporate strategy. They have introduced the need for a New IT platform that will bring business and IT together. IT needs to adopt more agile methods to meet the voracious demands of businesses. The authors inferred that Business and IT must work incessantly, synergically, and progressively [9].

III. METHODOLGY

As from articles and ongoing requirements in an organization it has been observed that analytics play a vital role in entrenching a successful Business and IT decision-making effectiveness. It has been made obligatory by many organizations to adopt this analytics-driven program to magnate their industry. This research methodology evinces the demand and importance of analytics-driven decision-making effectiveness to bridge the gap between Business and IT.

It is observed that few organizations find it difficult to implement the strategy for better decision-making effectiveness. According to the analysis made from articles and observations of the authors from their research, it has been implied that the best strategy to overcome this difficulty and Business/IT chasm is to get used to analytics-driven decision-making. Firstly, Business leaders and Tech leaders should have a basic knowledge about each other. For this Business Analyst (to bridge the communication gap between the business and technology), Data Analyst (to work on data provided by the businesses for better insights, which helps to understand the data and business conveniently) and also as said earlier, Business Intelligence Centre (to work on a defined common set of objectives with right management system in place) should be entrenched.

Analysts in the company should have a thorough knowledge of different software that is used for analysis, as it helps IT to have a broad and bigger idea of what is going on in the business with the insights and patterns that are generated. According to the research, it was observed that analysts in the technology had once worked on the business section and vice versa [1]. And it was also inferred that the analyst in the domain that works on both the side, should have a great knowledge of databases. Secondly, the most important aspect is to eliminate the distrust between the two. Lastly, as discussed above, applying appropriate analytics considering the business requirements, results in a successful outcome.

C. ANALYTICS

As technology is evolving day by day, Analytics and data mining have become more important than before. Analytics is a scientific process used to examine the raw data, to draw a meaningful and logical sense from it. It is also a combination of technologies, skills, and practices used to analyse an organization's data as a way to have a proper insight and make decisions based on the data given, in the future using statistical analysis [12].

Fig. 3. Complexity vs Value

ISSN 2394 - 7780

There are four types of analytics used in an IT organization for providing better business decisions by cleaning, dissecting, and absorbing in a way that helps in creating better solutions for businesses. They are as follows:

1. **Descriptive analytics**: This is a type of analytics that deals with the question "What is going on? What has happened?" Descriptive analytics breaks the huge volume of data into smaller pieces to give valuable insights. It aims at crunching relevant information. It uses arithmetic operations.
2. **Diagnostic analytics**: This type of analytics states "Why did it happen?" It dives deeper to comprehend the basic explanation and cause of any events. It is used to identify the source of the problem. It is based on probabilities and distributions of outcomes by using principal component analysis, sensitivity analysis, and regression analysis.
3. **Predictive analytics**: This is a type of analytics that depicts "What could have happened?" It derives the prospect of the outcomes, looks to the longer term, and is predicted based on probability. It converts the raw data into functioning information which is then used to uplift the firm ahead.
4. **Prescriptive analytics**: This type of analytics talks about "What should happen?" It uses both Descriptive analytics and Predictive analytics. It updates the relationship between action and outcome using statistical algorithms.

As it happens, the more complex the analytics is, the more value it brings.

D. Advanced Analytics

Data-driven Advanced Analytics allows organizations to have an absolute or "360-degree" view of the operations and consumers. The insights that they procure from such analyses are used for optimization, to direct and automate the decision-making process successful to achieve the organizational goals [8]. In an article published by the leading analytic firm in the world, Gartner, Advanced Analytics is the semi-autonomous or autonomous inspection of data or content using pertinent techniques, software, and tools, typically afar those of conventional business intelligence, to discover deep insights, generate predictions, or make recommendations [14].

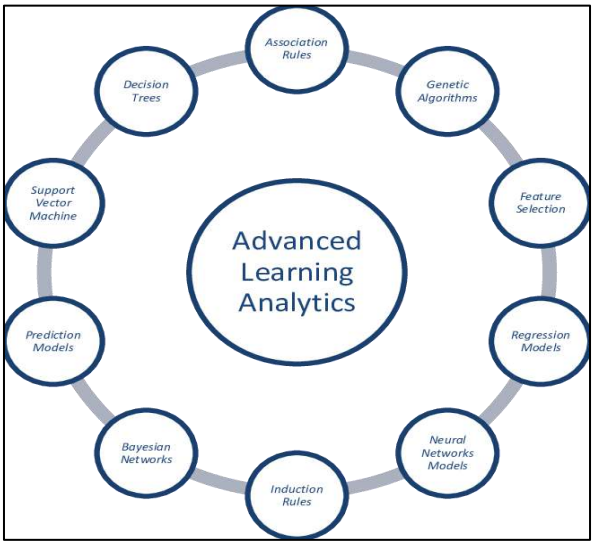


Fig. 4. Advance Learning Analytics

Advanced Analytics elucidates data analysis that goes afar basic mathematical calculations as sums and averages of data, filtering and sorting of data. It uses formulas with statistics and algorithms to produce new information, to perceive patterns, and also to speculate outcomes and their probabilities [15r]. Taking advantage of such techniques helps in building such a strong foundation for advanced analytics to be thorough and mature [16]. Some of the methods include:

1. **Data Mining**: It is a process of analysing a large amount of data to generate patterns and establish relationships with the data for the businesses to help solve the problems, mitigate risks with the help of business intelligence.
2. **Machine Learning**: It is a technique that uses computational methods to find patterns in data, and automates analytical model building without human intervention.

ISSN 2394 - 7780

3. **Cohort Analytics:** It is a subset of Behavioural Analytics that generates broadly applicable insights by examining the nature and behaviour of the people in a group.
4. **Cluster Analytics:** It mainly focuses on differences and similarities in a dataset in such a way that objects in one group are similar to each other than the ones in other groups.
5. **Retention Analytics:** It is used to comprehend cohorts of users or customers. It is a way to gain an insight on maintaining a beneficent customer base by improvising retention and acquisition rate.
6. **Complex Data Analysis:** It is a process that aggregates and analyses the data that are coming from different sources when any event is fired.

IV. CONCLUSION

To prosper with effectiveness, organizations should bridge the gap between Business and IT by creating a centre for Analysts. In this data-driven world, analysts need to look through two lenses at the same time. One is through understanding business requirements and the other by communicating them through analytics-driven decision-making effectiveness. For this decision-making, one must adopt germane analytics by the contemplation of data. The analyst must understand how to use analytical tools and software to transform the raw data into a dashboard of meaningful metrics which can be understood by both businesses and IT. It can be inferred that both parties must be able to communicate in the language of revenue and cost and also through technological strategies.

REFERENCES

[1] Watson, H. J. (2009). Bridging the IT/Business Culture Chasm. Business Intelligence Journal, 14(1), 4-7.

[2] Cao, G., Duan, Y., & Li, G. (2015). Linking business analytics to decision-making effectiveness: A path model analysis. IEEE Transactions on Engineering Management, 62(3), 384- 395.

[3] Parks, R., & Thambusamy, R. (2017). Understanding business analytics success and impact: A qualitative study. Information Systems Education Journal, 15(6), 43.

[4] do Amaral, F. L. (2017). Internet Software and Services Fernanda Luvizotto do Amaral University of West Florida Information Resources and Industry Analysis-GEB5930 December 10, 2017.

[5] Bose, R. (2009). Advanced analytics: opportunities and challenges. Industrial Management & Data Systems.

[6] Martin, V. A., Lycett, M., & Macredie, R. (2003). Exploring the gap between business and IT: an information culture approach. Proceedings of ALOIS, 265-80.

[7] Barton, Dominic, and David Court. "Making advanced analytics work for you." Harvard business review 90, no. 10 (2012): 78-83.

[8] apmdigest.com, “apm bridge the gap between it and business”.

[9] pwc.com, “digital iq new it platform final.pdf”.

[10] towardsdatascience, “using analytics for better decision making”.

[11] prodevbase.com, “types of data analytics to improve decision making”.

[12]https://www.entrepreneur.com/article/34178

[13]https://www.icas.com/thought-leadership/technology/10-companies-using-big-data

[14]https://www.gartner.com/en/information-technology/glossary/advanced-analytics

[15]https://bi-survey.com/predictive-analytics#def

[16]https://looker.com/definitions/advanced-analytics

[17] Fig.1 from google.com image search results

[18] Fig.3 from google.com image search results

[19] Fig.4 from google.com image search results

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

SENTIMENT ANALYSIS

Rikin Shah
Master in Science in InformationTechnology, Usha Pravin Gandhi College of Arts, Science, and Commerce
Vile Parle west, Mumbai, India

ABSTRACT
Sentiment analysis is the substantial task of NLP (Natural Language Processing). In this paper, the basic objective is to develop a web-based system with the capability of providing user insights, if a product is worth purchasing or not, based on the user review on one click. A general process includes the use of text-blob and performs the sentiment analysis to demonstrate its usefulness. In this work, A detailed description and literature survey on sentiment analysis, different models, their accuracy, and drawbacks are discussed. In this study, the sample datasets from online product reviews were collected from Amazon.com, Flipkart.com, and data repositories such as https://www.kaggle.com is used. At last, The future pointer for further research is also presented.

I. INTRODUCTION
The sentiment is an emotion or attitude prompted by the feelings of the customer. It is also called opinion mining [1] which studies people's opinions towards the product. The dataset is collected from the website. It examines the problem of studying texts, like posts and reviews, uploaded by users on microblogging platforms, forums, and electronic businesses, regarding the opinions they have about a product, service, event, person, or idea. The social web has made enormous amounts of information available to users globally at just the click of a button. Consumers often tend to rely on such text, especially those in the form of opinions or experiences regarding a particular product which makes it essential that this information should be available in a systematic manner. The most common use of Sentiment Analysis is this of classifying a text to a class. Depending on the dataset and the reason, Sentiment Classification can be binary (positive or negative or neutral) or multi-class (3 or more classes) problem. (See Fig 1).

Fig 1 - Sentiment Analysis Classes

II. BRIEF DESCRIPTION OF THE PROBLEM
The objective, purpose, and the proposed system, Data Set acquisition are described.

A. OBJECTIVE
The basic objective is to develop a web-based system with the capability of providing user insights, if a product is worth purchasing or not, based on the user review on one click. The availability of this huge volume of data is its diversity and its structural non-uniformity which makes it more difficult for a human to read. This project helps to put all data into the structural format. This makes the process quicker and easier than ever before, thanks to real-time monitoring capabilities.

B. PURPOSE
Sentiment analysis is extremely useful in online monitoring as it allows us to gain an overview of the wider public opinion behind a particular product. The ability to extract insights from online data is a practice that is being widely adopted by organizations across the world. Therefore, it makes it easy for the customer or user and helps to select the correct online E-product without any difficulties.

C. PROPOSED SYSTEM
Sentiment Analysis is a web-based system, on the Reviews of E–Products taken from a different website and after that, it is used to generate the output in the form of emoji or sentiments. This allows users to decide

ISSN 2394 - 7780

whether a product is good or bad based on the analysis. The applications of sentiment analysis are broad and powerful. This will be done using the Python Text-blob library [7], which checks the subjectivity and polarity. It is elaborated ina further section.

D. Data Set Acquisition

The dataset was obtained from Kaggle and amzon.com. The proposed methodology uses a diverse dataset. The dataset contains reviews of mobile phones and other details such as names, prices, and other details.

III. LITERATURE REVIEW

In this section, several similar reviews of the other research paper to understand the concept in a more precise way is being discussed.

A. Research Paper on "Sentiment analysis using productreview data" By journal of big data. *[2]*

Sentiment analysis or opinion mining is a field of studythat analyses people's sentiments, attitudes, or emotions towards certain entities. This paper tackles the problem of sentiment analysis, sentiment polarity categorization. Based on the Online product reviews from Amazon.com are selected as data used for this study. A sentiment polarity categorization process (See Fig 2) has been proposed along with detailed descriptions of each step. Experiments for bothsentence-level categorization and review-level categorization have been performed. In this paper [2], they have taken three categorized methods such Naïve Bayesian classifier, Random Forest, Support vector machine to perform the analysis. The outcome as per the paper was SVM Model and Naïve Bayes's performance was identical. The analysis was done was both in terms of sentence-level and review level categorization. There were certain limitations such as Review level categorization becoming quite difficult when they are classifying reviews as per the star-scaled rating. This study relies on sentiments token, andit may not work well for implicit statements.

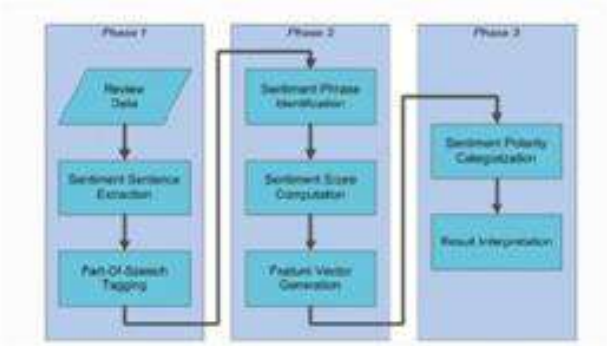


Fig 2 - Sentiment Analysis Classes

B. Opinion Mining and Sentiment Analysis – By Pang B,Lee L. [3]

They have examined the relation between subjectivity detection and polarity classification, showing that subjectivity detection can compress reviews into much shorter extracts that still retain polarity information at a levelcomparable to that of the full review. For example, in the data we used, boundaries may have been missed due tomalformed HTML. fact, for the Naive Bayes polarity classifier, the subjectivity extracts are shown to be more effective input than the originating document, which suggests that they are not only shorter but also "cleaner" representations of the intended polarity. We have also shown that employing the minimum-cut framework results in the development of efficient algorithms for sentiment analysis. Utilizing contextual information via this framework can lead to statistically significant improvement in polarity- classification accuracy. Directions for future research include developing parameter selection techniques, incorporating other sources of contextual cues besides sentence proximity, and investigating other means for modeling such information.

C. entiment Analysis of Review Datasets using Naïve Bayes' and K-NN Classifier" by Lopamudra Dey, Sanjay Chakraborty [5]

The study aims to evaluate the performance of sentiment classification in terms of accuracy, precision, and recall. In this paper, they have compared two supervised machine learning algorithms of Naïve Bayes' and KNN for sentiment classification of the movie reviews and hotel reviews. The experimental results show that the classifiers yielded better results for the movie reviews with the Naïve Bayes' approach giving above 80% accuracies and outperforming the k-NN approach. However, for the hotel reviews, the accuracies are much lower, and both the classifiers yielded similar results. Thus, this paper says Naïve Bayes' classifier can be used successfully to analyze movie reviews. (See Fig 3) Accuracies of the classifiers with the 2 datasets, for further work they would like to compare try and come up with an efficient sentiment analyzer like the

ISSN 2394 - 7780

random forest, Support vector machine, etc. And try to implement a new algorithm utilizing the benefits of both algorithms so that it can be used effectively in data forecasting.

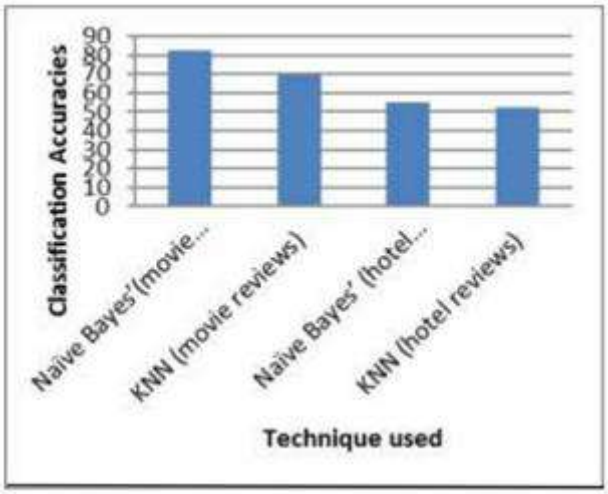


Fig 3: Analysis of movie reviews

D. Mining and Summarizing Customer by Hu, M., and Liu,[6].

This paper has the textual and linguistic features which are been extracted and classified to develop a categorize blog posts with respect to sentiment analysis. We ask whether a given blog post expresses subjectivity (vs. objectivity), and whether the sentiment a post expresses represents positive ("good") or negative ("bad") polarity. They have simplified the expression of subjectivity vs. objectivity, as well as positive and negative polarity, intobinary classification tasks. They make use of verb-class information in the sentiment classification task, by exploiting lexical information contained in verbs. To obtain this information they use Semantics an automatic text analyzer that groups verbs according to classes that often correspond to their polarity classification. Additionally, they have utilized Wikipedia's online dictionary, the Wiktionary 1, to determine the polarity ofadjectives seen throughout the posts. They propagate thislexical information up to the post level by a machine learning classification algorithm in which a binary classification (e.g., objective/subjective) is made for eachblog post.

E. Sentiment Analysis of Mobile Reviews using Sentiwordnet" by Gaurav Dudeja and Kapil Sharma [4]

This document-level plans executed by Gaurav Dudeja and Kapil Sharma include the utilization of 'Adverb+Adjective' consolidated only and the utilization of 'Adverb+Verb' combined with the 'Adverb+Adjective' combination. This is done to explorethe opinionated value of distinctive linguistic features of a review and discover a way As a result, many of the sentiment calculations were highly influenced by the tacit assumption is that a review describes only best aggregate all the opinionated information in a review together to produce the document level sentiment summary. Theresults demonstrate that joining the sentiment score of 'Adverb+Verb' joins to the commonly used 'Adverb+Adjective' joined further improves the accuracyof the sentiment analysis result. The best weightage factor for verb scores got through multiple experimental runs is 30%. The aspect-level sentiment analysis algorithmic formulation designed by us is a novel and unique way of obtaining a complete sentiment profile of a phone from multiple reviews on different aspects of evaluation. The resultant sentiment profile is informative,easy to understand, and extremely useful for users.

IV. PROPOSED METHODOLOGY

A. TEXTBLOB [8]

In this first phase of development, we are working on the implementation of the project using the Text-Blob Library and in detail concept.

1. INTRODUCTION

Text-Blob is a Python library used for processing textualdata. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of- speech tagging, noun phrase extraction, sentiment analysis, classification, translation, and more. Text-Blob stands on thegiant shoulders of NLTK and pattern and plays nicely with both.

Text-Blob aims to provide access to common text- processing operations through a familiar interface. You can treat Text-Blob objects as if they were Python strings that learned how to do Natural Language Processing.

ISSN 2394 - 7780

Sentiment analysis is the process of determining the attitude or the emotion of the writer, i.e., whether it is positive or negative or neutral, and which can be done using Text-Blob.

2. Working of Text-Blob

The sentiment function of text-Blob returns twoproperties polarity and subjectivity. Let us See what Polarityand Subjectivity are.

The polarity score is afloat within the range [-1.0, 1.0]. (SeeFig 4)

The subjectivity is afloat within the range [0.0, 1.0] where

0.0 is very objective and 1.0 is very subjective. ForReference Fig: below is the Triangle representation.

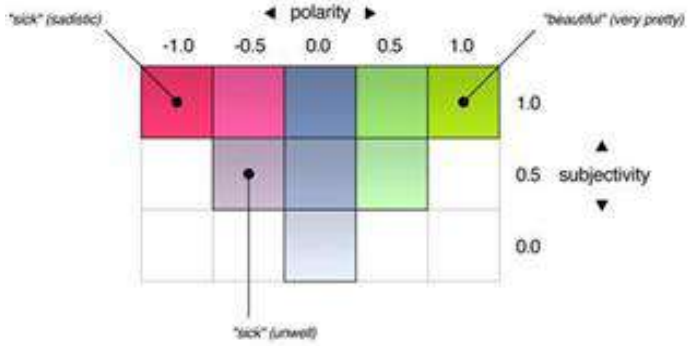


Fig: 4 Working of Text Blob

B. Naïve Base Classifier [9]

Once, the successful implementation of the project We can enhance the project accuracy level using the Naïve BasedModel. In the above section [III], Literature review about theaccuracy of Naïve based classifiers. Here is the concept of a Naïve Based Classifier.

1. INTRODUCTION

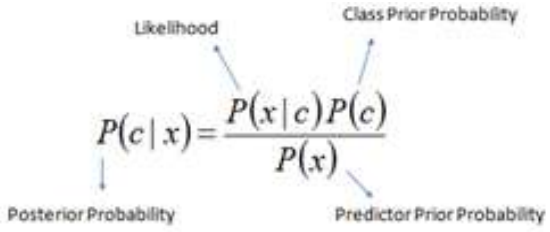
It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruitis an apple and that is why it is known as 'Naive'.

2. Formula: [7]

The Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below Fig [5]: Naïve bayes formula



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Fig 5: Naïve bayes formula

3. Working of Naïve based classifier

Let us understand it using an example.

Below I have a training data set of weather and corresponding target variable 'Play' (suggesting possibilities of playing). Now, we need to classify whether players will play or not based on weather conditions. Let us follow the below steps to perform it.

ISSN 2394 - 7780

Step 1: Convert the data set into a frequency table.

Step 2: Create a Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.

Step 3: Now, use the Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of the prediction.

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

Fig 6: Naïve Bayes example

V. RESULT AND DISCUSSION

Below are the results on the above library Text Blob project implementation, I have distinguished two graphs, one is the Gauge graph, and another is the Bar graph. In each of the graphs, the data has been visualized in a specific manner.

In gauge graph: The output is on a whole dataset as an average by using text blob based on the 3 classes using Polarity and Subjectivity. Further dividing into other 5 different parts. [See Fig 7]. The different parts will indicate the product positivity and the category of a product where itstands in terms of reviews.

Fig 7: Gauge Graph – Outcome

In Bar Graph (See Fig 8) the output is in such a manner where it takes how much the number of are Positive, Negative, and Neutral sentences. Further, it's has divided itinto the different parts such as Excellent, Very good, Satisfy, Average, and Poor.

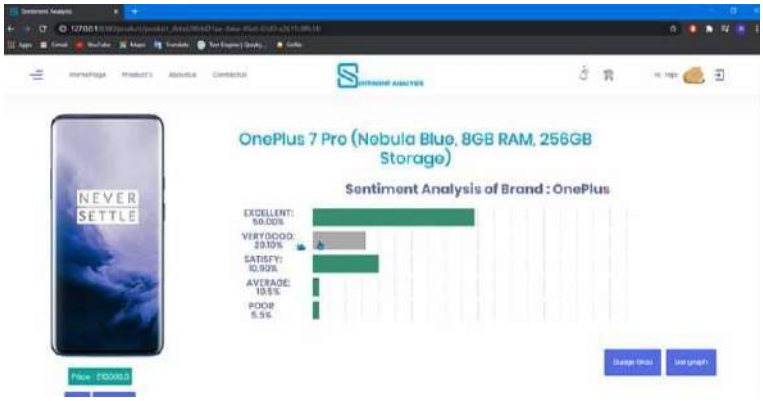


Fig 8: Bar Graph – Outcome

ISSN 2394 - 7780

VI. CONCLUSION AND FUTURE ENHANCEMENT

This website will use to choose your sentiment of theproduct based on the reviews to feed to the machine.

Text Blob is used Framework as Django, Programming Language as python, and Database as SQL is used in this project. the product, and it is based on NLTK. This project could be more analyzed in-depth, and the accuracy of the project can be improved in future scope.

This Project has several limitations as it is developed ona very basic level, but it could be developed on a more rigid level and with more accuracy in the future scope. Due to Unauthenticated or Spam user reviews, this may give a false output. So, therefore, the accuracy could be improved of this project and hence it could be improved using the trained models, authentication of user reviews,and other ways, but currently, it solves the objective of the product Hence this project solves the problem for the customer to choose a vice product from an online website, In just one click.

REFERENCES

[1] F. Bodendorf and C. Kaiser, "Mining Customer Opinions on the Internet - A Case Study in the Automotive Industry," 2010 Third International Conference on Knowledge Discovery and Data Mining, 2010, pp. 24-27, doi: 10.1109/WKDD.2010.129.

[2] Fang, X., Zhan, J. Sentiment analysis using product review data. Journal of Big Data 2, 2015. https://doi.org/10.1186/s40537-015-0015-2

[3] BoPang and LillianLee,“A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts”. Department of Computer ScienceCornell University Ithaca, 2004.

[4] Gaurav Dudeja and Dr. Kapil Sharma, “Sentiment Analysis of Mobile Reviews using Sentiwordnet”. Technological University Delhi, 2015

[5] Lopamudra Dey, Sanjay Chakraborty, “ Sentiment Analysis of Review Datasets using Naïve Bayes’ and K-NNClassifier”. Department of Computer Science & EngineeringHeritage Institute of Technology Institute of Engineering & Management Kolkata, 2016

[6] Minqing Hu and Bing Liu ,“Mining and SummarizingCustomer Review”. Department of Computer Science University of Illinois : Chicago, 2004

[7] www.machinelearningplus.com/predictive- modeling/how-naive-bayes-algorithmworks-with-example- and-full-code/

[8] https://textblob.readthedocs.io/en/dev

[9] http://en.wikipedia.org/w/index.php?oldid=422757005

.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

RAINFALL PREDICTION OF INDIA USING MACHINE LEARNING

Manisha Divate and Dipali Shah
Usha Pravin Gandhi College of Arts, Commerce and Science, Vile parle, Mumbai, Maharashtra, India

ABSTRACT
Rainfall is the primary source for agriculture in our country. As global warming is increasing, it is getting very difficult to predict the rainfall which will create a major problem in India. Due to lack of proper technology several districts are facing water crisis which affect the crops. Rainfall prediction can help them in several purposes such as farming, agriculture, drinking, health and many other. Many techniques came into existence to predict rainfall. Rainfall can be predicted using Machine Learning Techniques. These techniques can be useful to take precautions in order to protect crops. For rainfall prediction, we can use few machine learning algorithms. Some of the major machine learning algorithms are Logistic Regression, K-means, Linear regression, Support Vector Machine. These algorithms will help us to predict the rainfall for coming years. This will represent a review of different rainfall prediction techniques for the early prediction of rainfall.

INTRODUCTION
Global warming has been hyperbolic so has earth's temperature then the daily climate change. In India, there's no stable season now. we tend to receive rain at any point of the year. Our primary supply of water is rain, however excess of rain will cause harm to the farming business because the crops might get damaged. Agriculture is only obsessed on precipitation so it's important to understand the quantity of rainfall received by the districts and states. The dynamic atmospheric condition and therefore the increasing greenhouse emissions have created it troublesome for the people in general and the planet earth to expertise the required amount of rainfall that is needed to satisfy the human desires Associate in Nursing its uninterrupted use in everyday life.

The rainfall prediction would also assist in coming up with the policies and techniques to take care of the increasing global issue of gas depletion. The changing patterns of rainfall are associated abundant with the world warming; that's increasing of the earth's temperature because of hyperbolic Chlorofluorocarbons emitting from the refrigerators, air conditioners, deodorants and printers etc. that are the many a part of everyday life. This analysis proposes a study and analysis of precipitation victimization machine learning algorithms and compares the performance. However, rainfall prediction with cc using Random Forest regression, Linear regression, provision Regression, k-means & call tree is meant to supply precise and a lot of correct forecasts. The predictions might be utilized for a most vary of functions and therefore will play an important role in minimizing issues} related to water reserves, agricultural problems with dynamic atmospheric condition and flood management.

MACHINE LEARNING
Machine learning is kind of AI that permits package applications to grant correct predicting outcomes. Machine Learning may be a major factor within the field of information science. Machine learning conjointly has intimate ties to optimization: several learning issues are developed as diminution of some loss perform on a coaching set of examples. Loss functions categorical the discrepancy between the predictions of the model being trained and also the actual downside instances.

It turns the dataset into a model that helps in prediction. looking on the problem, the character of information and resources available, it tries to grant the most effective algorithmic program which can help to urge correct result. There are three sorts of Machine learning algorithmic programs.

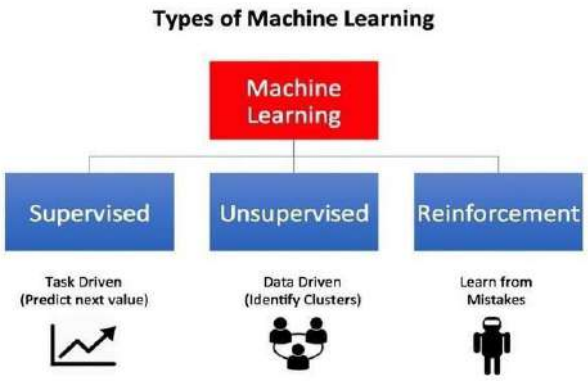


Figure 1: Types of Machine Learning 1

ISSN 2394 - 7780

Supervised Learning
Supervised learning algorithms take direct feedback for the prediction. supervised learning will be categorized in classification and regression methods. KNN, DT, SVM, LR, Artificial Neural Network (ANN), Naive Bayes (NB) etc., are some in style algorithm of supervised learning.

Unsupervised Learning
Unsupervised learning algorithms don't take any feedback for the prediction. This learning finds the hidden patterns in data. 2 easy ideas i.e. principal element Associate in (PCA) and cluster analysis are utilized in unsupervised ML. PCAs eliminate very associated options by exploitation variance matrix, eigenvalues and eigenvectors.

REINFORCEMENT
Reinforcement learning is that the coaching of machine learning models to create a sequence of decisions. The agent learns to attain a goal in an uncertain, probably advanced environment. In reinforcement learning, a synthetic intelligence faces a game-like situation. The computer employs trial and error to return up with an answer to the problem.

Problem Statement
The correct and particular rainfall prediction remains missing which can help in numerous fields like agriculture, water reservation and flood prediction. The difficulty is to formulate the calculations for the rainfall prediction that could be primarily based totally at the preceding findings and similarities and could provide the output predictions which might be dependable and appropriate. The vague and erroneous predictions aren't handiest the waste of time however additionally the lack of sources and cause inefficient control of disaster like terrible agriculture, terrible water reserves and terrible control of floods.

LITERATURE REVIEW
There are many researches and research withinside the literature for predicting the rainfall. In this section, we can talk a number of the paintings associated with our proposed technique.

IN [1], this paper deep neural community algorithms are used to are expecting rainfall that offers higher overall performance. The use of standardized information with already skilled information also can offer interesting outcomes.

IN [2], this paper used the logistic regression set of rules to are expecting the rain for subsequent day. The information gathered for this studies changed into primarily based totally on transactional real-lifestyles information of a main Global firm, and private information changed into saved confidential.

IN [3] paper, K-manner clustering set of rules is used in conjunction with IDW clustering set of rules to calculate the are expecting values. This paper accommodates an exam of numerous strategies and an assessment of every method on the premise of sure overall performance standards.

IN [4], linear regression set of rules is used, the primary goal of this paper is to assist farmers for crop protection. Here, they are expecting the decrease correlation among numerous vegetation season and information outcomes.

Dataset Features
This examine is specializing in common temperature, air stress, wind pace, wind route and humidity. The parameters are described as:

1. TEMPERATURE
It is the diploma to which some thing is warm or cold. It is suffering from numerous elements like latitude, altitude, and distance from the sea, wind with ocean current. It impacts the rainfall in a manner that the quantity of water vapor withinside the ecosystem determines the quantity of rainfall as mild or with a heavy shower.

2. WIND DIRECTION
It is the beginning of the wind from wherein it generates. The route of the wind determines the precipitation. Therefore, it's also essential to research the route of the wind as a parameter for the rainfall prediction.

3. WIND SPEED:
It is the rate of the air transferring from excessive stress to low stress. The wind pace is in near correlation to the rainfall and it's miles important to recollect it as an enter for prediction. The growth withinside the wind pace decreases the depth of the rainfall.

ISSN 2394 - 7780

4**. HUMIDITY**

It is the quantity of water vapor gift withinside the air for a given temperature this is giant for saturation of vapours to shape clouds. The accelerated moisture withinside the air will accommodate excessive saturation. The humid weathers revel in extra precipitation than the dry weather.

5. AIR PRESSURE

The air stress is the stress withinside the ecosystem of the earth. The air stress decreases with the growth withinside the elevation from the earth surface. For example: the air stress on mountains is much less compared to plains.

METHODS USED

1. LINEAR REGRESSION

Linear regression is one of the famous and understood set of rules in device gaining knowledge of. It is primarily based totally on supervised gaining knowledge of. This regression allows to discover the linear dating among structured variable and unbiased variable. It is used to are expecting a structured variable fee primarily based totally on a given unbiased variable.

$y = mx + c$

wherein,

y = structured variable

m = slope

x = Independent variable

c = Intercept.

Univariate linear regression: Univariate linear regression focuses on determining relationship between one independent (explanatory variable) variable and one dependent variable. Regression comes handy mainly in situation where the relationship between two features is not obvious to the naked eye.

Multivariate Linear Regression : This is quite similar to the simple linear regression model, but with multiple independent variables contributing to the dependent variable and hence multiple coefficients to determine and complex computation due to the added variables.

2. **Logistic Regression**

Logistic regression is used for specific variables. This set of rules is primarily based totally on supervised gaining knowledge of. Linear regression is sort of a straight-line regression, however logistic regression generates a sigmoid curve. Based on a predictor or an final results variable, the logistic regression produces sigmoid curves which constitute 0 to at least one fee primarily based totally on logarithmic functions.

Univariate logistic analysis: When there is one dependent variable, and one independent variable; both are categorical; generally produce Unadjusted model (crude odds ratio) by taking just one independent variable at a time..

Multivariate logistic: When there is one dependent variable, and more than one independent variables; All are categorical; produce Adjusted model (adjusted odds ratio) by taking all the independent variables at a time. [i.e. univariate logistic becomes multivariate when there are more than one independent variables at a model.]

3. SVM (Support Vector Machine)

SVM is a technique of device gaining knowledge of regression and category. Modelling SVM includes steps: education a facts set to construct a version, after which predicting records from a take a look at facts set with the aid of using the usage of that version. The SVM version depicts the education facts factors as factors withinside the n-dynamically spatial variety, then maps them in a manner that separates the factors of diverse lessons from the broadest variety possible. Kernel plays a vital role in classification and is used to analyse some patterns in the given dataset. They are very helpful in solving a non-linear problem by using a linear classifier**.** The function of a kernel is to require data as input and transform it into the desired form. Different kinds of kernel functions are used in different SVM algorithms. Some of kernel functions are:

i. Gaussian Radial Basis Function (RBF)

It is one of the most preferred and used kernel functions in SVM. It is usually chosen for non-linear data. It helps to make proper separation when there is no prior knowledge of data.

ISSN 2394 - 7780

Formula:

F(x, xj) = exp(-gamma * ||x - xj||^2)

ii. **Sigmoid Kernel**

It is mostly preferred for **neural networks**. This kernel function is similar to a two-layer perceptron model of the neural network, which works as an **activation function** for neurons.

Formula:

F(x, xj) = tanh(αxay + c)

iii. Linear Kernel:

It is the most basic type of kernel, usually one dimensional in nature. It proves to be the best function when there are lots of features. The linear kernel is mostly preferred for **text-classification problems** as most of these kinds of classification problems can be linearly separated.

Formula:

F(x, xj) = sum(x.xj)

4. Random Forest Regression:

This set of rules is primarily based totally on supervised gaining knowledge of which makes use of ensemble method to get the correct results. It makes use of a couple of selection timber and method to mix a couple of selection timber with a view to get correct very last output instead of counting on unmarried selection tree. Major downside of selection tree is that it reasons over becoming. This downside is minimized even as the usage of a couple of selection tree in random wooded area regression.

5. **K-means clustering:**

K- means clustering is an unmanaged gaining knowledge of set of rules that is used to clear up the clustering issues. It organizations the unlabelled dataset into exclusive clusters. Unlabelled facts is portions of facts that don't have precise traits, properties, etc. This set of rules is centroid primarily based totally set of rules, wherein every cluster is related to the centroid. The goal of this set of rules is to discover the minimal sum of distances among the facts factor and their corresponding clusters.

Flowchart for rainfall prediction:

The under facts flowchart outlines the choice of enter and the technique to get the output that during this observe is the rainfall prediction. The flowchart illustrates step one as a choice of inputs which might be parameters, processing of those inputs and finishing education, trying out and validation for correct and specific output this is rain forecast.

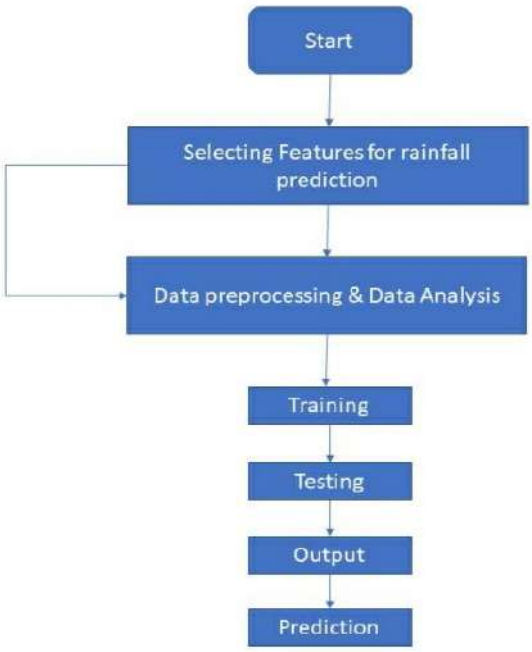


Figure 2: Flowchart 1

ISSN 2394 - 7780

Evaluation Techniques for prediction of rainfall

1. Confusion matrix

A confusion matrix is N x N matrix used for comparing the overall performance of a category version, wherein N is the wide variety of goal classes. The matrix compares the real goal values with the ones expected through the system studying version. This offers us a holistic view of the way properly our category version is acting and what forms of mistakes it's miles making

2. PRECISION

Precision refers back to the wide variety of actual positives divided through the whole wide variety of nice predictions (i.e., the wide variety of actual positives plus the wide variety of fake positives). Real international fashions by no means have 100% precision.

3. RECALL

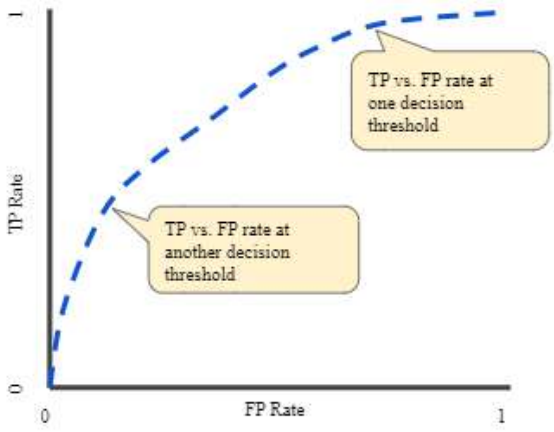
Recall on this context is described because the wide variety of actual positives divided through the whole wide variety of factors that truly belong to the nice magnificence (i.e. the sum of actual positives and fake negatives, which can be gadgets which have been now no longer labelled as belonging to the nice magnificence however must have been).

4. ACCURACY

Accuracy is a metric that summarizes the overall performance of a category version because the wide variety of accurate predictions divided through the whole wide variety of predictions.

6. Receiver operating characteristic curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives. The following figure shows a typical ROC curve.



ROC Curve 1

- True Positive Rate:

True Positive Rate is a synonym for recall.

$$TPR = \frac{TP}{TP + FN}$$

- False Positive Rate:

False positive rate (FPR) is a measure of accuracy for a test.

$$FPR = \frac{FP}{FP + TN}$$

i.

ISSN 2394 - 7780

i. True Positive: True Positive: the truth is positive, and the test predicts a positive.

ii. True Negative: The truth is negative, and the test predicts a negative.

iii. False Negative: The truth is positive, but the test predicts a negative.

iv. False Positive: The truth is negative, but the test predicts a positive.

METHODOLOGY

Initially, historic facts become accumulated which consist of various attributes. One of the demanding situations that face the know-how discovery technique in rainfall facts is negative facts quality. We get rid of the lacking price records. In our facts we've little lacking, due to the fact we're operating with rainfall facts. Applying pre-processing and remodeling the rainfall facts specific algorithms are used. Algorithms which we used on this prediction is Linear regression, Logistic regression, Support Vector Machine (SVM), Random Forest regression and K-method fashions. Precision & Recall could be examined and Comparison could be made. The dataset includes rainfall facts from year 1901 to 2019. It considers handiest monsoon months i.e. June, July, August, September & common rainfall from June-September.

Here four regression fashions & 1 category version is used. 80% of the facts set is applied for education to evaluate those fashions, at the same time as 20% are used for validation and trying out. The overall performance of the 4 classifiers is classed the usage of accuracy, sensitivity, specificity, precision. In each set of a pattern the actual nice, actual bad, fake nice and fake bad charges are represented withinside the desk under and a confusion matrix layout is likewise shown. The precision and specificity rankings of numerous actual negatives are inaccurately excessive withinside the desk.

CONCLUSION

From the study, We would like to conclude that the algorithms which we used can be used to perform prediction model. Here, we have taken features, perform analysis and test the algorithm to get accurate results.

REFERENCES

1. Akshay R. Naik; A. V. Deorankar; Premchand B. Ambhore, 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA).
2. Ogochukwu Ejike; David L. Ndzi; Abdul-Hadi Al-Hassani, 2021 International Conference on Communication & Information Technology (ICICT).
3. Heuristic prediction of rainfall using machine learning techniques Chandrasegar Thirumalai; K Sri Harsha;M Lakshmi Deepak; K Chaitanya Krishna 2017 International Conference on Trends in Electronics and Informatics (ICEI).
4. A Novel Study of Rainfall in the Indian States and Predictive Analysis using Machine Learning Algorithms Nikhil Tiwari; Anmol Singh 2020 International Conference on Computational Performance Evaluation (ComPE).
5. A support vector regression model for forecasting rainfall. Nasimul Hasan;Nayan Chandra Nath;Risul Islam Rasel 2015 2nd International Conference on Electrical Information and Communication Technologies (EICT)
6. S Aswin, P Geetha and R Vinayakumar, "Deep Learning Models for the Prediction of Rainfall", International Conference on Communication and Signal Processing, pp. 0657-0661, April 3–5, 2018.
7. https://en.wikipedia.org/wiki/Machine_learning#Artificial_intelligence.
8. Moulana Mohammed, Roshitha Kola Palli, Niharika Golla, Siva Sai Maturi, “Prediction of Rainfall Using Machine Learning Techniques.”, International Journal Of Scientific & Technology Research Volume 9, Issue 01, January 2020.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

MORPHOLOGICAL ANALYZER FOR ENGLISH NOUN FORMS

Manisha Divate, Sonal Marthak and Viral Malvania
Usha Pravin Gandhi College of Arts, Science and Commerce, Vile Parle, Mumbai

ABSTRACT

Stemming is a method of reducing inflected words to their root or stem form. This is applicable for both the suffix as well as prefix Stemming is a text mining preprocessing step that is commonly used for Natural Language Processing (NLP). A stemmer can perform the operation of morphologically identical words being changed to root words without performing morphological analysis on that term. It is useful for improving performance in many areas of computational linguistics and information retrieval work. The idea of stemming is to improve the efficiency of information retrieval. Stemming is the way of eliminating a word's inflectional and sometimes derivationally related forms to a single root form. In this paper we have discussed about Suffix-prefix stripping method, and we obtained accuracy of 65.5% for morph analysis of noun phrases.

Keywords: Natural Language Processing, Morphological analysis, Stemming, WordNet Lemmatize

INTRODUCTION

Morphology is the part of linguistics that analyses word structure. The set of categories and rules involved in word formation is known as morphology. [9] Understanding words is fascinating because it is vital for human society. It is hard to imagine a human language with no words at all. Many definitions of word have been established, and they can be found in dictionaries or linguistic textbooks.

Words are the smallest meaningful unit of a language [9]. The words are distinct in both pronunciation and meaning. It signifies that the term has the simplest possible meaning in linguistics and may stand alone without further explanation. For example: word; sleep, teach, write, etc. The words "sleep," "teach," and "write" cannot be split into smaller parts that can convey meaning on their own. They are just a part of a sentence that has a function to convey the meaning if they stand with other elements in a sentence [12].

As indicated above, Inflectional and Derivational morphemes comprise of bound morphemes. The morphemes that do not create new meaning are known as inflectional morphemes. The syntactic category of the words or morphemes to which they are linked is never changed by these morphemes. They simply enhance the previously established meaning of the words to which they are attached by refining and providing more grammatical information [12].

What is Morphological Analysis? What is the use of Morphological Analyzer?

LITERATURE WORK

Vocabulary knowledge and learning strategies

Words are significant aspects of linguistic knowledge and are part of a speaker's mental grammars. The smallest meaningful unit of a language are words. People can create phrases, paragraphs, texts, and even discourses by starting with words. When people study words, they discover certain parts of the words, such as the root, stem, base, morpheme, syllable, prefix, and suffix. Morphology, a part of linguistics that analyses the nature and arrangement of morphemes to create words, includes all of them. Morphology is concerned with the process of word formation, specifically the methods in which various types of words are generated. It also explains the phonological process that occurs during word formation. Morphology discusses the methods that lead to the formation of words, specifically the patterns in which various kinds of words are formed. In short, morphology is the study of word formation, whereas morphological analysis is the study of how words are formed. The majority of English words are created through affixation, which involves adding a prefix, a suffix, and an infix to the root of the words [9].

MORPHOLOGY

Morphology has vital impact on the development and perception of English Words. Morphemes are the smallest elements of word that convey meaning. That include roots, stems, prefixes and suffixes. The capacity to use this moderate level of dialect is essential to build an overwhelming vocabulary and grasping English content. Morphology refers to the implementation of morphemes, the components of words that contains the values. The precise role of morphology varies by language, depending on the word arrangement forms used in each regional language. In terms of English, morphemes supply the raw materials for the creation of new words, and expertise of morphemes leads to the English language's generative force. Numerous new words are universally recognizable since they are derived from well-known morphemes [7].

ISSN 2394 - 7780

- Basic word classes (parts of speech)
- Nouns: people, library, industry, etc.
- Verbs: start, walk, think, etc.
- Adjectives: sweet, warm, tough, etc.
- Adverbs: carefully, innocently, etc.

There are different types of Morphemes

Inflectional Morpheme – Various combinations of the same word are created by inflection. For example:

Verbs: to be, being, I am, you are, he is, I was.

Nouns: One book, two books.

Derivational Morpheme – Derivation creates different words from the same lemma:

grace → disgrace → disgraceful → disgracefully

Prefixes and Suffixes are the most common affixes[8].

Prefix: An affix that goes before a root(base) word.

Examples re-, un- (re-read, un-loved)

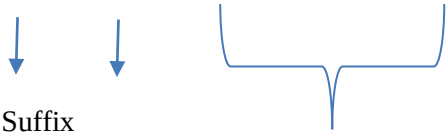
Suffix: An affix that goes after a root(base) word.

Examples -est, -er, -s (quick-est, quick-er, read-s, book-s)

sign + s → signs

base suffix

de + sign + ate/ + ion → designation



Suffix

prefix base

RESEARCH QUESTIONS

Morphological awareness refers to the ability to identify the significance and structure of morphemes that are part of or linked with a word. The relationship between morphology and vocabulary knowledge will be examined in this study. Two sets of research questions were addressed in this study:
1. Does morphology contribute to vocabulary teaching and learning process?

2. Do morphological strategies help learners to enlarge their vocabulary knowledge?

METHODOLOGY

The proposed method contains suffix-prefix stripping algorithm for Noun form of English Language.

After understanding linguistic behavior of English grammar, the authors have identified the most appearing affixes and studied their behavior. Affixes are as shown in Table 2.

Some of the most frequently used Suffixes are shown in Table 3 and Prefixes are shown in Table 4.

TABLE 1. Stemming Analysis Of Different Languages.

System	Methods	Approach	Language	Dataset
[1]	Longest Matched	Rule based	Hindi	Online Hindi News,

ISSN 2394 - 7780

				Magazines, Films, Health, business, Sports & Politics
[2]	Take-all-split method	Hand-crafted Suffixes	Gujarati	EMILLE Corpus
[3]	n-gram Method	Suffix Stripping	Marathi	Marathi Corpus internet
[4]	Take-all-split, POS based	Suffix Stripping, Rule Based	Gujarati	EMILLE corpus
[5]	Brute force technique	Suffix Stripping	Punjabi	Online Newspaper, Dictionaries, Articles
[6]	Longest Matched	Rule Based	Gujarati	EMILLE corpus

TABLE 2. Affixes Used In Word Formation

Prefixes	Suffixes
Auto-, Dis-, Mid-, Non-, Over-, Out-, Super-.	-ies, -es, -ss, -s, -ment, -eer, -ee, -or, -ion, -ship, -th, -ix, -i, -tion, -sion, -ant, -ent, ity, -ness, -ing, -ance, -al, -ial, -ism.

TABLE 3. Most Common Suffixes Used In Word Formation

Suffix	Example	Occurrence	Accuracy
-ing	Wedding, Warning, Holding, Undertaking, Meeting, Working, Shooting, Opening.	38	100%
-ity	Personality, Security, Reality, University, Quantity, Possibility	15	36%
-ment	Agreement, Employment, Judgment, Management, Measurement	32	70%
-ion	Action, Association, Collection, Connection, Inspection, Selection	26	48%

TABLE 4. Most Common Prefix

Prefix	Example	Occurrence	Accuracy
Auto-	Autosuggestion, autocorrect, automobile, autorotation	24	55%
Over-	Overcontrol, Overcharged, Overconfident, Overdressing	20	48%

ISSN 2394 - 7780

Fig. 1 Describes the approach used in English Stemmer.

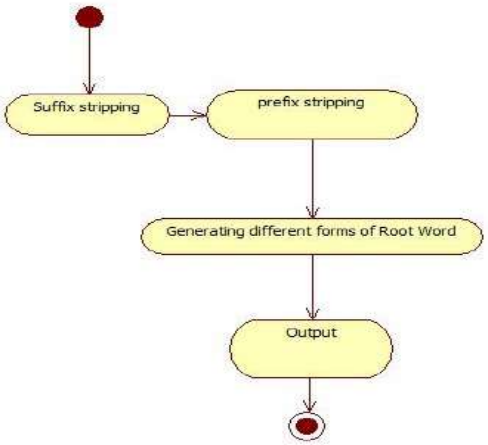


Fig. 1. Suffix-Prefix Stripping Approach

RESULT AND DISCUSSION

English has a morphologically diverse vocabulary. Which has so far included execution of morphological investigations in the English language. Most of them are descriptive in nature. The morphology of English has not been thoroughly studied from a computational perspective. Due to which there exist some insufficient work from a computational point of view.

After having been corrected and analyzed, some errors were found and most of the errors are about derivational in class word. Proposed work of English Morphological Analyzer using Brute Force Technique considering Suffix-Prefix stripping approach has an accuracy of 65.5% as shown in Table 5. The problem faced in remaining 35.5% was some exceptional cases which needed different set of rules.

TABLE 5. Result Analysis Of Proposed System

Tested on Language	Total Words	Proposed Method	Approach Used	Accuracy
English	1000	Brute Force Technique	Suffix-Prefix Stripping	65.5%

The proposed stemmer uses nltk,pattern lemmatizer as a base tool which fails to produce proper lemma due to which our system fails in identifying the proper lemma form a word . For example if we give “Application” word as input to the system and it should give “Apply” as output but it fails to lemmatize .

CONCLUSION

This paper presents conceptual framework concerning Morphological analysis. This analyzer is based on Suffix-Prefix Striping method. The stemmer performs with an accuracy of 65.5% considering noun form of words with Suffix and prefix. The Analyzer can classify the Root word, Parts of Speech (POS), Suffix and Prefix. n the stripping of the morphemes the various morphemes pattern combinations are tested. POS tagging is very much depended on Morphological Analysis and lexical rules of each category. The stemmer can be improved so as to minimize under stemming and over stemming by enriching the rule set. Apart of this, a hybrid approach, i.e. usage of rule set along with statistical approach can be applied so as to generalize the stemming process [10].

REFERENCES

[1]. Ramanathan, A. And Rao, D. 2003. “A lightweight stemmer for Hindi”. In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL), on Computational Linguistics for South Asian Languages (Budapest, Apr.) workshop.

[2]. Pratikkumar patel kashyap popat “hybrid stemmer for gujarati” in proc. of the 1st workshop on south and southeast Asian natural language processing (wssanlp), pages 51–55, the 23rd international conference on computational linguistics (coling), Beijing, august 2010.

[3]. Mudassar M. Majgaonker, Tanveer J Siddiqui. “Discovering suffixes: A Case Study for Marathi Language” in International Journal on Computer Science and Engineering. Vol. 02, No. 08, 2010

ISSN 2394 - 7780

[4]. Kartik Suba, Dipti Jiandani and Pushpak Bhattacharyya "Hybrid Inflectional Stemmer and Rule-based Derivational Stemmer for Gujarati" 2011

[5]. Dinesh Kumar, Prince Rana. (2011) "stemming of Punjabi words by using brute force technique", International Journal of Engineering Science and Technology (IJEST) ISSN

[6]. Juhi Ameta, Nisheeth Joshi, Iti Mathur (2012) "A Lightweight Stemmer for Gujarati"

[7]. Akbulut, D. F. (2017),"Effects of Morphological Awareness on Second Language Vocabulary Knowledge. Journal of Language and Linguistic Studies", Published by JLLS, 13(1), 2017

[8]. Dr. John R. Kirbcy and Peter N. Bowers "What works Research into practise" published by Literacy and Numeracy Secretariat and the Ontario Association of Deans of Education , june 2012

[9]. Adisti Herliningtyas, "A Morphological Analysis on English Derived Verbs Using the Suffix -ize," July 2008

[10].Jikitsha Sheth and Bankim Patel, "Dhiya: A stemmer for morphological level analysis of Gujarati language," February 2014

[11].R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[12].Nurhikmah, Widya Lestari and , Drs. Sigit Haryanto, M. Hum and , Dra. Siti Zuhriah, M. Hum (2018) An Analysis On The Derivational And Inflectional Affixes Found In Facebook 'Bbc News' September, 8 2017. Skripsi thesis, Universitas Muhammadiyah Surakarta.

[13].M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

ROBOTIC PROCESS AUTOMATION: TRANSACTIONAL PROCESS OPTIMIZATION USING DATABASE THROUGHMULTI-BOT ARCHITECTURE

Joveria Mirza
Beg Security Consultant, Cyber, IAM

ABSTRACT

Robotic Process Automation (RPA), in other words, this is the technology that deals with the applications of machines and computers to generate good amount of revenue and profits. With the help of this tool, one can automate the processes which are rule based and repetitive in nature and does not require any kind of human intervention. So, the present study focuses on UiPath Tool which is based on RPA. UiPath is a tool for RPA technology partners who develop and deliver software that helps to automate business processes. UiPath Orchestrator is a web- based application that runs and manages UiPath Robots. It is capable of deploying multiple Robots, making queues, assets, getting interactive with the action center, storage bucket and monitoring and inspecting the activities of UiPath Robots. UiPath Robot is an application service. One of the important components of UiPath is that it can open interactive/non-interactive window sessions to execute processes which are developed using UiPath Studio as well as UiPath Recording method. Sometimes, it is also called a performance agent as it carries out automation projects, or even called a runtime robot as it executes process generated by developing or recording processes in UiPath Studio. Author is in the RPA industry, and has worked on many clients' side projects (mostly Banks) and faced many problems while developing transaction based operations such as: - a)Per day frequency of transaction is high, b) Due to security reasons, Banks will not allow outside web-portal accesses such as Orchestrator, c) To transact per data, UiPath Robot takes 2-5 mins to complete it, d)Bank prefers to use only one robot per process, they won't be able to use another robot for another process because process is not made in multi-bot architecture even though many robots are available to utilize, e) Banks mostly use database to store the data. To overcome the above-mentioned problems, we have come up with Multi-Bot Architecture using Database which has never been done before, it will also revolutionise thetransactional processes which uses database.

Keywords: RPA, UiPath Tool, Robot, Multi-Bot, Architectures, Database, SQL, TransactionItem, Stored Procedure, Queues.

I. INTRODUCTION

As we are in the RPA industry from the past 2years and have mostly worked with banking clientsand also s ince the banking sector is the most securedsector than any other sector - we have analyzed thatm o s t o f t h e process of a bank is transactional, and the frequencyof transactions vary from process toprocess and banks to banks but one thing is commonamong them is number of transactions we get toprocess is high and mostly repetitive as banks have millions of customers/Accounts. Due tothe above mentioned reasons some of the banks donot find i t promising to Implement RPA in thebanking Sector. Hence, we came up with Multi-Bot Architecture using Database solution to gain the trustof m o r e n u m b e r o f banking sectors in orderto implement RPA. However, without a doubt - outof many sectors, banking sectors are the one's who adapted RPA the most.

Now-a-days companies who provide RPA as services are mostly targeting the banking sectors and the reason is that maximum number of tasks which are repetitive in nature can be found. Since we have faced this challenge on so many locations, we decided to find the permanent solution for it. Also, the banks that we have worked for are not using one of the UiPath Components i.e. Orchestrator. If they would have been using that component, it would have been much easier for us to implement the Multi- bot Architecture. But that is not the case over here, in all the processes they are using database to keepand store the processed data. Consider one example, The frequency of per day transaction is around 1,000 and these transactions needs to be completed within a day only. To complete the whole process, every bank has an entire departmentto achieve the per day goals. Manual operation takes around 10-15 minutes to complete each transaction, whereas the same task if performed by the robot would take around 5-7 minutes to complete each transaction. Here, we are saving almost 50% of the execution time, yet it is not feasible by the bot to complete these many transactions per day.

Since we are facing the same challenge over and over again, we decided to create our own Multi-bot Architecture using Database. The reason behindadopting this method is to reduce the process execution time and optimize it to such a level that even if in the future, the frequency of the transactions will increase but it will not affect the process execution time that much.

ISSN 2394 - 7780

II. BACKGROUND AND MOTIVATION

A. Background

Now-a-days, there is not a single which are not affected by the automation system. Few of the examples involve washing machines, microwave ovens, autopilot mode for automobiles and airplanes, nestle using Robots to sell coffee pods in stores in Japan, Amazon using drones to deliver products in the US, our bank checks being sorted using (OCR Optical Character Recognition), and ATMs.

The term automation is derived from the Greek words, autos meaning self, and motos, meaning moving. Automation, in other words, this is the technology that deals with the applications of machines and computers to generate good amount of revenue and profits. With the help of this tool, we can automate the processes which are rule based and repetitive in nature that does not require any kind of human intervention.

With the arrival of computers, many software systems have been developed to accomplish tasks that were previously done on paper to manage businesses, or not being done at all due to the lack of tools. Some of these are bookkeeping, inventory management, and communications management which are now has been replaced by the latest tools are called as Enterprise Resource Planning (ERP) like SAP, Oracle cloud ERP, NetSuite ERP, Microsoft Dynamics andmany more.

B. Motivation

In digital marketing and e-business, RPA organizations want to target large amount of customer with high successrate of consumer retaining.

Banks have large number of repetitive tasks, which can be handled by Robots easily and efficiently. So, we want toachieve this task with more efficiency, accuracy, timesaving, Long-term profits to end-user.

III. RELATED WORK

Related work of this particular research is based on thereal-life experience in which we have mentioned few of thebelow points:

1. What is Orchestrator, Queue and Assets
2. How multi-bot is been implemented using queue
3. Limitations and Disadvantages of Multi-bot Architecture using Queue
4. How we can overcome the limitations and disadvantages of Orchestrator Queue using Database Multi-bot Architecture

A. UiPath Orchestrator

It is a web-based application that lets you manipulate or manage your robots, hence the name Orchestrator. It runs on a webserver and connects to all the Robots which are connected to that particular network or orchestrator, whether attended, Unattended, or Free. It has a web-based interface that enables the orchestration and management ofhundreds of Robots with a click. Orchestrator lets you manage the creation, monitoring, and deployment of resources in your environment, and also can manage to divide the operation evenly among the robots which can help to achieve to complete the process not only on time but before it, action center which allows to assign particular document understanding or form task to other users which are available in the orchestrator. Orchestrator acting in the same way as an integration point with third- party applications. Orchestrator's main capabilities:

a) It helps in creating and maintaining the connection between Robots

b) Itensures the correct delivery of the packages to Robots

c) It helps in managing the queues

d) It helps in keeping track of the Robot identification and generate the reports for of the robots like results of successful and unsuccessful transaction which can be useful for the end- user to find out the accuracy of the bots and helps to generate reports of return on investment (ROI) e) It stores and indexes the logs to SQL or Elasticsearch.

B. QUEUE

Queues work as a data bucket that stores tasksthat need to be implemented and executed by the robots. To simplify the same in simple words we can take one example for the same, consider there is group of people formed a line for to purchase ticket in frontof ticketing counter. The logic of queue will work like the person

ISSN 2394 - 7780

who is standing line first will get the ticket first and will be out from the same line first. The same thing applies in the queue. Consider line as queue and people are task that need to be perform, then the ticketing person will be assume as robot who will call person one by one and complete its task according to their requirements. This method is also known as first in First out (FIFO).

Similarly, in the case of Robots, when we have a number of operations that are to be performed and when the server is busy, then tasks are moved in a queue and waits until the server is free once the server is free then robot will perform its work accordingly, and they are implemented on the same logic First in First Out (FIFO). To create a new queue, search for the Queue option in the Orchestrator Server listed on the Top side of the website in Shared Folder or My Workspace Folder or in My Folder then inside the Queue page, you can add one. It gives you the permission to access all those Queues that have already been created or been created by the user to perform necessary operations on it. It contains some information about the task such as the remaining time, progress time, average time, description, and so on. Wecan trigger another process with help of queue base trigger. When new transaction has been added inside the queue, the queue will trigger the process which has been mapped with the particular queue name.

We can also add queue items from UiPath Studio and there are various activities that helps to add data in the queue, which are listed as follows: a) Add Queue Item: This activity is helps to add new item in thequeue and its status will be set as new. b) Add Transaction Item: This activity is used to add an item to the queue to start the transaction and set the new item status as "In Progress". We can also add custom reference for each respective transaction. c) Get Transaction Item: The purpose of this activity is to get an item from the queue to execute and process on it and set its status as In Progress so that another robot does not take the same transaction to process to avoid the same transaction process execution. d) Postpone Transaction Item: The purpose of this activity is to postpone the transaction to be executed and set its status as postpone and as input parameter it takes "QueueItem" datatype. e) Set Transaction Progress: The purpose of this activity is to set the custom progress of the particular transaction item and change the status of "In Progress" transaction item as per user requirement. f) Set Transaction Status: The purpose of this activity Is to change the status of particular transaction item to either successful or failed.

C. Assets

Assets are of two types:

1. Get Asset
2. Get Credential

The Get Asset and Get Credential activities are used in Studio, and the purpose of this activities are to request information from Orchestrator about a specific asset, based on its provided Assets Name. The name of the asset is required in order to fetch the data from already existed assets which are available or created in orchestrator database, so that robot can fetch the information from asset in order to use the same data in the process. To perform the same to get asset information from orchestrator robots needs to have a permission to fetch information which are beingused in automation project.

D. How Multi-bot Architecture is been implementedusing Queue

First, we build a process according to its businesslogic, and then we integrate the implemented process in the RE-framework architecture. Then we deploy this same integrated RE-Framework process into different machines. After the deployment, all the machines will hit the same orchestrator queue and get the transaction item from it and start processing the data as per process. While one transaction is in process the other machines will not pick the same transaction data, theywill pick the next transaction data from the QueueItems and start processing the same. And after the transaction is completed the same will be updated to the orchestrator.

This process will repeat until all the transactionsin queue are not completed. This is an in-built functionality of orchestrator queue, once the particular transaction item comes into action it will change itsstate from new to pending in background and on the basis of that bot will be able to identify whether particular transaction is in useor not, if it is in use then robot will skip that transaction item and take another one if it is available.

E. Limitations and Disadvantage in Multi-bot Architecture using Queue

At least one in-built function has some the disadvantages and here we have the live case for the same.Consider that one process has been published on multiple production environment to execute the same task and divide the transaction items to complete the process as early as possible. Before we fetch the data form transaction item, we have to publish the data to the queues then and only robot will be able to fetch the same data in the flow, and

ISSN 2394 - 7780

here comes the major disadvantage that is data duplication. While uploading the data in queue, the queue does not have the functionality like database toavoid data duplication.

Even if we pushed the same data again, we could do nothing about it unless the same transaction gets complete, and queue get empty. Whereas we can handlethe same scenario in database level and put some validation point where we can check for the data duplication and avoid the same.

As we said earlier, consider we have frequency of per day transaction is huge then we cannot afford to perform the operation on the same transaction item. Also, if we consider that, this process is of bankingsectors then bank will either face the loss of money or the loss of time. So, for the same reason this could be the biggest disadvantage of the queue based multi-bot architecture.

To use the orchestrator queue, we need to install orchestrator in the environment. Since the orchestratoris such an expensive tool, which is not possible some of the clients to purchase the same. Standard Orchestrator (basic) cost around 25000$ which is current price, and it goes up to the 50000$ according to the feature and functionality you want in theorchestrator, but we can avoid this part and still canbuild the same using database and it is not that muchcostly that client could not afford.

F. How we can overcome the limitations anddisadvantages of Orchestrator Queue using Database based Multi-bot Architecture

In the previous section we discussed about the limitations and disadvantages of queue based multi- bot architecture, where we found some of the disadvantage of the architecture which could lead the process to major loss of any kind with respect to its sector. With the help of database, we can overcome all the points which has been mentioned in previous point i.e., Data Duplication, no need to setup and internet- based application (Orchestrator), intrawork network would be fine if we use database and so on.

We can resolve so many obstacles by just changing the one application and we can still achieve the same thing which we can achieve in Queue based Multi-bot Architecture.

IV. METHODOLOGY

Here we propose Re-framework architecture which is in-built framework in UiPath Studio. We have re-structured this framework so that we can integrate the multi-bot architecture into it. RE-Framework basically has 4 states i.e.,

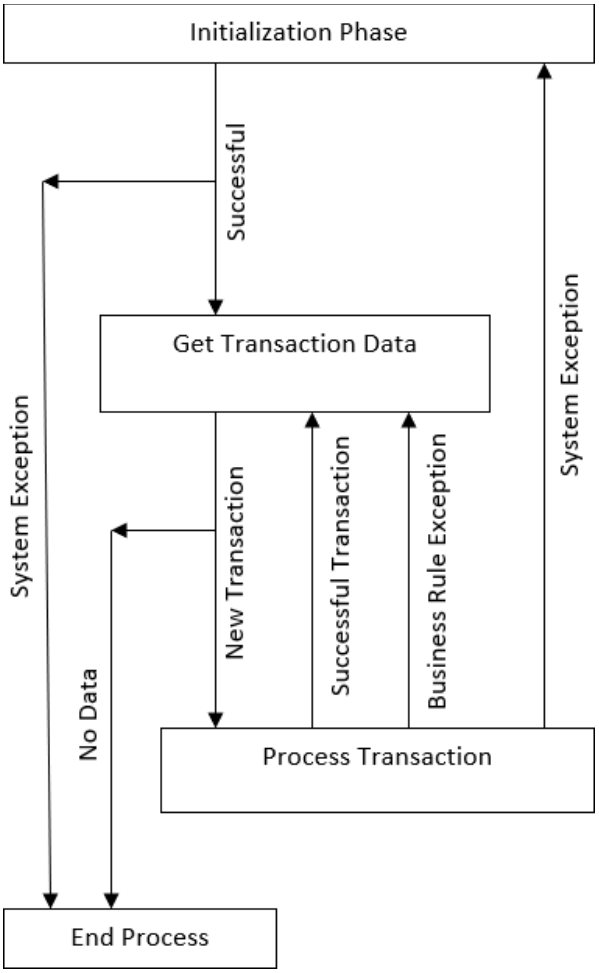
1) Initialization State, 2) Get Transaction State, 3) Process State and 4) End Process State.

The working mechanism of RE-Framework is like, first the bot will go to the Init State in which robot initialize the necessary variables, read the configurationfiles, initialize, and store the data into the dictionary and initialize the application that need to be open.

Once this state is done then robot will go to theGet Transaction State in which robot will fetch thesingle data from database and set the status of samedata into 'In Progress' and store inside the'TransactionItem' variable. Then robot will go to the Process State and perform task which has been designed according to business logic. Once the processstate is executed, after that, 3 transitions come into thepicture where robot will navigate to each transition is dependent upon the process state. If process state is executed successfully then robot will navigate to theSuccessful Transition and update the status of fetcheddata to 'Completed' with no exception message. If process state is executed with business exception, then robot will navigate to BusinessException Transition and update the status of fetched data to 'Failed' with business exception message and navigate back to GetTransaction State to get next data.

If process state is executed with system exception, then robot will navigate to SystemException Transition and update the status of fetched data to 'Failed' with system exception message and navigate back to Get Transaction State to get next data. Afterthat robot will navigate to Get Transaction State and try to fetch the data, if data is available then robot will perform the same steps mentioned above and if no data available, then robot will navigate to End Process State where we close the all the opened applications and finish the process. We can even generate the fallout report or exception report in the End Process state.

ISSN 2394 - 7780



Algorithm 4.1: Algorithm for Get Single Query at a Time.
Input: RoboName, ProcessedByOutput: Refined table.
Step 1: Declare the Variable RoboName Varchar(MAX)and ProcessedBy Varchar (MAX).

Step 2: Fetch the single data from its desired table.

DECLARE @Sbl Varchar (50)

SELECT TOP 1 @Sbl = Symbol from Yahoo_Finance WHERE [Robo Name] is null AND Status is null

Step 3: Update the same fetched query and set its status to 'In Progress' state

Update Yahoo_Finance SET [Robo Name] = @RoboName, [Processed By] = @ProcessedBy, Status ='in progress' WHERE Symbol = @Sbl

Step 4: Again, select the same query to perform operation on it.

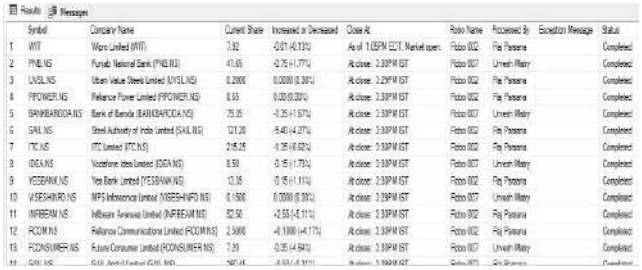
SELECT TOP 1 * FROM Yahoo_Finance WHERE [Robo Name] is not null and Status = 'in progress' and Symbol = @Sbl

Step 5: Refined Table.

This algorithm 4.1 first fetches the top 1 query from table whose RoboName and Status is null and store it into one variable for further query use. Then we will update the data of RoboName = @RoboName, ProcessedBy = @ProcessedBy, Status = 'In Progress'column using where condition in which we will pass@Sbl variable which we declared earlier, and againselect the same data whose status we update as 'In Progress' and give the same as output datatable to proceed further to capture the necessary details form website.

ISSN 2394 - 7780

TABLE 1



	Symbol	Company Name	Current Share	Increased or Decreased	Close At	Robo Name	Processed By	Exception Message	Status
1	WIT	Wipro Limited (WIT)	7.92	-0.01 (-0.13%)	As of 1:05PM EDT. Market open.	Robo 002	Raj Parsana		Completed
2	PNB.NS	Punjab National Bank (PNB.NS)	41.65	-0.75 (-1.77%)	At close: 3:30PM IST	Robo 007	Umesh Mistry		Completed
3	UVSL.NS	Uttam Value Steels Limited (UVSL.NS)	0.2000	0.0000 (0.00%)	At close: 3:29PM IST	Robo 002	Raj Parsana		Completed
4	RPOWER.NS	Reliance Power Limited (RPOWER.NS)	0.55	0.00 (0.00%)	At close: 3:30PM IST	Robo 002	Raj Parsana		Completed
5	BANKBARODA.NS	Bank of Baroda (BANKBARODA.NS)	75.35	-1.35 (-1.67%)	At close: 3:30PM IST	Robo 007	Umesh Mistry		Completed
6	SAIL.NS	Steel Authority of India Limited (SAIL.NS)	121.20	-5.40 (-4.27%)	At close: 3:30PM IST	Robo 002	Raj Parsana		Completed
7	ITC.NS	ITC Limited (ITC.NS)	215.25	-1.35 (-0.62%)	At close: 3:30PM IST	Robo 002	Raj Parsana		Completed
8	IDEA.NS	Vodafone Idea Limited (IDEA.NS)	8.50	-0.15 (-1.73%)	At close: 3:30PM IST	Robo 007	Umesh Mistry		Completed
9	YESBANK.NS	Yes Bank Limited (YESBANK.NS)	13.35	-0.15 (-1.11%)	At close: 3:30PM IST	Robo 002	Raj Parsana		Completed
10	VISESHINFO.NS	MPS Infoscience Limited (VISESHINFO.NS)	0.1500	0.0000 (0.00%)	At close: 3:29PM IST	Robo 007	Umesh Mistry		Completed
11	INFIBEAM.NS	Infibeam Avenues Limited (INFIBEAM.NS)	82.50	+2.55 (+3.11%)	At close: 3:30PM IST	Robo 002	Raj Parsana		Completed
12	RCOM.NS	Reliance Communications Limited (RCOM.NS)	2.5000	+0.1000 (+4.17%)	At close: 3:30PM IST	Robo 002	Raj Parsana		Completed
13	FCONSUMER.NS	Future Consumer Limited (FCONSUMER.NS)	7.20	-0.35 (-4.64%)	At close: 3:30PM IST	Robo 007	Umesh Mistry		Completed

Refined Output of Algorithm 4.1

	Symbol	Company Name	Current Share	Increased or Decreased	Close At	Robo Name	Processed By	Exception Message	Status
1	WIT	NULL	NULL	NULL	NULL	Robo 007	Umesh Mistry	NULL	In Progress

Algorithm 4.2: Algorithm for Updating Transaction Data.

Input: Symbol, Company Name, Current Share,Increased or Decreased, Close At, Exception Message,Status.

Step 1: Update the Transaction Item with the Captured Data

UPDATE Yahoo_Finance SET

[C o m p a n y N a m e] = @CompanyName, [Current Share] =@CurrentShare,

[I n c r e a s e d o r D e c r e a s e d] = @INC_OR_DEC,[Close At] = @CloseAt, [Exception Message] = @ExceptionMessage,Status = @Status WHERE Symbol@Symbol

This Algorithm 4.2 will update the necessary columns which needs to be updated according to the business loginand these fields will be updated on the basis of the data which has been extracted form webpage like Company Name, Current Share, Increased or Decreased, Close At, Exception Message and Status to keep the track of recordsand status.

V. CONCLUSION

In this paper, we studied the better approach for multi-bot processes, which could best fit in the banking sectors and other relevant sectors. Also, howwe can minimize or nullify the disadvantage of queue based multi- bot architecture with the help of database based multi-bot architecture.

Database based multi-bot architecture approach reduces the processing time, increases the accuracy, avoid the data duplication unlike orchestratorqueue.

With this approach, if we increase the number of robots who work simultaneously on the same process will decrease the huge amount of time depending upon the number of robots allocated.

At Last, we have analyzed and got to the conclusion that :-

1. For Yahoo Process we have built as a sample with Multi-Bot Architecture with 1 Robot takes approx. 30:00 Mins.
2. For Yahoo Process we have built as a sample with Multi-Bot Architecture with 2 Robot takes approx. 12:00 Mins.

More than half of a time is saved.

Here we conclude that, Multi-bot architecture using database is better than Multi-bot Architecture using Queues with the concrete results.

ACKNOWLEDGEMENT

The authors would like to thank anonymous reviewers for their helpful suggestions and comments; this has led to significant improvements in this paper and numerous invaluable suggestions leading to a much better version of this paper.

REFERENCES

[1]. Alok Mani Tripathi,"Learning Robotic Process Automation". Packt, 2018.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

A PERCEPTION STUDY ON THE EMERGING TREND OF CREDIBILITY ASSOCIATED BY ADOLESCENTS TO POLITICAL NEWS STORIES ON DIGITAL MEDIA

Mayur Sarfare
Assistant Professor, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce

ABSTRACT
Social media has completely revolutionized the way we consume content, especially since the pandemic. This has had major implications and ramifications upon the production, distribution and consumption of online news content. The accessibility of online news content is incomparably swift when one looks at other kind of traditional media; however, this has also had a significant impact upon its accuracy, which in turn has affected its credibility. Of all the news beats appearing on social media, politics is the one that has drawn less scrutiny of researchers with respect to the level of credibility associated by its users, especially the adolescents. According to several studies, adolescents have displayed greater vulnerability when it comes to placing trust in in political news stories. This study intends to understand the perception of credibility associated by adolescents with respect to political news stories on the following social media platforms: WhatsApp, Facebook and Instagram. The research approach of this study is qualitatively quantitative and the primary aim is to measure Credibility on the basis of three sub dimensions: Believability, Fairness and Accuracy, using Likert's scale. The method of data collection shall be by the use of a questionnaire, which will be administered to one hundred respondents in the age group of 17-22. The findings of this study are that the level of credibility associated by young adults to political news stories appearing on the above-mentioned social media platforms are on the lower side.

Keywords: News, social media, politics, credibility, believability, accuracy, young adults, adolescents

INTRODUCTION
The rise of social media has democratized content creation and has made it easy for everybody to share and spread information online.

Social media has become one of the important platforms for interaction, information sharing and news consumption. Considering the medium is dynamic oriented, the authenticity and credibility associated with social media and its content is generally questionable.

Social media appears to become a collective voice of agreement, where users and friends within the network have the power to recommend good media contents amongst them, and trust and believability gets associated with the content and across the networks or communities residing in the Internet (Shahrinaz & Latif, 2013).

Often age affects credibility attitudes towards online content. Even personal preferences can affect the perceived credibility of internet.

The majority of the public follows news and information with great interest and frequency. And trust seems to be an important component related to how much people interact with news in general.

It's interesting to find out whether political news on social media platforms offer credible, accurate and objective news by the adolescents.

LITERATURE REVIEW
With increase in the popularity of social media, more and more people are consuming news from social media instead of traditional media.

Social media helps in creating propaganda through an established network. Generally, a person believes information on social media because the people he chooses to follow share things that is in synchrony with his belief system. Then, that person tries to share the same information with others in his network. With enough sharing, a social network accepts the propaganda story in effect. (Jarred, 2017).

New form of social technologies like Twitter, Facebook have facilitated rapid information sharing and cascades that can spread misinformation or information that is inaccurate or misleading (Soroush, Roy, & Aral, 2017).

Social networking sites and blogs has helped the news become a social experience in fresh ways for consumers. People use their social networks and social networking technology to filter, assess, and react to news. The ascent of mobile connectivity via smart phones has turned news gathering and news awareness into an anytime, anywhere affair for a segment of avid news watchers (Purcell, Mitchell, Rosenstiel, & Olmstead, 2010).

ISSN 2394 - 7780

Age is a chief contributor in the way people evaluate a news organization's online content. News readers of a relatively younger age group attach greater significance than older news readers. (2016)

People appear to judge attitude-consistent sources and balanced sources as comparably, if not equally, credible, and both credibility and selective exposure should be weighed equally (Metzger, Hartsell, & Flanagin, 2015).

College students heavily rely on the internet for academic and general information and they verify it only marginally. They find the information on net to be more credible than adult population (Metzger, Flanagin, & Zwarun, 2003).

OBJECTIVE

The aim is to investigate the degree of credibility associated by adolescents with social media for political news consumption on the basis of the three dimensions: Believability, Fairness, and Accuracy

HYPOTHESIS

H_0: Adolescents associate a high level of credibility with social media consumption for political news

Ha: Adolescents don't associate a high level of credibility with social media consumption for political news

RESEARCH QUESTION:

Is the credibility associated by adolescents significantly high with respect to political news stories on social media?

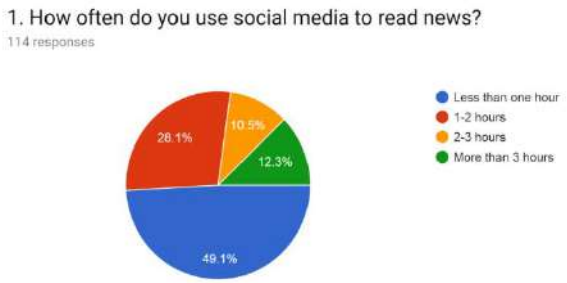
RESEARCH METHODOLOGY:

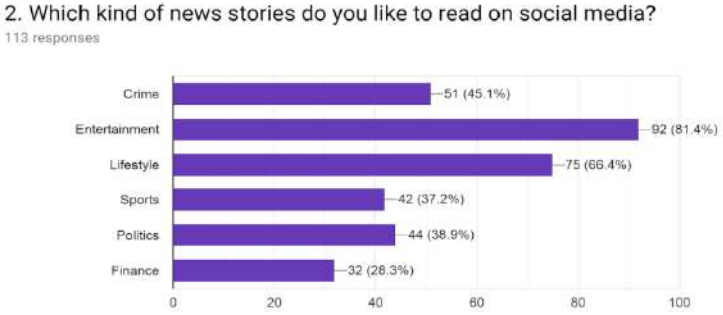
The approach adopted is quantitative as one of the objectives is to mathematically test the hypothesis. The tool of data collection used was a survey technique administered through the platform - Online Google forms. The questionnaire enables the researcher to determine whether the adolescents feel political news on various social media platforms of Whatsapp, Facebook and Instagram are they believable, credible and accurate. However, this approach doesn't not allow to obtain an in-depth exploration of the subjective experience derived by the respondents while consuming the news content.

Sampling Technique

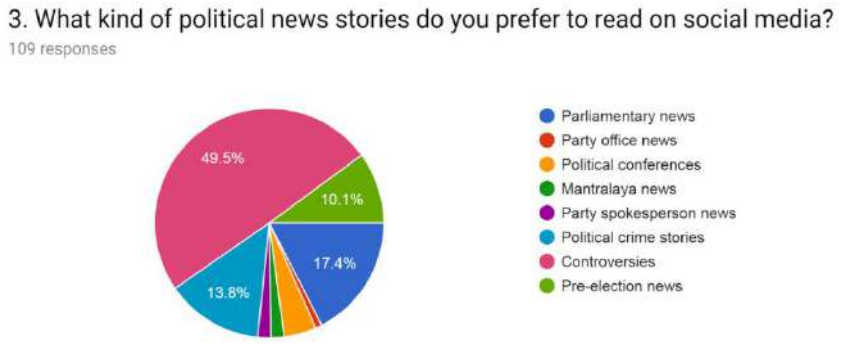
In order to conduct research among adolescents between age group of 18-20. Under non-probability sampling techniques, convenience sampling was chosen. The sample size decided upon was 114. The adolescents were students belonging to various streams of Bachelors of Management Studies, BSC IT and First year Bachelor of Mass Media. Care was taken that students who have been introduced to research and research methodology and various aspects of media and journalism have not been included in the research in order to avoid bias into the research.

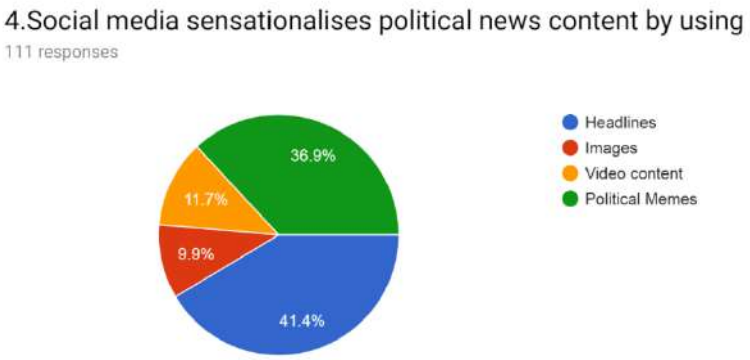
Observation



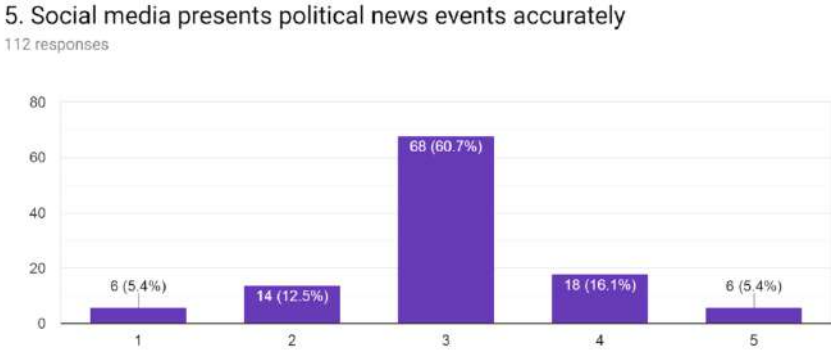


ISSN 2394 - 7780

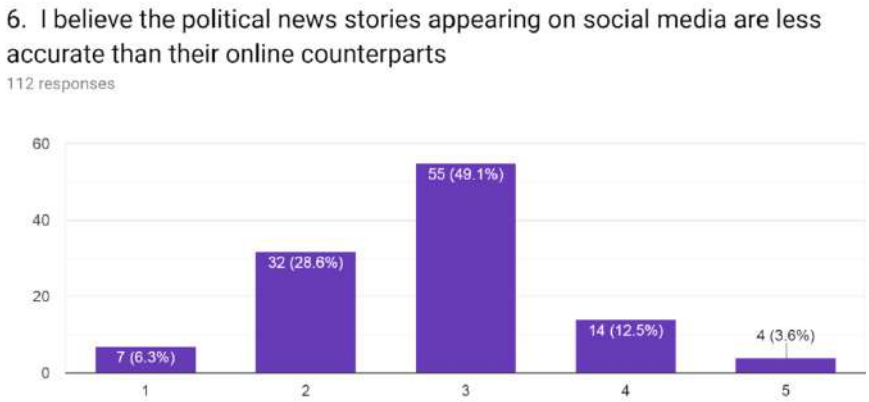




Component: Believability

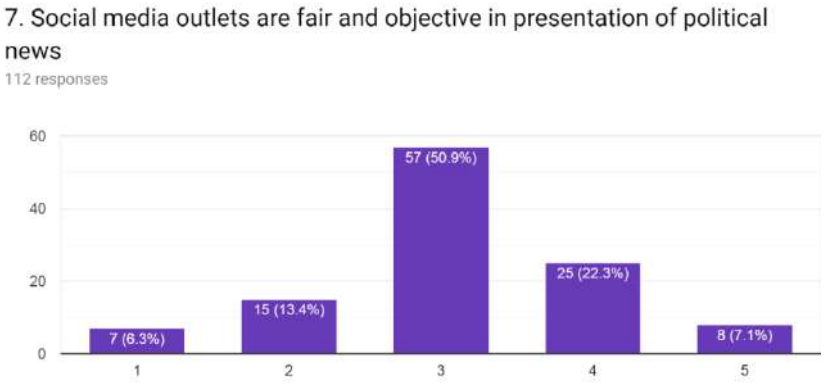


Component: Accuracy

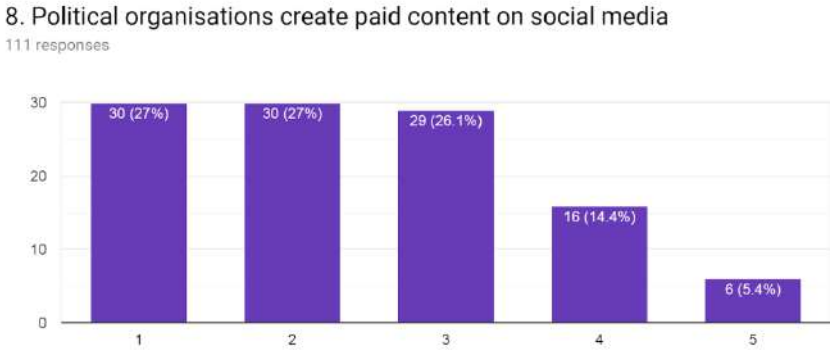


ISSN 2394 - 7780

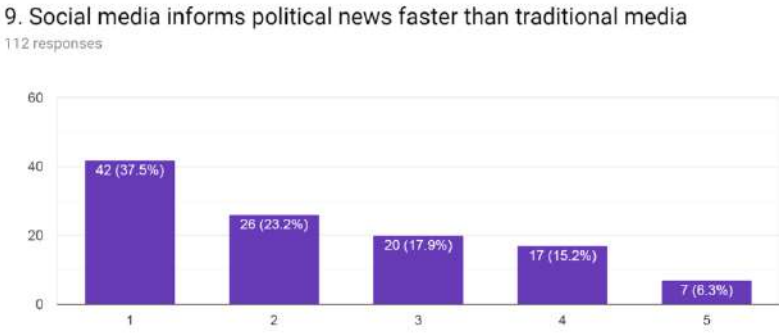
Component: Accuracy

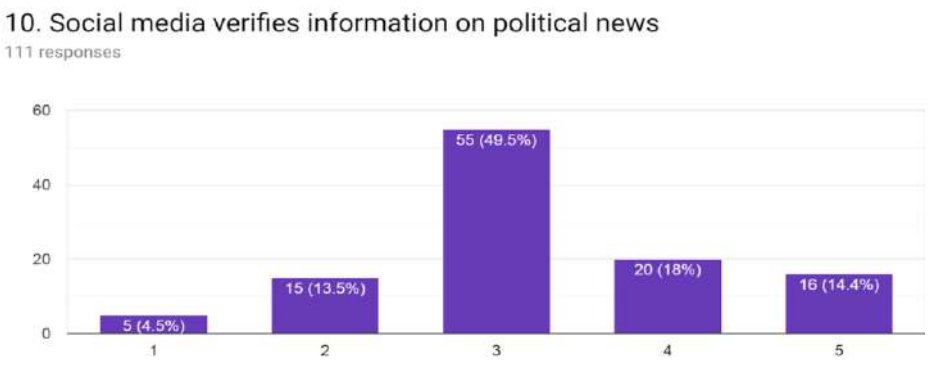


Component: Fairness



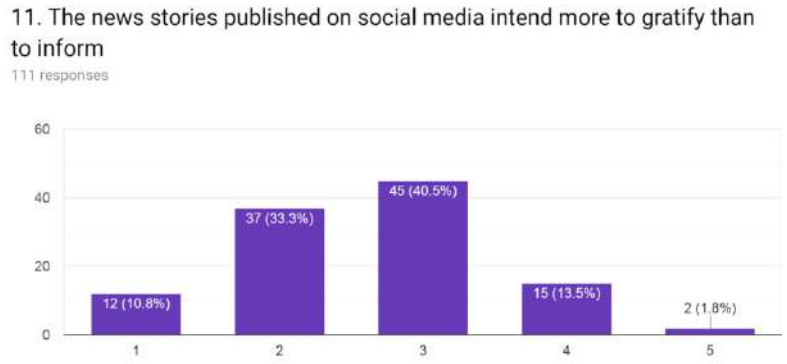
Component: Believability



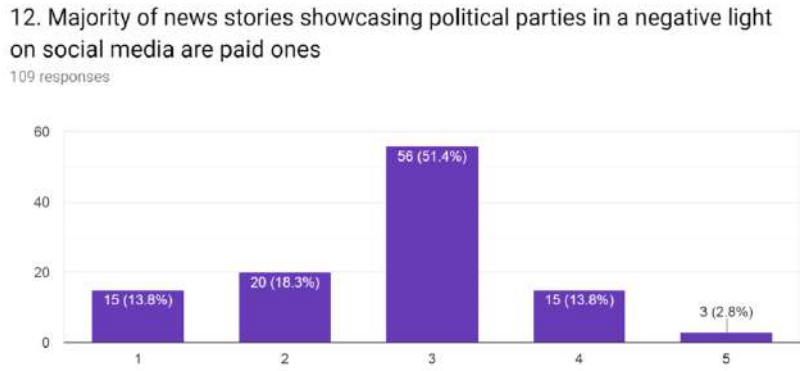


ISSN 2394 - 7780

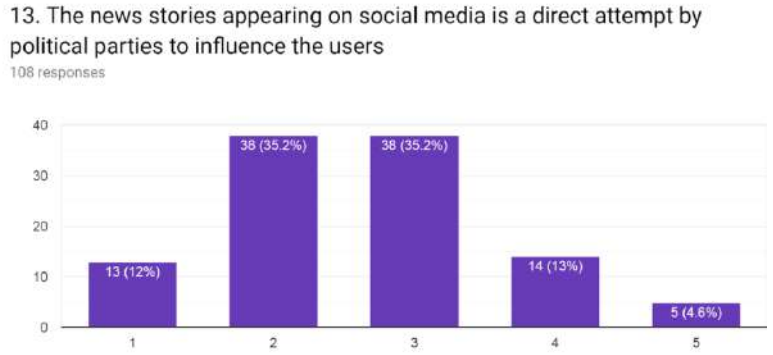
Component: Believability and Accuracy



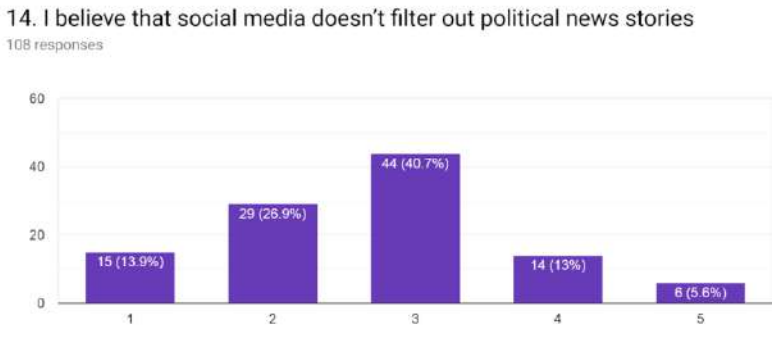
Component: Believability and Accuracy



Component: Believability and Fairness

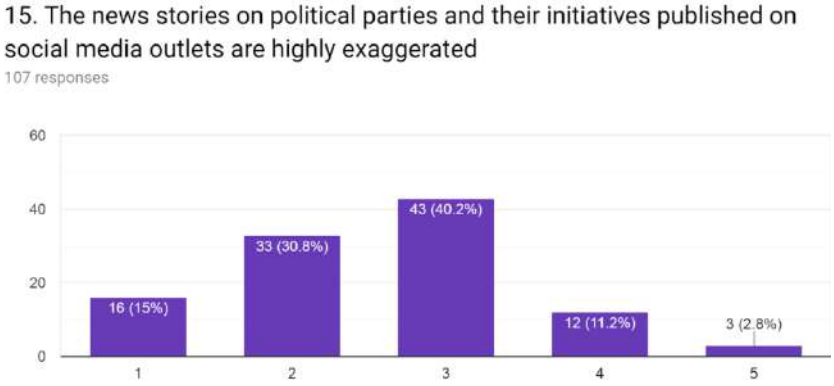


Component: Believability

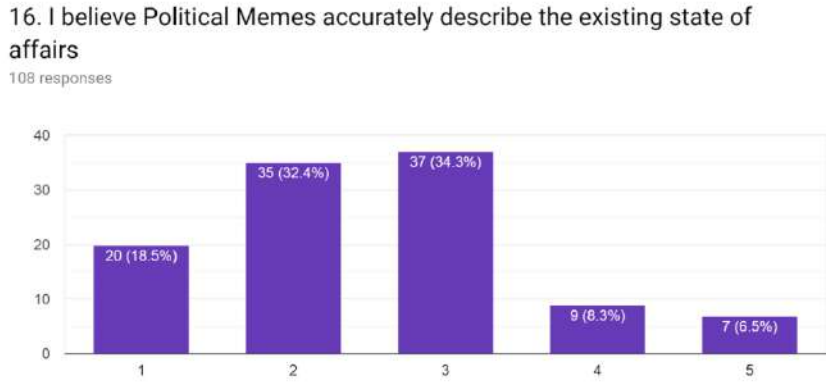


ISSN 2394 - 7780

Component: Believability



Component: Accuracy



Component: Accuracy

DATA ANALYSIS

Of all the news beats consumed on social media, entertainment is most read newsbeat and politics is at the fourth most preceded by Lifestyle and Crime. The data suggest political controversies as the most preferred kind of political news story on social media and pre-election news being the lowest. Political Memes and Headlines together perceived to be as the dominant means used by the Social media to sensationalize political news stories. The perception of the accuracy and objectivity of political news events and news stories carried in social media sites are found to be neutral. This suggests a lack of a deep involvement of the respondents with political news content appearing on the social media.

It's interesting to note that majority agreed that political organisations create paid content on social media platforms and that social media informs political news faster than traditional media. But it was found that they are neutral, rather unsure and ignorant about social media's verification on information on political news. They couldn't take a stand whether the news stories on social media gratifies or informs them. Majority decided not to comment on negative political parties stories as being paid information. The respondents generally agree that the news stories on social media is a direct attempt by political parties to influence users. Filtration of political news stories is a different aspect in itself. The respondents preferred to stay neutral on it. It is possible that the respondents couldn't comment on it as they were ignorant about it. Although respondents considered that political memes are one of the ways in order to sensationalise news yet they preferred to stay neutral regarding political memes describing the current state of affairs.

CONCLUSION

The analysis of data directs us towards a lack of definitiveness displayed by adolescents with respect to the dimension of credibility as a whole associated to social media consumption of political news stories. Therefore, the null hypothesis stands rejected. However, when one looks at the sub-components, perception of *Believability* and *Accuracy* lies in the territory of neutrality tending towards agreement which is higher than the component of

ISSN 2394 - 7780

Fairness. Adolescents perceive the component of *Fairness* to be lesser than *Accuracy* and *Believability* associated with political news articles on social media.

FURTHER SCOPE

The perception study on credibility associated by adolescents can be studied with respect to social media news consumption on various news beats in journalism. Also, the research can be a qualitative exploratory study.

REFERENCES

Anonymous. (2016, April). A New Understanding: What Makes People Trust and Rely on News.

Jarred, P. (2017, Winter). Commanding the Trend: Social Media as Information Warfare. Strategic Studies Quarterly, 11(4), 56.

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003). College student web use, perceptions of information credibility, and verification behaviour. Computers and Education, 271.

Metzger, M. J., Hartsell, E. H., & Flanagin, A. J. (2015). Cognitive Dissonance or Credibility?: A Comparison of Two Theoretical Explanations for Selective Exposure to Partisan News. Communication Research.

Purcell, K., Mitchell, A., Rosenstiel, T., & Olmstead, K. (2010, March 1). Understanding the participatory news consumer.

Shahrinaz, I., & Latif, R. A. (2013). Authenticity Issues of Social Media: Credibility, Quality and Reality . World Academy of Science, Engineering and Technology, 254-261.

Soroush, V., Roy, D., & Aral, S. (2017). The Spread of True and False News Online. MIT Initiative on Digital Economy Research Brief, 1-5.

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022
ISSN 2394 - 7780

STUDYING THE DILEMMAS OF VOICE ASSISTANTS

[1]Dr. Swapnali Lotlikar and [2]Bhakti Chotalia
[1]Assistant Professor and [2]MScIT Student, Information Technology Department, Usha Pravin Gandhi College of Arts, Science and Commerce

ABSTRACT
Technology is rapidly automating all the manual work. Voice Assistant (VA) is one such application that can transcribe human speech, exploits provided verbal information and logical data and gives the user significant information about his inquiry via symphonized speech and has proved to be a good fortune to the technology. VA is used by both native and non – native English Speakers. Users can ask questions, perform basic day – to – day task such as setting reminders, create to – do lists, send emails and texts, manage calendars, control home automation, web searches on the internet and media playback using voice commands. VAs such as Amazon's Alexa, Google's Assistant, Microsoft's Cortana, Apple's Siri are few of the most prevailing, prominent and accustomed used VAs which are ingrained in smartphones, smart watches or dedicated home speakers. VAs have become a principal part of the technology gadget experience and this amplifying competition has led to an across – the – board improvement as well.In this review, I have plunged into the most popular VAs against each other, to find out which one is the BEST! I will also discuss the privacy and security concerns related to the VAs and also future prospect uses of these VAs.

Keywords: Voice Assistants (VA), Google's Assistant, Siri, Cortana, Alexa.

I INTRODUCTION
The Covid – 19 Pandemic has increased reliance on telehealth and additional digital tools for distant healthcare delivery. In a recent article published in the *NJP Digital Medicine,* researchersabstracted the posture of VAs as a dawning tool for remote care delivery and put forward the aptness of health systems and technology saviors to advocate these tools during the current health confrontation and beyond. One of the research authors, *Ujjwal Ramtekkar, MD,* a child and adolescent psychiatrist at Nationwide Children's and medical director of Tele/Virtual Health for Behavioral Health, says "*Voice assistant systems are already a part of our daily lives*." [14]

A new leading – edge technology in every family home is a Voice Assistant, such as a Google Home or an Amazon Echo. These gadgets are usually quartered in the drawing – room as a source for music, jokes, keeping timers for a family game or even asking general knowledge questions, in the cookery as a source for cooking appetizing food recipes or even in master – bedrooms, to control home automation of devices while you kick – start after an exhausting day. [9] In penned definitions, Voice Assistants (VAs) are defined as, *"voice-based interface that relies on voice commands supported by artificial intelligence to have verbal interaction with the end users for performing a variety of tasks."* [7] VAs interminably listens fora key word to awaken. Once it hears the key – word, it distinguishes the user's voice, consigns it to a dedicated server, the server processes and annotates it as a command. Bound by the command, the server releases the VA with particularized intelligence to consummate assignment with profuse services and devices requisitioned by the user.

Google's Assistant was effectuated by Google in 2016, molded to be verbose and reciprocal experience. Siri was invented by Stanford Research Institute Cohorts, which was acquired byApple in 2011 and was assimilated in the iPhone 4S as a spanking new feature. Amazon's Alexa was unveiled in 2014 along as an accessory to Amazon Echo speaker. In January 2015, Microsoft publicized the availability of Cortana for Windows 10 desktops and mobile devices as part of fusingWindows Phone into the operating system principally. [11]

A VA is bygone a character in science. Humans have always desired to talk to computers ensuing its inception. Just a decade ago, having human – like synergetic conversation with computers resembled day – dreaming but here we are, having VAs brainier, humanistic and illuminative than ever. All these VAs have their own base apps or an enthusiastic home screen page or a widget and are responsive to display supplementary information like weather, news and more. It is the quickest – growing and unremittingly oscillating technology over the last few years.

II OBJECTIVE
Voice assistants can execute voluminous activities and every voice assistant has its own inimitable component although they do proportion out some homogeneity and are knowledgeable to carry out fundamental pursuits such as setting reminders, timers, alarms, making phone calls, read and send messages, making to do lists,

ISSN 2394 - 7780

weather forecasting, basic calculations, control media playback from connected devices such as Amazon's Echo Dot, Spotify, Netflix etc., manage home automation devices such as lights, locks etc., respond to basic information queries. Apart from these common features, VAs also cling to proficiency symptomatic to themselves. For example: Amazon's Alexa is qualified of ordering your usual local favorite drink from Starbucks, booking an Uber. Google's Assistant uses web services such as Tasker and IFTTT (If This Then That) that's able to systematize multiple tasks, with the use one single key – word. [3] To illustrate, using the key – word "I am Home" could activate considerable proceedings such as turn off the security camera, turn on the lights and air conditioner, lock the garage doors, turn on the coffee maker, read voicemails, remind the user of his scheduled medical appointments, if any.

Every Voice Solution is not a Voice Assistant, but every Voice Assistant is a Voice Solution. To be categorized as a Voice Assistants, Voice Solution should effectuate these three conditions:

1. **Voice as Input:** Primary Mode of Input for a VA should be Voice.
2. **Conversational:** VAs should be capable to have two – way natural and contextual communication with the user.
3. **Confirmational:** VAs should be able to confirm, clarify and answer with context to the user's request. [15]

Maintaining the illusion of fluid and natural interactions has forged a breach between the user expectation and the real jurisdiction of this VAs which the source is time and time again of partakes of frustration. The primal ambition is to substantiate the real potential of the Voice Assistants, which the founders have called attention to and acclaimed the most. Based on the heterogeneity in the susceptibility of so many trending and ingenious Voice Assistants, it pushes the user into a dilemma while culling the most coherent Voice Assistant, pertaining to user experience, security and privacy, personalization, and to draw a conclusion which one prevails to be the best in real life.

III LITERATURE SURVEY

A. Revamping Voice Assistants

Have you ever asked a Voice Assistant – Google's Assistant, for instance, about her age? Sometimes she replies "*I came into this world in 2016, so I am technically a baby, but I don't throw tantrums, and I am super good at helping others.*" Asking Alexa, the same question, she replies "*I'm 6, I celebrated my Golden Birthday which means I turned 6 on the 6th of November*." whereas Cortana replies "*Well, my birthday is April 2, 2014, so I'm really a spring chicken. Except I'm not a chicken*." Asking Google's Assistant about her gender, she replies "*I am all inclusive. I try to stay neutral*." Asking Alexa, she replies "*As an AI, I don't have a gender*." And Cortana replies "*Well, technically I'm a cloud of infinitesimal data computation.*"

Voice Assistants are increasingly common in current generation Smart Devices and are compatible with many IoT makeshifts that run a managed Operating System. Siri is exclusively available on iOS devices such as the iPhone, iPad, iPod, Apple Watch, Mac, Apple Watches, Home Pod, and Apple TV. Alexa from Amazon works automatically with Amazon's Resounds, Fire Stick, Dash Thing Family, and a plethora of clever contraptions operating on iOS and Android, such as modern speakers, smart televisions, smart watches, headphones, indoor coolers, and so on. Cortana from Microsoft works with Windows 10, Xbox One, Android, Skype, iOS, and Windows Mixed Reality devices. Google Home, Philips Hue White Starter Kit, Google Chromecast, Nest Cam IQ, Sengled Smart LED Starter Kit, and a slew of other devices are compatible with Google Assistant. [13]

The last decade has reshaped people's school of thought on Voice Assistants. From measured users to unfathomable users, VAs have now matured to be a vital component of our daily lives. Today, we are perpetual users of Apple's Siri, Amazon's Alexa, Google's Assistant, any many more. Things recuperated with the launch of Google Home, Apple's Home Pod and Amazon Echo Dot. All these descend on that VAs ratify themselves as driving force for technology not only in household, but also in business quarters too. The use of VAs is parading at a rampant rate across all verticals, ranging from healthcare to banking to production and manufacturing industries.

B. Voice Assistants Trends and Statistics

The growing global sales of smart speakers such as Google Home and Amazon Echo are fueling these VA honours. This also implies that virtual assistants will dominate the web search area in the next years. Current statistics show that 3.25 billion individuals utilise VAs from coast to coast. That equates to around half of the world's population. According to reports, the figure might reach 4 billion in 2020, 5 billion in 2021, and 6.4 billion in 2022. According to Gen Z statistics, 38% of them want to buy via voice-activated ordering. By 2021,

ISSN 2394 - 7780

virtual employee VAs will be used by 16.5% of digital employees. According to Gartner voice search data, VAs will be integrated into commercial and consumer settings, with a $3.5 billion market value by 2021. VAs will be used to optimise testing, medicine, and patient data in industries including as healthcare, remote diagnostics, and elder-care applications. At the moment, data show that the procurement of cutting – edge technological equipment is pre – eminent in the United States, Canada, and Mexico. [16]

According to a research conducted by Uberall, 21% of surveyees were utilising virtual assistants at the end of the week. Almost 10% of queries began as questions – who, what, when, where, why, and how – compared to just 3.7 percent of text-based searches. While driving, 52 percent of those polled used a virtual assistant (VA). Cortana voice search accounts for 25% of desktop searches in Microsoft Windows 10. 52 percent of the VAs were assigned to drawing rooms, 25 percent to bedrooms, and 22 percent to the kitchen. VAs usually sought contrastive activities from surveyees. The market for virtual assistants is anticipated to reach $15.8 billion by 2021. [12]

2.0 Comparative Study of the Voice Assistants

A lot of people reckon Siri, Google Assistant, Alexa, and Cortana, are all just alternatives of the same VAs, setting aside each has its own individuality, imperfection, and stability. Take a glance at the comparison table below. [10]

Parameter	Siri	Google's Assistant	Alexa	Cortana
Parent	Apple Inc	Google	Amazon	Microsoft
Initial Release	14th October, 2011	18th May, 2018	November, 2014	2nd April, 2014
Awaken Key – Word	"Hey Siri"	"Ok Google"	"Alexa"	"Hey Cortana"
Operating System	iOS 5 onwards, macOS sierra onwards, tvOs, watchOS	Android, iOS, KaiOS(jio Phone)	Fire OS 5.0 or later, iOS 8.0 or later, Android 4.4 or later	Windows, iOS, Android, Xbox OS
Compatible With	iOS Devices such as Mac, Apple Watch, Apple TV, Home Pod	Android 4.1 and higher, Google Home, Chromecast, Philips Hue Starter Kit, Nest Cam IQ	Android, iOS, Fire Stick, Dash, Amazon Resounds	Windows 10, XBOX,ONE, Skype, Windows Mixed Reality
Geofencing	Limited	Y	N	Y
Web Search	Y	Y	Y	Y
Search Engine Used	Google Search	Google Search	Bing Search	Bing Search
3rd Party App Support	Y	Y	Y	Y
Play Music	Y	Y	Y	Y
Set Reminders / Alarms	Y	Y	Y	Y
Make Calls	Y	Y	Y	Y
Weather Forecasting	Y	Y	Y	Y
Send and Receive Messages / Emails	Y	Y	Y	Y
Get an Uber	Y	Y	Y	Y
Look Up Directions	Y	Y	Y	Y
Manage Home Automation	Y	Y	Y	Y

Table I. Comparative Study of the Voice Assistants.

Parameter	Siri	Google's Assistant	Alexa	Cortana
Basic Search	5/5	5/5	5/5	5/5
Performing Simple Tasks	5/5	5/5	5/5	5/5
Compatibility	5/5	5/5	5/5	3/5
AI Capability	4/5	4/5	4/5	5/5
Diversity (Accents / Languages)	5/5	5/5	4/5	3/5

ISSN 2394 - 7780

1) Data Analysis :

The most suited survey type is thought to be an online survey. As a result, the target demographic of users and potential users is mostly made up of present Internet users. An online survey proved beneficial in terms of facilitating time-efficient and effective participation. The poll was completed by 40 people, who answered 18 questions on the VAs they use, when they use them, and their privacy concerns about these VAs. The survey is circulated using Google Form by sending the link in various WhatsApp groups. The survey answers were evaluated based on which VA they thought was the best and the device they used to access each VA.

IV USERS

Out of the 40 respondents, our users comprised of 47.5% Male, 37.5% Female and 15% of the users did not wish to disclose their gender. 75% of the users belonged to the age group of 18 –34 years, 15% belonged to the 35 – 54 years age group, 7.5% belonged to the greater than 54 years age group. This proves that Voice Assistants are most famous within the technology expert teenage group.

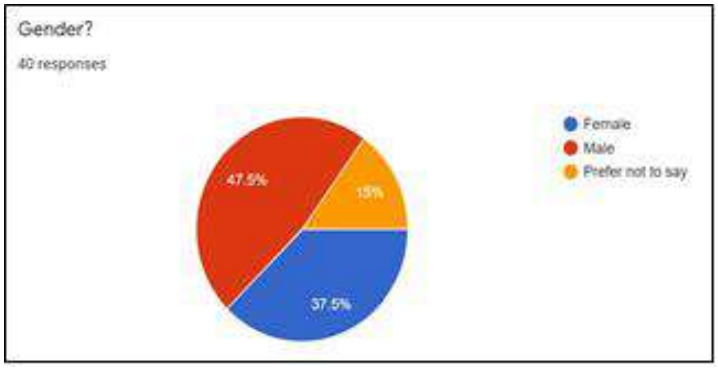


Fig 5.0.1.1 Gender Survey of Users

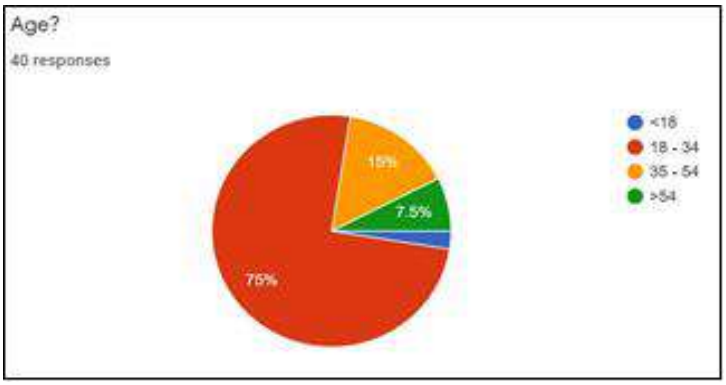
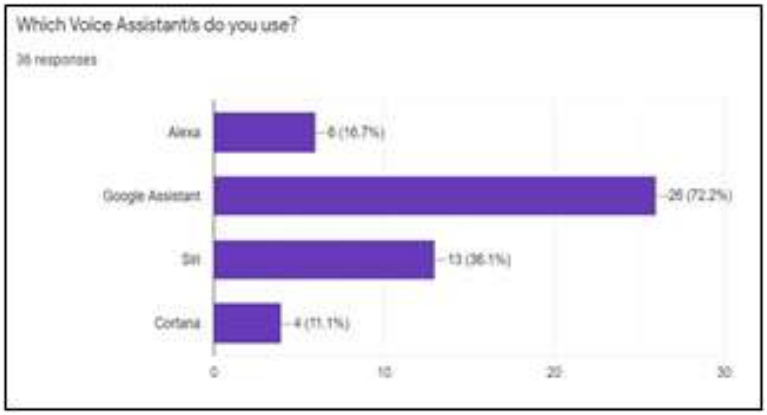


Fig. 5.0.1.2 Age Survey of Users

1) **VA and the Device used by User :**

72.2% of the respondents used Google Assistant, 36.1% users used Siri, 16.1% users used Alexa and 11.1% users used Cortana. Hence as derived from the survey Google Assistant is the most used Voice Assistant, followed by Siri and Alexa. Cortana remains the least favored Voice Assistant. 89.5% respondents used their Personal Mobile Phone to use Voice Assistants, 21.1% used Laptop / Computers and the remaining 18.4% users used Smart Speakers. As derived from the survey, most users use VAs on the Smartphone as it is always handy.



5.0.2.1 Voice Assistants used by users

ISSN 2394 - 7780

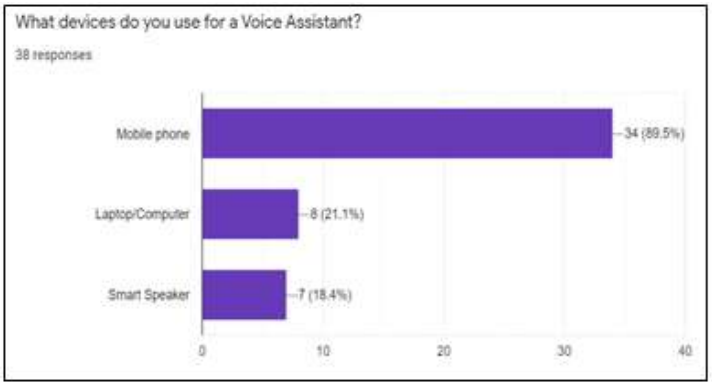


Fig. 5.0.2.2 Devices used by users

2) **Frequency, How and When are VAs used :**

37.85% users used VAs few times in a day, 21.6% users used daily and weekly respectively and 18.9% users used it monthly. Voice Assistants are majorly used by our users to play music (56.4%), to browse web (43.6%) and to provide information about weather, traffic, news or sports and ask general questions (35.9%). Most of the users agreed to use VAs while at home (74.4%), while driving (30.8%) and while at work (17.9%).

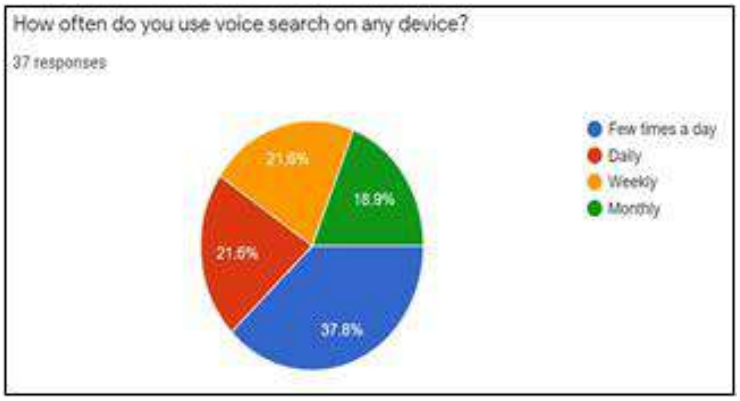


Fig. 5.0.3.1 Frequency Survey on use of VAs

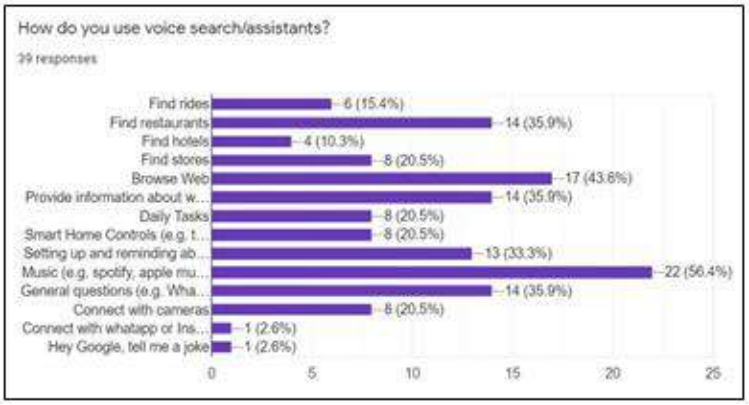


Fig. 5.0.3.2 Survey on most common uses of VAs

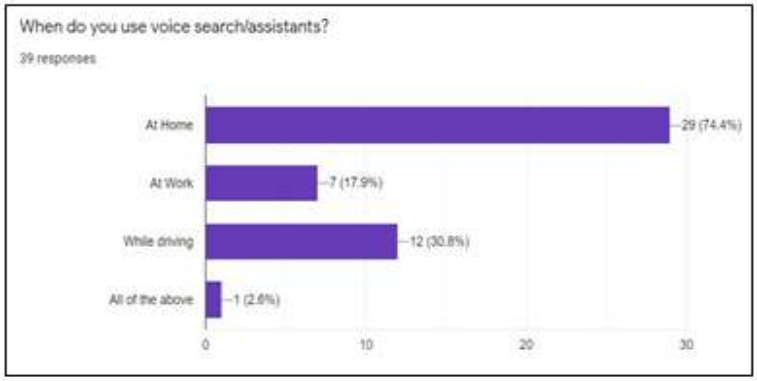


Fig. 5.0.3.3 Survey of when are VAs commonly used

ISSN 2394 - 7780

Usability Study

As per the survey, most of our users seem comfortable using the VAs. The users have agreed that the VAs are easy to use, clear and understandable, and is easy to become skillful using the VAs. Among the four Voice Assistants surveyed, Google Assistant is the most trusted VA leading with 60.5% score, followed by Siri 31.6% and Alexa with 5.6%. The lowest scored VA is Cortana, 2.6%. It becomes apparent that Google Assistant is the most widely used Voice Assistant in Mobile Phone, and on a few times a single day basis. The Voice Assistant used less (in terms of frequency) is Alexa and Cortana.

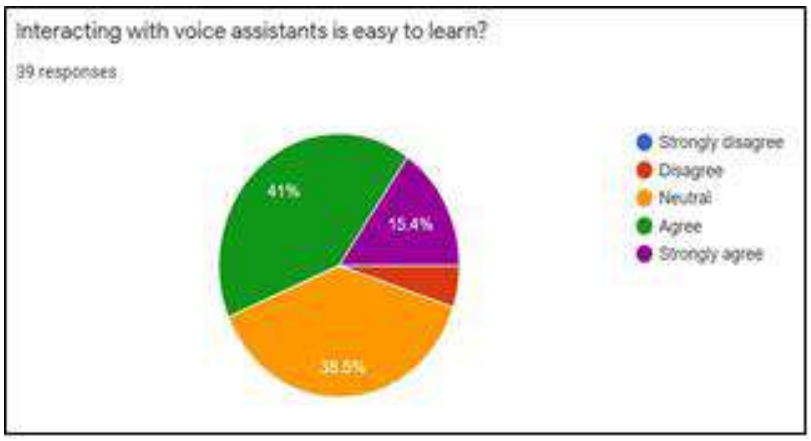


Fig. 5.0.4.1 Survey to know if VAs are easy to learn

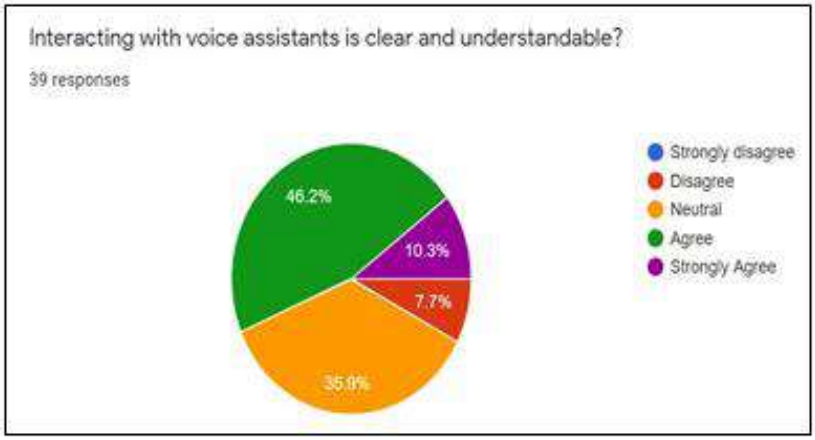


Fig. 5.0.4.2 Survey on most clear and understandable VA

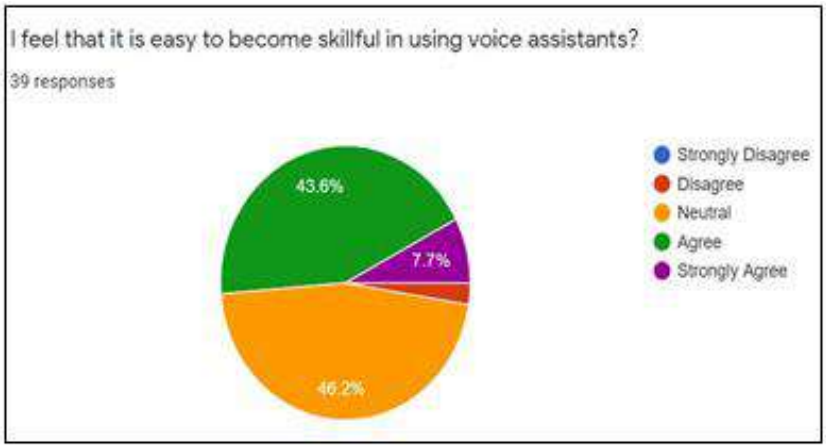


Fig. 5.0.4.3 Survey to know if one can become skillful with use of VA

ISSN 2394 - 7780

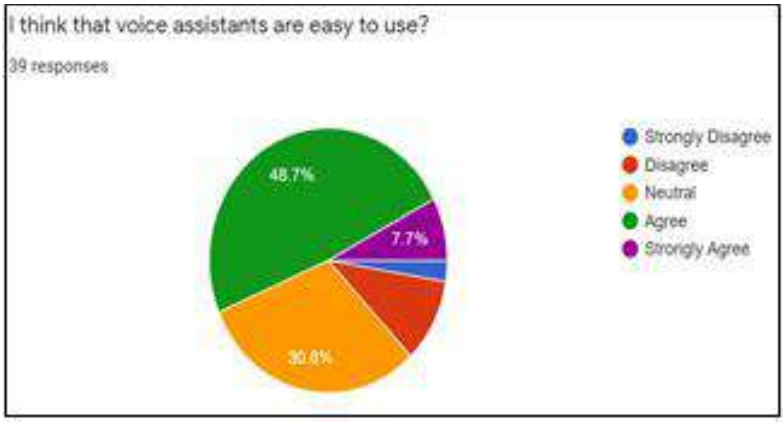


Fig. 5.0.4.4 Survey on most easy–to–use VA

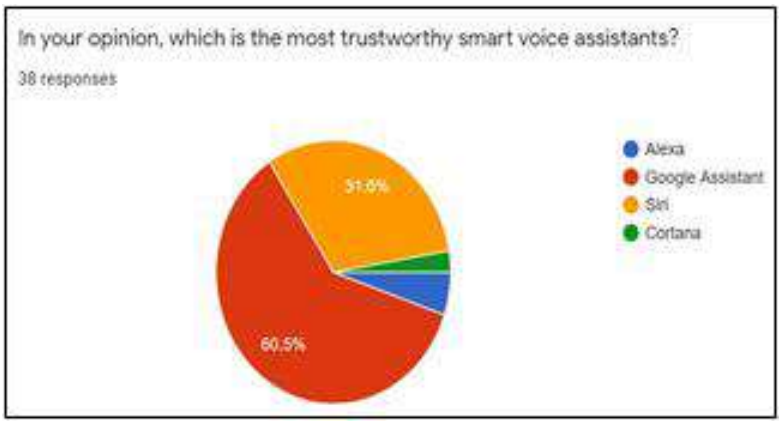


Fig. 5.0.4.5. Survey on most trustworthy VA

5.0.1 **Security and Privacy Concerns with VAs**

According to these findings, the Data Security element is highly regarded, with a medium perceived ease-of-use. 53.8 percent of users were concerned that the VAs were listening in on their private talks in an unethical manner, while 17.9 percent thought that their chats were private in the presence of the VAs. 51.3 percent of users indicated that the information transmitted between the device and the service provider was not encrypted, implying that the user's private information was passed on to theservice provider without their knowledge. 28.2 percent of users trusted Voice Assistants and felt that no communicationoccurred between the device and the service provider without the user's authorization. The majority of users, 69.2 percent, have observed the untimely activation of the VA (Voice Assistant activated without the use of the exclusive activating Key – Word), indicating that the microphone on these devices is always active, with or without the use of the Activating key – word, posing a risk of data breach. Among these, 61.5 percent have raised worries about their data being transmitted to third parties and the Voice Assistants gaining access to this unethically gathered data.

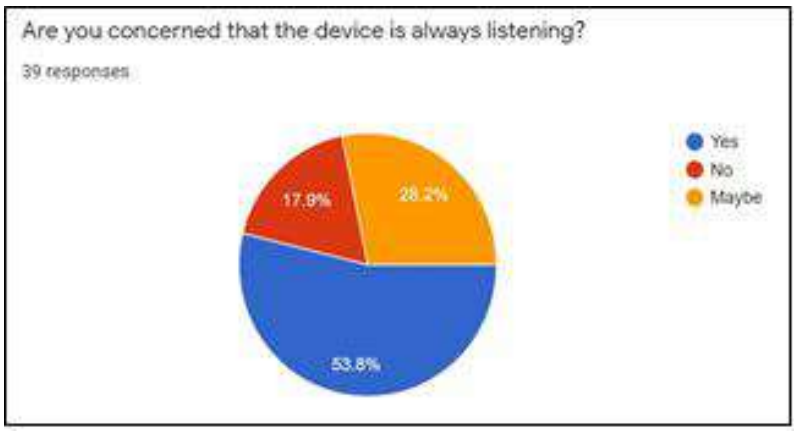


Fig 5.0.5.1 Survey on unauthorized listening by VA

ISSN 2394 - 7780

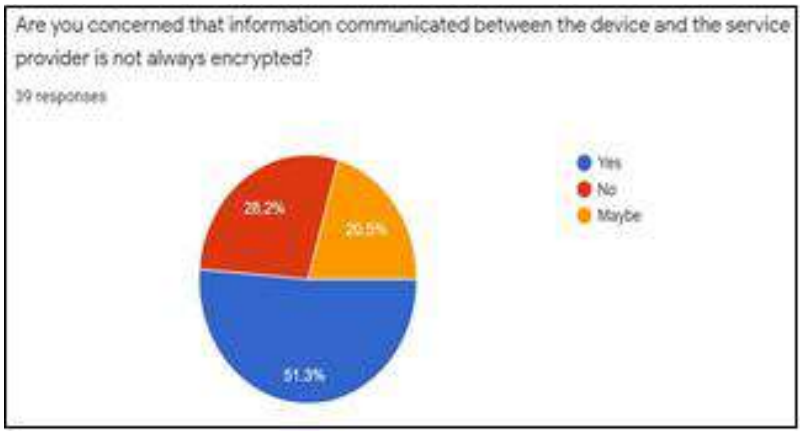


Fig. 5.0.5.2 Survey on Information Communication via VAs

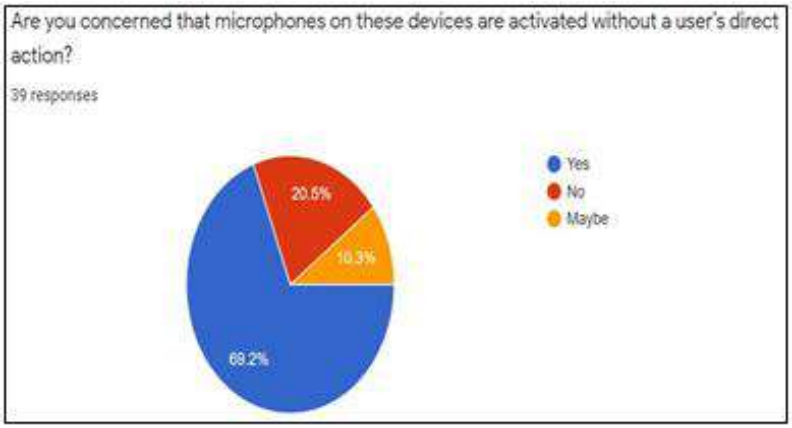


Fig. 5.0.5.3 Survey on Activation of VA

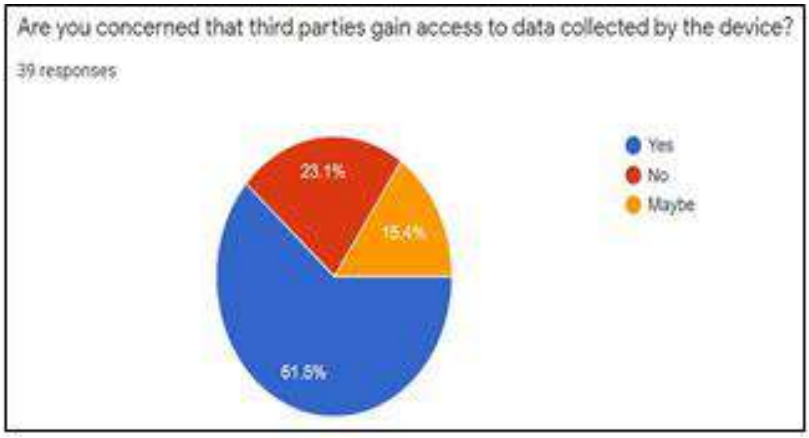


Fig. 5.0.5.4 Survey on Data Collection by VA

5.0.2 **Future Expectations from Voice Assistants**

As per the user expectations, 78.4% says that Voice Assistants need to upskill themselves to converse more like humans and have natural conversational skills. The Voice Assistants need to available in more different languages and support more regional languages as well making them available to use in lesser developed and rural areas where Voice Assistants and technology is still in backward positions. 54.1% users prompted Voice Assistants to have the ability to call and assist with the relevant and required emergency services so as to help in duringunforeseen conditions.

ISSN 2394 - 7780

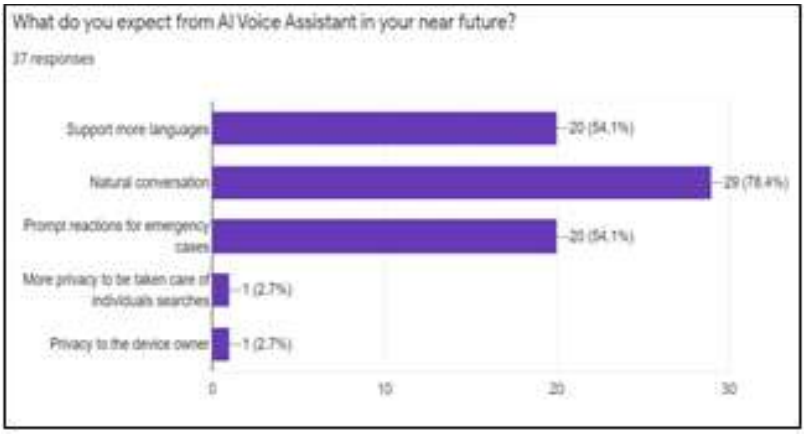


Fig. 5.0.6.1 Survey on Future Expectations from VA

VI CONCLUSION

The primary conclusions of the survey's research with a sample size of 40 respondents revealed the advantages and downsides, privacy issues, and the most trusted Voice Assistant. This research presented the findings of an examination of four resourceful VAs in order to choose the best VA based on how accurate their replies were. The investigation included the most prominent virtual assistants on the market: Siri, Cortana, Alexa, and Google Assistant. The results demonstrate that Google Assistant and Siri are much ahead of Alexa and Cortana. Thereis no statistically significant conflict in claiming that Siri issuperior to Google's Assistant or vice versa. Despite the fact thatour results are uncontaminated, similar replication should be conducted in a variety of settings in order to get additional first person testimony on the usage of VAs. The study undertaken in this publication has certain limitations due to a lack of resources. To begin with, the majority of poll respondents are students between the ages of 18 and 34, which may reduce the sample's representation of the overall population. Furthermore, due to a lack of individuals from various age groups, cross- cultural differences could not be noticed. This is an intriguing starting point for future study on confidence in voice assistants. Options for aged care and their implications for trust might be studied, particularly in light of a caregiver shortage. Finally, data security-focused marketing is an area to investigate further. It would be interesting to expand on this research in the futureby using additional VAs.

REFERENCES

[1] Tulshan, A. S., & Dhage, S. N. (2018, September). Survey on virtual assistant: Google assistant, siri, cortana, alexa. In International symposium on signal processing and intelligent recognition systems (pp. 190-201). Springer, Singapore.

[2] Pal, D., Arpnikanondt, C., Funilkul, S., & Varadarajan, V. (2019, July). User experience with smart voice assistants: the accent perspective. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.

[3] Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: an introduction to voice assistants. Medical reference services quarterly, 37(1), 81-88.

[4] Kaye, J. J., Fischer, J., Hong, J., Bentley, F. R., Munteanu, C., Hiniker, A., ... & Ammari, T. (2018, April). Panel: voice assistants, UX design and research. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-5).

[5] Lahoual, D., & Frejus, M. (2019, May). When Users Assist the Voice Assistants: From Supervision to Failure Resolution. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-8).

[6] Berdasco, A., López, G., Diaz, I., Quesada, L., & Guerrero, L. A. (2019). User experience comparison of intelligent personal assistants: Alexa, Google Assistant, Siri and Cortana. In Multidisciplinary Digital Publishing Institute Proceedings (Vol. 31, No. 1, p. 51).

[7] Branham, S. M., & Mukkath Roy, A. R. (2019, October). Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In The 21st International ACM SIGACCESS Conference on Computers and Accessibility (pp. 446-458).

[8] Choudhari, Y. D., There, A., Bawane, M., Bais, R., Rajgire, R., & Balpande, H. Review On Virtual Assistant With Emotion Recognition.

ISSN 2394 - 7780

[9] Beirl, D., Yuill, N., & Rogers, Y. (2019). Using voice assistant skills infamily life.

[10] https://medium.com/softweb-solutions-inc/face-off-of-the-voice- assistants-siri-vs-cortana-vs-alexa-vs-google-assistant-6ead5cb6bbee

[11] https://medium.com/techtalkers/google-vs-siri-vs-alexa-vs-cortana-which- reigns-supreme-faf7143ccbca

[12] https://backlinko.com/voice-search-stats

[13] https://www.tomsguide.com/best-picks/google-home-compatible-devices

[14] https://pediatricsnationwide.org/2020/11/10/the-role-of-health-care-voice- assistants-during-a-pandemic-and-beyond/

[15] https://www.slanglabs.in/voice-assistants

[16] https://review42.com/resources/voice-search-stats/

International Journal of Advance and Innovative Research
Volume 9, Issue 2 (VI) April – June 2022

ISSN 2394 - 7780

GENERATION OF RANDOM CURVE INSIDE AN ISOTHETIC COVER OF DIGITAL OBJECT

Raina Paul[1], Apurba Sarkar[2], Md. A. A. A. Aman[3] and Arindam Biswas[4]
[1,2,3]Department of Computer Science and Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India
[4]Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah, India

***Abstract*—This paper presents an algorithm to generate a simple random digital curve inside a closed restricted boundary. The curve thus generated is of finite length and non-interesting in nature, and the points on the curve define a closed digital path. In this work we have applied our algorithm inside the inner cover of different digital objects, considering the inner cover as the restricted boundary. The algorithm has been tested on several different objects and the results are very promising.**

***Index Terms*—Cover, IIC, Digital grid, Random Digital Curve**

I. INTRODUCTION

Construction of random curve can be traced back to Brownian motion [1], [2] and random walk [3]–[5]. In the recent literature, a lot of research have been done on generation of closed random curves [6]–[9], [14], [16]. The applications of random curve is in drawing random polygons, which are used as inputs in different algorithms [10]–[13], also in various areas of image processing, computational geometry, and graph algorithms [10]–[12]. Rourke et. al. [8] proposed a heuristic to draw a simple random polygon along with various methods to test the uniformity of random curve generator. Several heuristics were proposed by Auer et. al. [6] to randomly generate simple and star-shaped polygons on a given set of points. The importance of random curve lies in testing the correctness of the algorithms that operate on polygons and to calculate the actual CPU-consumption of such algorithms experimentally. An application of random curves is given by Bhowmick et. al. [20] for artistic emulation of objects. They used randomization factor while sketching a figure from a set of irreducible digital curve segments. Zhu et. al. [9] proposed an $O(n)$ time and space algorithm, after $O(k)$ preprocessing time for generating a random x-monotone polygon on a given set of n vertices, where where $n < k < n^2$ is the number of edges of visibility graph of the x-monotone chain of the given vertex set. They also proposed an $O(n^3)$ time algorithm for generating a random convex polygon whose vertices are a subset of a given set of n points. Tomas et. al. [14] proposes two algorithms, a polynomial time algorithm to generate polygons with increasing number of vertices starting from a unit square. The other follows a constraint approach and find all n-vertex orthogonal polygons with no collinear edges that can be drawn in an $n/2 \times n/2$ grid, for small n. In [16] an algorithm is proposed to generate polygons that are random and representative of the class of all $n-$ *gons* in $O(n^2 \log n)$ time. In [15] an algorithm to generate random digital curves of finite length is reported. It generates points of a digital path on the fly which never intersects or touches itself and hence becomes simple and irreducible. A random curve is called orthogonal (isothetic or 4-connected) when it consists of alternate axis parallel edges whereas in case of an 8-connected random digital curve eight directions (including the diagonal directions) are permissible. Sarkar et. al. [21] proposed a linear time combinatorial algorithm that generates a random simple orthogonal closed curve imposed on the background grid. The combinatorial technique is modified to generate a random simple 8-connected closed digital curve. Also, Sarkar et.al. [17] proposed an algorithm to generate random digital curve in triangular grid. This paper presents an algorithm to generate an orthogonal (4-*connected*) random digital curve inside the inner cover of an digital object. A linear time algorithm to find the inner cover of a digital object is presented by Biswas et. al. [18].

The paper is organized as follows. Required definitions are presented in Sec. II. The procedure to generate the curve and the corresponding algorithm is discussed in Sec. III. Section IV presents the analysis of the algorithm with experimental results in Sec. V and the conclusion is presented in Sec. VI.

II. DEFINITIONS

Few terminologies and definitions are given that are required to explain the proposed algorithm.

ISSN 2394 - 7780

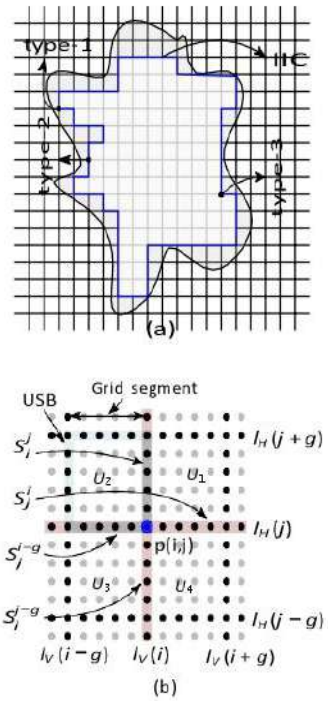


Fig. 1. (a) The digital object(black, *IIC* (blue) with type-1, type-2 and type-3 vertex for *IIC*, (b) Grid points and UGBs.

Definition 1: Digital Object: A digital object is a finite subset of Z^2 consisting of one or more *k*-connected components.

Definition 2: Two points *p* and *q* are said to be k-connected ($k = 4$ or 8) in a set *S* if and only if there exists a sequence $h_p = p_0, p_1, ..., p_n = q_i \in S$ such that $p_i \in N_k(p_{i-1})$ for $1 \leq i \leq n$. The 4-neighborhood of a point (x, y) is given by $N_4(x, y) = (x', y') : |x - x'| + |y - y'| = 1$ and its 8-neighborhood by $N_8(x, y) = (x', y') : max(|x - x'|, |y - y'|) = 1$.

In this paper we have considered the curve for 4-connected objects.

Definition 3: Digital Grid: A digital grid (henceforth referred simply as a grid) $G := (H, V)$ consists of a set H of horizontal (digital) grid lines and a set V of vertical (digital) grid lines, where $H = \{..., l_H(j-2g), l_H(j-g), l_H(j), l_H(j+g), l_H(j+2g), ...\} \subset Z^2$ and $V = \{..., l_V(i-2g), l_V(i-g), l_V(i), l_V(i+g), l_V(i+2g), ...\} \subset Z^2$, for a grid size, $g \in Z^+$. Here, $l_H(j) = \{(i, j) : i \in Z\}$ denotes the horizontal grid line and $l_V(i) = \{(i, j) : j \in Z\}$ denotes the vertical grid line intersecting at the point $(i, j) \in Z^2$, called the grid point, where *i* and *j* are multiples of *g* [18].

Definition 4: Unit Grid Block (UGB): The horizontal digital line $l_H(j)$ divides Z^2 into two half-planes (each closed at one side by $l_H(j)$), which are given by $h_H^+(j) := \{(i, j) \in Z^2 : j \geq j\}$ and $h_H^-(j) := \{(i, j) \in Z^2 : j \leq j\}$. Similarly, $l_V(i)$ divides Z^2 into two half-planes, which are $h_V^+(i) := \{(i, j) \in Z^2 : i \geq i\}$ and $h_V^-(i) := \{(i, j) \in Z^2 : i \leq i\}$. Then the set, given by $(h_V^+(i) \cap h_V^-(i+g)) \cap (h_H^+(j) \cap h_H^-(j+g))$, is defined as the unit grid block, UGB(*i*, *j*). The interior of $U_1 = UGB(i, j)$ is given by $U_1 \setminus (s_i^j \cup s_{i+g}^j \cup s_j^i \cup s_{j+g}^i)$ [18]. For a given grid point, $p(i, j)$, there are four neighboring UGBs, namely, $U_1 = UGB(i, j)$, $U_2 := UGB(i-g, j)$, $U_3 := UGB(i-g, j-g)$, and $U_4 := UGB(i, j-g)$ as shown in Fig. 1(b).

The digital grid and UGBs are shown in Fig. 1(b) for $g = 4$. The four UGBs are shown w.r.t. a grid point $p(i, j)$ where (i, j) are the coordinates of *p*.

Definition 5: Inner Isothetic Cover: The inner (isothetic) cover (IIC) of a digital object *Q*, denoted by $\underline{P}(Q)$, is the maximum-area isothetic polygon inscribed in *Q*.

Classification of Vertices for the Inner Isothetic Cover (IIC). The construction of *IIC* of an object is detailed in the work by Biswas et. al. [18], where they have classified the vertices based the occupancy of the *UGBs*. Fig. 1(a) represents the *IIC* of a given object along with the vertex type.

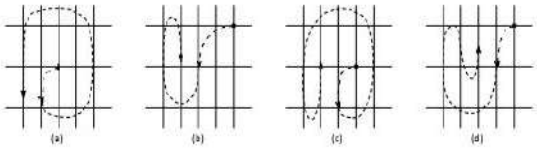


Fig. 2. (a) and (b) Curve in same direction will never reach start point(for consecutive grid points), (c) and (d) Curve in opposite direction will reach the start point(for consecutive grid points).

III. Generating the Curve

The proposed method generates a random curve inside the *IIC* of a digital object. *IIC* is obtained by using the algorithm proposed by Biswas et. al [18]. Once the *IIC* of the object is obtained, the curve starts from a random point inside the object and the curve propagates randomly till it returns to the start point thus giving a random and simple closed curve. The curve is allowed to move to the next point only if it is not visited earlier and satisfies all the conditions which guarantees that the curve will never run into a loop, will never cross the boundary and will be simple.

In this work, we have considered 4-connected curve, thus, there are four permissible directions viz. *top*(1), *left*(2), *bottom*(3), *right*(0) from a point. These directions are stored in an array *A*, which we call as direction vector and initialized to true in the beginning, i.e., all the directions are permissible directions initially. After the inner cover of the digital object is generated, a randomly generated grid point, P_s is chosen as the start point only if it falls within the inner cover. Two indexed lists L_h and L_v stores the coordinate values of vertical line segments and horizontal line segments respectively of the inner cover as shown in Fig. 3 (b). The list L_h stores

ISSN 2394 - 7780

the horizontal line segments of the *IIC* indexed on their y coordinate, similarly, the list L_v stores the vertical segments indexed on the x coordinate of the segment. Every time the next vertex v_n of the curve is generated it is verified to see that if it lies on the boundary of *IIC* by consulting the lists L_h or L_v. If the random direction generated from current point v_c is vertical i.e. 1 or 3 then the entry at x index of the list L_v is checked; if the direction is 0 or 2 i.e. horizontal the the y index of the list L_h is checked. When the point hits the boundary for the first time then the direction taken by the curve (clockwise/anti clockwise) is remembered and is followed every time it hits the boundary to avoid dead ends and backtracking. As shown in Fig. 3 (a), initially, all the four directions were permissible for the point $C(6, 8)$, but when the curve reaches point $C(6, 8)$ from the point $C(6, 7)$, it lies on the boundary and thus permissible directions are calculated. It can be seen that direction 1 cannot be allowed because it is outside the object boundary, and also the direction 3 cannot be allowed since it is the incoming direction, thus in the direction vector of point $C(6, 8)$ directions 1 and 3 are marked invalid (grey). Suppose the curve takes the direction 2. Since, this is the first time curve hits the boundary thus the direction taken by the curve and direction of the *IIC* is remembered. As the curve proceeds, it again hits the boundary at point $C(4, 5)$ from point $C(4, 6)$, the direction taken by the curve first time is anti-clockwise, and the direction of the *IIC* at that instance is anti-clockwise, thus here also curve will follow the same. Hence, for the point $C(4, 5)$ direction 1, 2, and 3 are marked invalid in the direction vector of $C(4, 5)$.

Now, to avoid any loop, and to make sure that the curve does not reaches a dead end, the algorithm is designed in such a way that the two consecutive grid points of C cannot have either same incoming or same outgoing direction. As shown in Fig. 2 (a) and (b) the direction followed by the curve is same and hence it propagates into a region from where it can never reach the start point. Whereas, in Fig. 2 (c) and (d) the direction of the curve is opposite to each other and hence it can reach the start point. After the point v_n is checked for the boundary condition, and it satisfies it, it is allowed to move further only if the direction falls in the permissible direction. Permissible directions are calculated by considering the front left f_l and front right f_r vertex of v_c. Depending on occupancy of those two vertices there may be three cases-

(i) **Neither f_l nor f_r of v_c is visited:** If none of the vertex f_l or f_r of v_c is visited then the curve can move in any direction except in the direction of the visited neighbors. As shown in Fig. 4(a), direction 0 is marked invalid in direction vector D. Thus, the curve can move in any of the valid directions chosen randomly from the direction vector D.

(ii) **Either of f_l or f_r of v_c is visited:** If any one of the vertex f_l or f_r is occupied, then the permissible directions are calculated by traversing the direction vector D starting from either the vertex f_l or f_r whichever is visited along the path followed by f_l or f_r till the vertex v_c, and invalidating all the directions encountered during the traversal. Remaining directions constitutes the permissible directions in the direction vector. For example in (Fig. 4(b), considering the current point v_c, vertex f_l is visited, thus, a traversal is made in D from f_l in the direction of next visited vertex from f_l till the farthest visited vertex upto v_c. During the traversal the directions 0, 2, 3 are marked invalid and the unmarked direction 1 is the only permissible direction from v_c.

(iii) **Both front-left f_l and front-right f_r of v_c is visited:** If both the vertices f_l and f_r of vertex v_c are visited then a traversal is made in the direction vector from the earliest visited vertex between f_l and f_r along the path taken by f_l or f_r till the current vertex v_c, invalidating all the neighbours of v_c encountered in the path. For example in (Fig. 4(c)) assume that the earliest visited vertex is f_l, thus a traversal is made in clockwise direction and the directions 0, 1, 3 in D are marked invalid, thus giving direction 2 as the permissible direction.

A. Algorithm

Algorithm *Random_Curve* takes three lists L, L_h, and L_v as input, where, the list L contains all corner the points of *IIC* of the object, list L_h contains the end points of the horizontal segments of *IIC*, and list L_v contains the end points of vertical line segments of *IIC*. Algorithm outputs another list L_rc as the ordered set of vertices of the generated random curve. In step-1 the list is initialized to ϕ. Function *StartPoint*() generates the start vertex v_s by checking whether the point v_s lies within the *IIC* in step-2. The algorithm starts producing further points from this start point. Step 4 initializes the current vertex v_c with v_s. For each vertex v_c, direction vector D is initialized to true in step 6, and the function *PermissibleDirection*() calculates the permissible directions in step-7. The permissible directions check for the boundary conditions i.e. the curve should be inside the *IIC* and allows only those directions following which the curve will be inside the boundary and it won't get into a trap. The procedure for calculating the permissible direction is explained in details in Sec. III. Among the permissible directions a random direction d is chosen in step-8 and the curve proceeds to the new vertex v_n along the direction d in step-9. The current vertex is then updated to v_n in step-10 and is appended to the list L_rc in step-11. These steps are repeated until the curve meets the start point thereby producing a closed curve.

IV. Time Complexity

This paper proposes a method to generate a random curve inside the *IIC* of a digital object. The pre-processing requires to draw the *IIC* of the object. The time taken to draw the *IIC* of an digital object is $(n g)$, where n is the number of the vertices and g is the grid size. Now, to generate the curve, the permissible directions are calculated along with the boundary check by consulting constant number of vertices. Thus, the running time of the algorithm depends on the length of the

ISSN 2394 - 7780

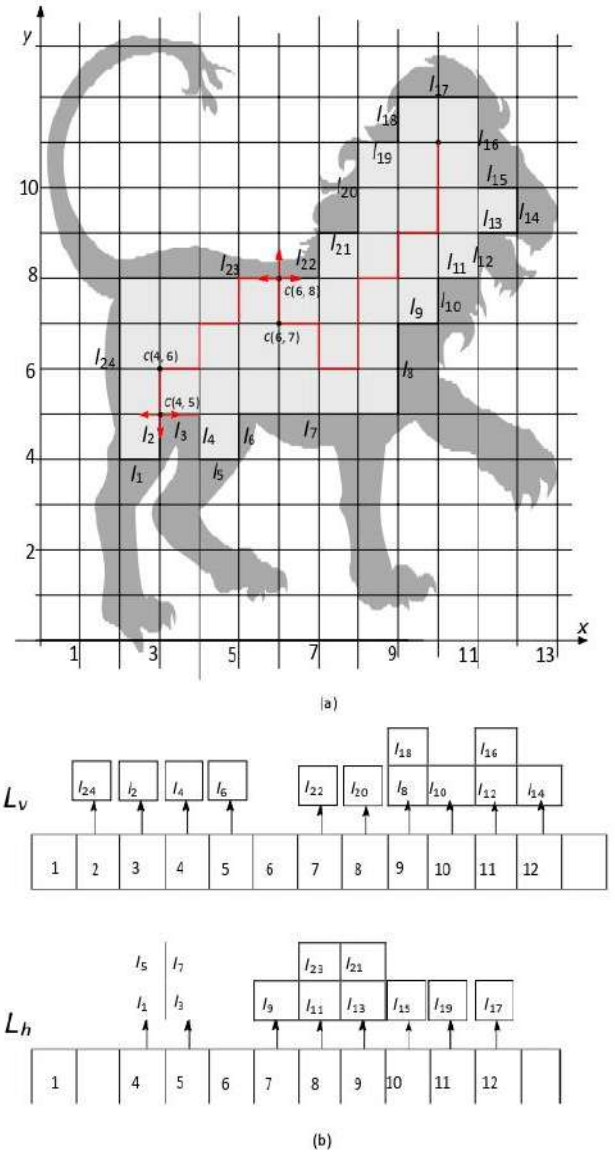


Fig. 3. (a) Digital object with IIC, and the direction vector for point $C(6, 8)$ (b) Lists L_h and L_v storing the lines of the IIC.

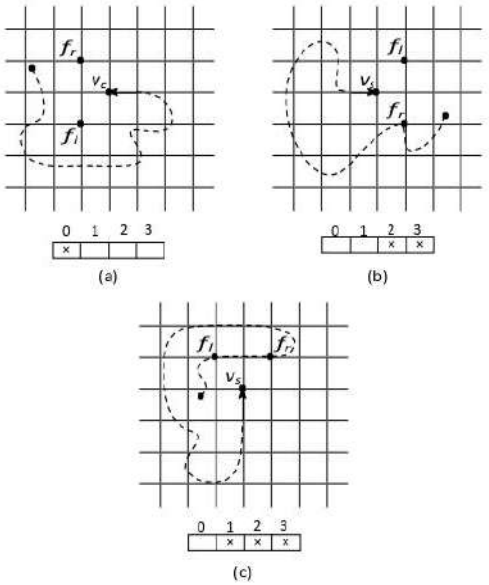


Fig. 4. (a)None of f_l and f_r is visited with respect to v_c, (b) only f_r is visited with respect to v_c, and (c) Both f_l and f_r is visited with respect to v_c.

Algorithm 1: *Generate_Random_Curve*

Input: L, L_h, L_v ← *IIC* of the object
Output: Random Curve ***(C)***

```
1  L_rc = ϕ;
2  v_s = StartPoint();
3  L_rc = v_s;
4  v_c = v_s;
5  do
6      D ← [1, 1, 1, 1];
7      D = PermissibleDirection(L_h, L_v, v_c);
8      d = Random(D);
9      v_n = vertex in direction d;
10     v_c = v_n;
11     L_rc = L_rc ∪ v_c;
12 while v_c ≠ v_s;
```

curve which equals $O(t \setminus g)$, where $|c|$ is the length of the curve. Thus, the total complexity is $O(n \setminus g) + O(t \setminus g)$.

V. Result and Analysis

The proposed algorithm is implemented in the computer system with i5-3230M CPU, 2.60 GHZ × 4 processor and OS Ubuntu 19.0 LTS 64-bit. The algorithm is tested exhaustively on the different images for efficiency and correctness of the algorithm. The Fig. 5 and Fig. 6 shows the random curve generated inside the *IIC* of few objects in different grid size. The first column shows the input objects, the second and third column shows the random curve(red lines) inside the *IIC*(blue lines) of each object. The table 5 shows the generated random curve for grid size 1 and 2 and table 6 shows the generated random curve for grid size 3 and 4. The curves generated depends on the grid size of the images taken. The grids selected for our algorithm is orthogonal grids.

VI. Conclusions and Future Directions

This work proposes a linear time combinatorial algorithm to generate a random simple isothetic curve inside *IIC* of digital object which does not need backtracking and runs in linear time. In this paper, we consider the digital objects as the region for generating random curves, but the algorithm can be used to generate random curves within any restricted isothetic polygon. Applications of random curve lies in experimentation on the study of behaviour of different algorithms that takes curve as input. They also hold importance in artistic emulations of sketches, in different maps used in various strategic games. The proposed algorithm is considered for any digital object that does not contain any hole or simple isothetic polygon. In future, the algorithm may be explored to generate a random curve within a digital object with a hole or any isothetic polygon.

ISSN 2394 - 7780

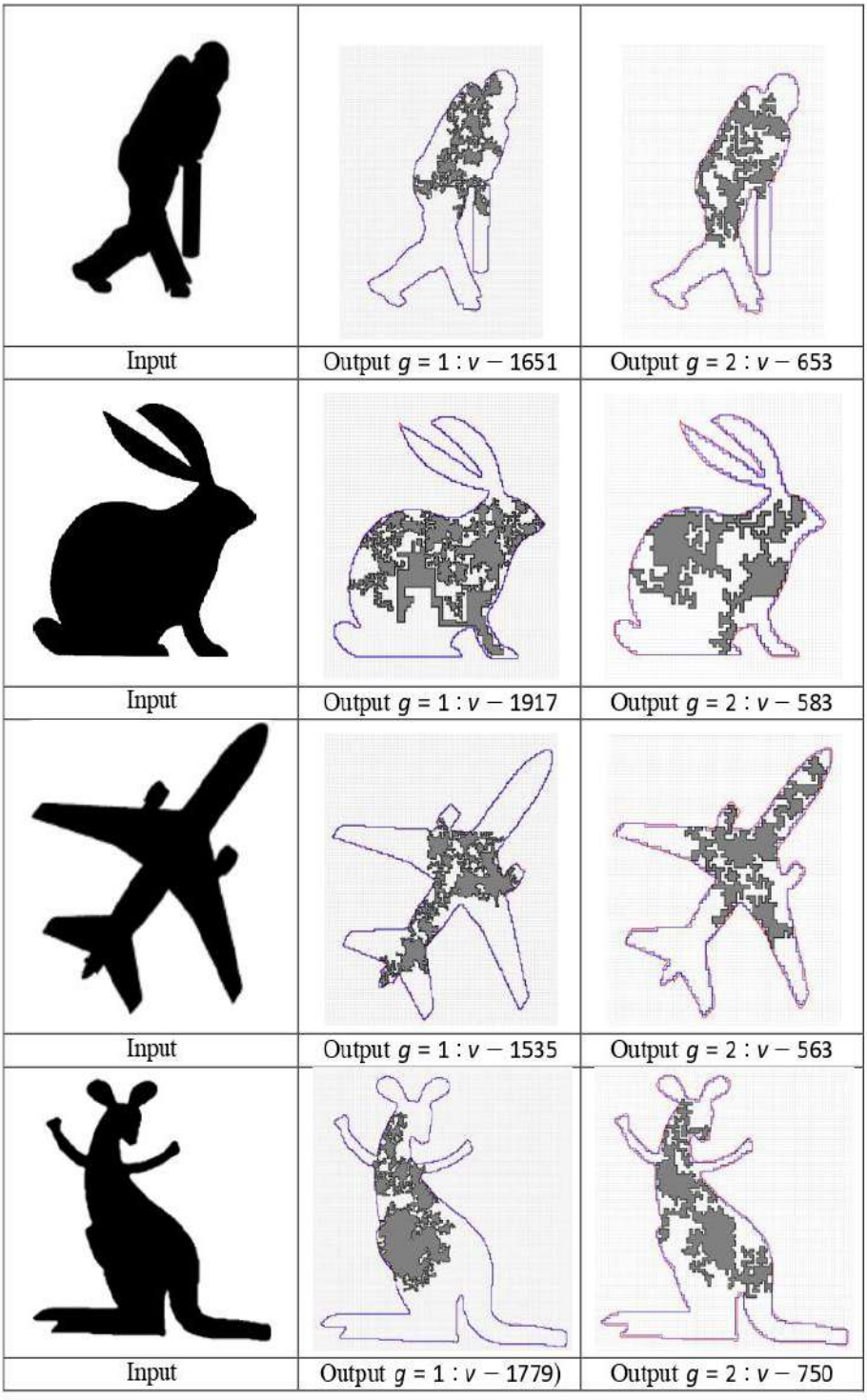


Fig. 5. Random Curve generated for grid size $g = 1$ and $g = 2$

REFERENCES

[1] Albert Einstein, "Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen," Annalen der physik, 1905.

[2] Albert Einstein, "Investigations on the Theory of the Brownian Movement," Courier Corporation, 1956.

[3] Steven R. Finch, "Mathematical constants," Cambridge university press, 2003.

[4] Georg Pólya,"Über eine Aufgabe der Wahrscheinlichkeitsrechnung betreffend die Irrfahrt im Straßennetz,"Mathematische Annalen, Springer, 1921.

[5] Steffen Rohde, and Oded Schramm, "Basic properties of SLE,"Springer, 2011.

[6] Thomas Auer, and Martin Held, "Rpg-heuristics for the generation of random polygons,"Proc. 8th Canada Conf. Comput. Geom. Ottawa, Canada, Citeseer, 1996.

[7] Peter Epstein, "Generating geometric objects at random,"Carleton University, 1992.

[8] Joseph Rourke, and Mandlra Vlrmanl, "Generating Random Polygons," 1991.

[9] Chong Zhu, and Gopalakrishnan Sundaram, and Jack Snoeyink, and Joseph SB Mitchell, "Generating random polygons with given ver-

ISSN 2394 - 7780

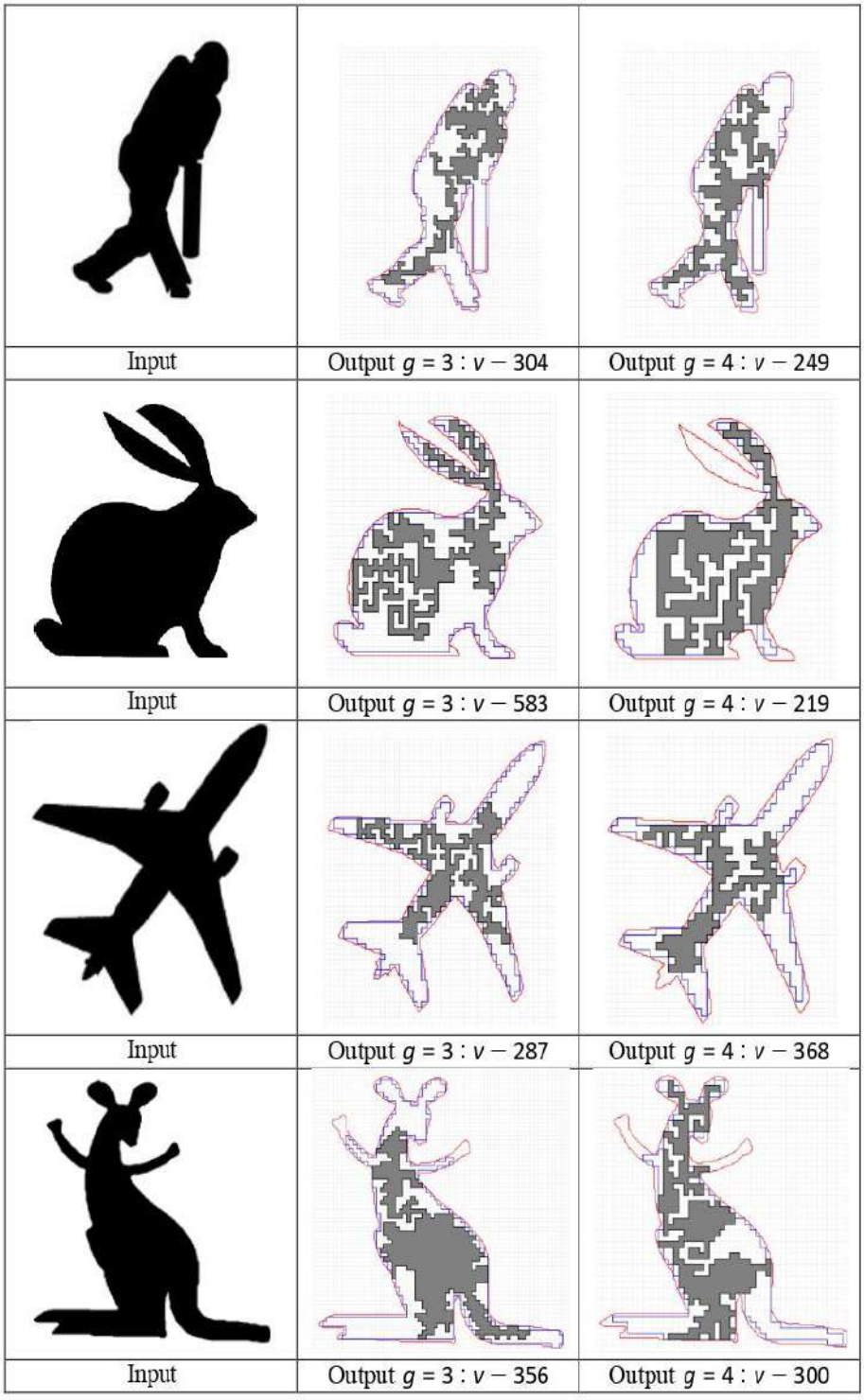


Fig. 6. Random Curve generated for grid size $g = 3$ and $g = 4$

tices,"Computational Geometry, Elsevier, 1996.

[10] Partha Bhowmick, and Bhargab B Bhattacharya, "Fast polygonal approximation of digital curves using relaxed straightness properties,"IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2007.

[11] Isabelle Debled-Rennesson, and Remy Jean-Luc, and Jocelyne Rouyer-Degli, "Segmentation of discrete curves into fuzzy segments." Electronic Notes in Discrete Mathematics 2003.

[12] Reinhard Klette and Azriel Rosenfeld. "Digital geometry: Geometric methods for digital picture analysis." Morgan Kaufmann, 2004.

[13] Smeulders and Leo Dorst." Decomposition of discrete curves into piecewise straight segments in linear time". Phycologia, 1991.

[14] Ana Paula Tomás, and António Leslie Bajuelos, "Generating random orthogonal polygons", In Conference on Technology Transfer, Springer,2003.

[15] Partha Bhowmick, Oishee Pal, and Reinhard Klette. "A linear-time algorithm for the generation of random digital curves", In 2010 Fourth Pacific-Rim Symposium on Image and Video Technology, IEEE, 2010.

[16] David Dailey and Deborah Whitfield. "Constructing Random Polygons," Association for Computing Machinery, 2008.

[17] Apurba Sarkar, Arindam Biswas, Mousumi Dutt, and Arnab Bhattacharya, " Generation of random triangular digital curves using ombinatorial techniques," In International Conference on Pattern Recognition

ISSN 2394 - 7780

and Machine Intelligence, Springer, 2015.

[18] Arindam Biswas, Partha Bhowmick, and Bhargab B. Bhattacharya. "Construction of Isothetic Covers of a Digital Object: A Combinatorial Approach," Journal of Visual Communication and Image Representation, 2010.

[19] Arindam Biswas and Partha Bhowmick and Bhargab B. Bhattacharya, "**TIPS**: On finding a Tight Isothetic Polygonal Shape covering a 2D object,"14th Scandinavian Conference on Image Analysis (SCIA),Springer, Berlin, Heidelberg, 2005.

[20] Partha Bhowmick, and Reinhard Klette, "Generation of Random Digital Simple Curves with Artistic Emulation," Journal of mathematical imaging and vision, 48, 1, 53-71, Springer, 2014.

[21] Apurba Sarkar, and Arindam Biswas, and Mousumi Dutt, and Arnab Bhattacharya, "Generation of random digital curves using combinatorial techniques,"Conference on Algorithms and Discrete Applied Mathematics, Springer, 2015.

MANUSCRIPT SUBMISSION

GUIDELINES FOR CONTRIBUTORS

1. Manuscripts should be submitted preferably through email and the research article / paper should preferably not exceed 8 – 10 pages in all.
2. Book review must contain the name of the author and the book reviewed, the place of publication and publisher, date of publication, number of pages and price.
3. Manuscripts should be typed in 12 font-size, Times New Roman, single spaced with 1" margin on a standard A4 size paper. Manuscripts should be organized in the following order: title, name(s) of author(s) and his/her (their) complete affiliation(s) including zip code(s), Abstract (not exceeding 350 words), Introduction, Main body of paper, Conclusion and References.
4. The title of the paper should be in capital letters, bold, size 16" and centered at the top of the first page. The author(s) and affiliations(s) should be centered, bold, size 14" and single-spaced, beginning from the second line below the title.

First Author Name1, Second Author Name2, Third Author Name3

1Author Designation, Department, Organization, City, email id

2Author Designation, Department, Organization, City, email id

3Author Designation, Department, Organization, City, email id

5. The abstract should summarize the context, content and conclusions of the paper in less than 350 words in 12 points italic Times New Roman. The abstract should have about five key words in alphabetical order separated by comma of 12 points italic Times New Roman.
6. Figures and tables should be centered, separately numbered, self explained. Please note that table titles must be above the table and sources of data should be mentioned below the table. The authors should ensure that tables and figures are referred to from the main text.

EXAMPLES OF REFERENCES

All references must be arranged first alphabetically and then it may be further sorted chronologically also.

- **Single author journal article:**

Fox, S. (1984). Empowerment as a catalyst for change: an example for the food industry. *Supply Chain Management*, *2*(3), 29–33.

Bateson, C. D.,(2006), 'Doing Business after the Fall: The Virtue of Moral Hypocrisy', Journal of Business Ethics, 66: 321 – 335

- **Multiple author journal article:**

Khan, M. R., Islam, A. F. M. M., & Das, D. (1886). A Factor Analytic Study on the Validity of a Union Commitment Scale. *Journal of Applied Psychology*, *12*(1), 129-136.

Liu, W.B, Wongcha A, & Peng, K.C. (2012), "Adopting Super-Efficiency And Tobit Model On Analyzing the Efficiency of Teacher's Colleges In Thailand", International Journal on New Trends In Education and Their Implications, Vol.3.3, 108 – 114.

- **Text Book:**

Simchi-Levi, D., Kaminsky, P., & Simchi-Levi, E. (2007). *Designing and Managing the Supply Chain: Concepts, Strategies and Case Studies* (3rd ed.). New York: McGraw-Hill.

S. Neelamegham," Marketing in India, Cases and Reading, Vikas Publishing House Pvt. Ltd, III Edition, 2000.

- **Edited book having one editor:**

Raine, A. (Ed.). (2006). *Crime and schizophrenia: Causes and cures.* New York: Nova Science.

- **Edited book having more than one editor:**

Greenspan, E. L., & Rosenberg, M. (Eds.). (2009). *Martin's annual criminal code:Student edition 2010.* Aurora, ON: Canada Law Book.

- **Chapter in edited book having one editor:**

Bessley, M., & Wilson, P. (1984). Public policy and small firms in Britain. In Levicki, C. (Ed.), *Small Business Theory and Policy* (pp. 111–126). London: Croom Helm.

- **Chapter in edited book having more than one editor:**

Young, M. E., & Wasserman, E. A. (2005). Theories of learning. In K. Lamberts, & R. L. Goldstone (Eds.), *Handbook of cognition* (pp. 161-182). Thousand Oaks, CA: Sage.

- **Electronic sources should include the URL of the website at which they may be found, as shown:**

Sillick, T. J., & Schutte, N. S. (2006). Emotional intelligence and self-esteem mediate between perceived early parental love and adult happiness. *E-Journal of Applied Psychology, 2*(2), 38-48. Retrieved from http://ojs.lib.swin.edu.au/index.php/ejap

- **Unpublished dissertation/ paper:**

Uddin, K. (2000). A Study of Corporate Governance in a Developing Country: A Case of Bangladesh (Unpublished Dissertation). Lingnan University, Hong Kong.

- **Article in newspaper:**

Yunus, M. (2005, March 23). Micro Credit and Poverty Alleviation in Bangladesh. *The Bangladesh Observer*, p. 9.

- **Article in magazine:**

Holloway, M. (2005, August 6). When extinct isn't. *Scientific American, 293*, 22-23.

- **Website of any institution:**

Central Bank of India (2005). *Income Recognition Norms Definition of NPA.* Retrieved August 10, 2005, from http://www.centralbankofindia.co.in/ home/index1.htm, viewed on

7. The submission implies that the work has not been published earlier elsewhere and is not under consideration to be published anywhere else if selected for publication in the journal of Indian Academicians and Researchers Association.

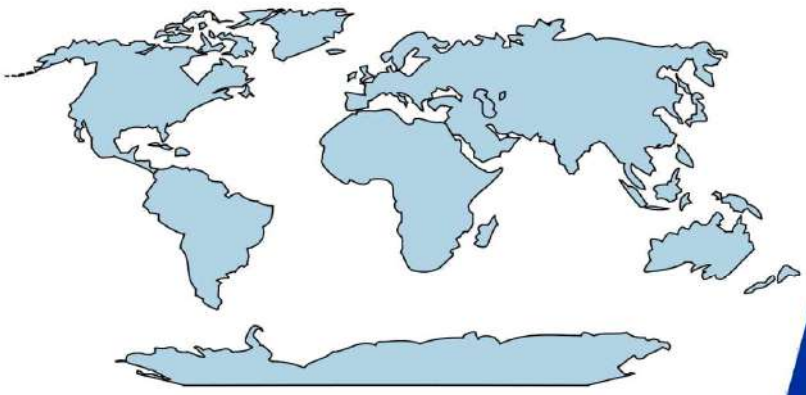
8. Decision of the Editorial Board regarding selection/rejection of the articles will be final.

www.iaraedu.com

Journal

ISSN 2322 - 0899

INTERNATIONAL JOURNAL OF RESEARCH IN MANAGEMENT & SOCIAL SCIENCE



Volume 8, Issue 2
April - June 2020

www.iaraedu.com

Journal

ISSN 2394 - 9554

International Journal of Research in Science and Technology

Volume 6, Issue 2: April - June 2019

Indian Academicians and Researchers Association
www.iaraedu.com

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Membership No: M / M – 1365

Certificate of Membership

This is to certify that

XXXXXXXX

is admitted as a

Fellow Member

of

Indian Academicians and Researchers Association

in recognition of commitment to Educational Research

and the objectives of the Association

Date: 27.01.2020

Director

President

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Membership No: M / M – 1365

Certificate of Membership

This is to certify that

XXXXXXXXXX

is admitted as a

Life Member

of

Indian Academicians and Researchers Association

in recognition of commitment to Educational Research

and the objectives of the Association

Date: 27.01.2020

Director

President

INDIAN ACADEMICIANS AND RESEARCHERS ASSOCIATION

Membership No: M / M – 1365

Certificate of Membership

This is to certify that

XXXXXXXX

is admitted as a

Member

of

Indian Academicians and Researchers Association

in recognition of commitment to Educational Research

and the objectives of the Association

Date: 27.01.2020

Director

President

IARA Organized its 1st International Dissertation & Doctoral Thesis Award in September'2019

1st International
Dissertation & Doctoral Thesis Award
(2019)

Organized By

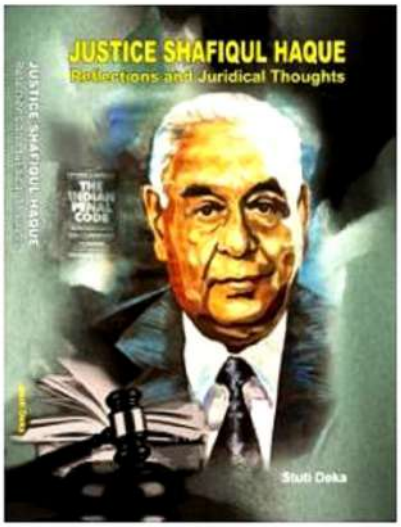
Indian Academicians and Researchers Association (IARA)

EMPYREAL PUBLISHING HOUSE

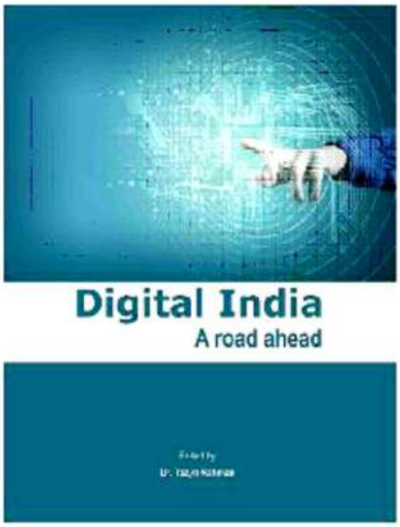
www.editedbook.in

Publish Your Book, Your Thesis into Book or Become an Editor of an Edited Book with ISBN

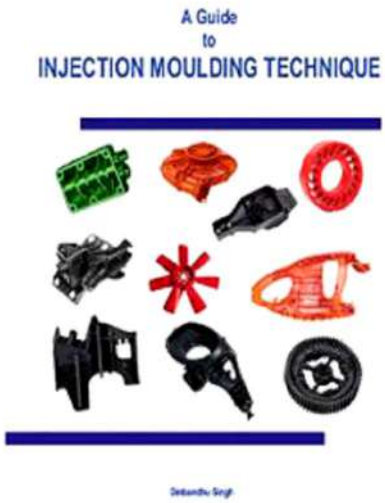
BOOKS PUBLISHED



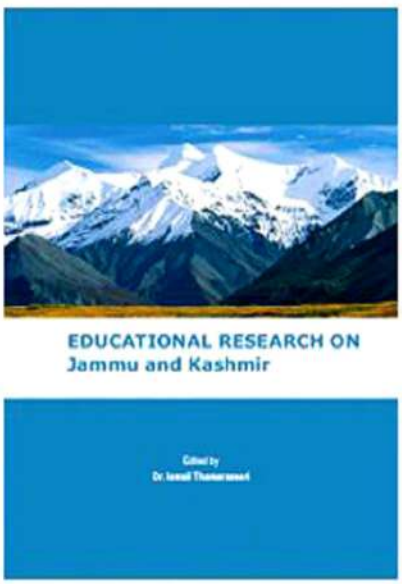
Dr. Stuti Deka
ISBN : 978-81-930928-1-1



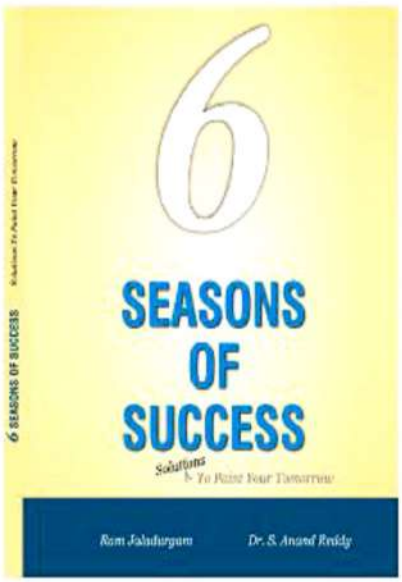
Dr. Tazyn Rahman
ISBN : 978-81-930928-0-4



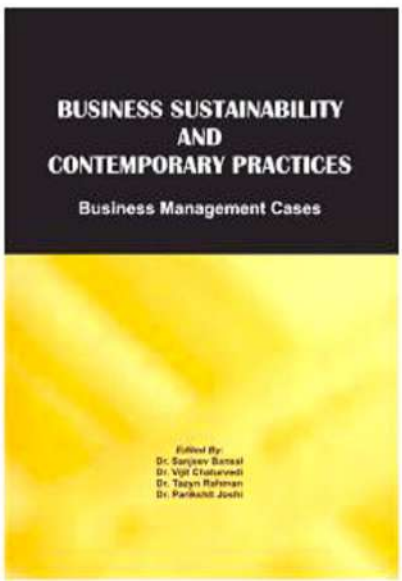
Mr. Dinbandhu Singh
ISBN : 978-81-930928-3-5



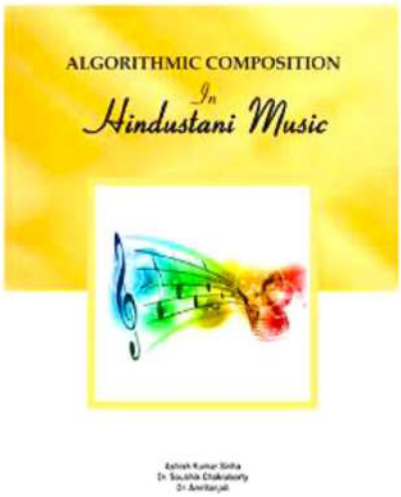
Dr. Ismail Thamarasseri
ISBN : 978-81-930928-2-8



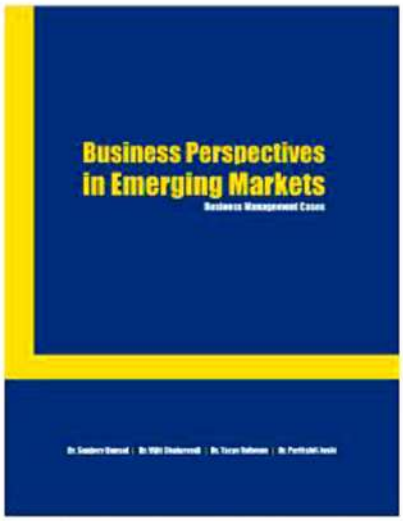
Ram Jaladurgam
Dr. S. Anand Reddy
ISBN : 978-81-930928-5-9



Dr. Sanjeev Bansal, Dr. Vijit Chaturvedi
Dr. Tazyn Rahman, Dr. Parikshit Joshi
ISBN : 978-81-930928-6-6

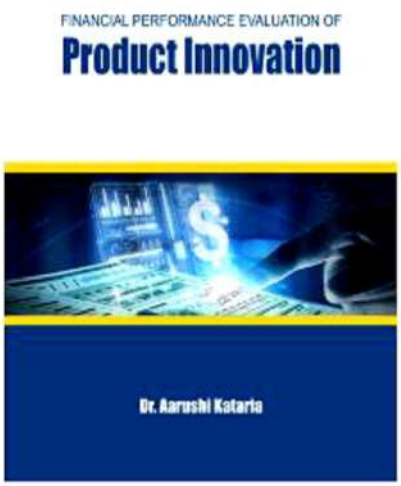


Ashish Kumar Sinha, Dr. Soubhik Chakraborty
Dr. Amritanjali
ISBN : 978-81-930928-8-0

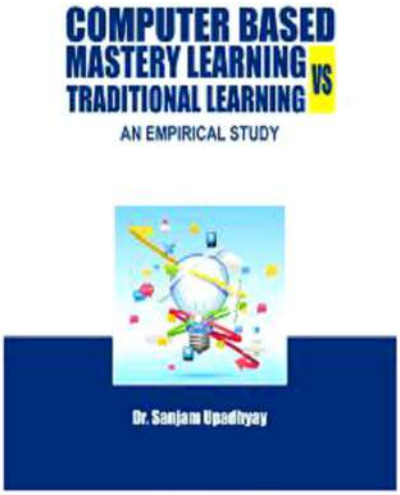


Dr Sanjeev Bansal, Dr. Vijit Chaturvedi
Dr. Tazyn Rahman, Dr. Parikshit Joshi
ISBN : 978-81-936264-0-5

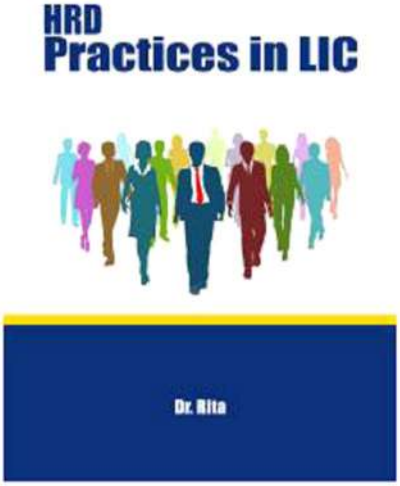
Dr. Jyotsna Golhar
Dr. Sujit Metre
ISBN : 978-81-936264-6-7



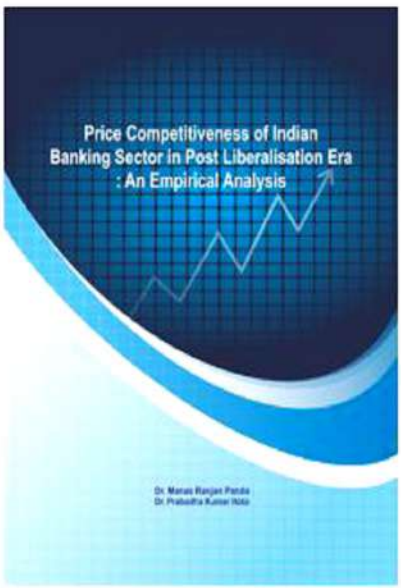
Dr. Aarushi Kataria
ISBN : 978-81-936264-3-6



Dr. Sanjam Upadhyay
ISBN : 978-81-936264-5-0



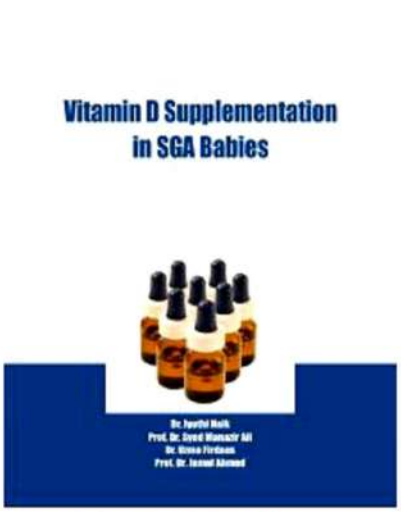
Dr. Rita
ISBN : 978-81-930928-7-3



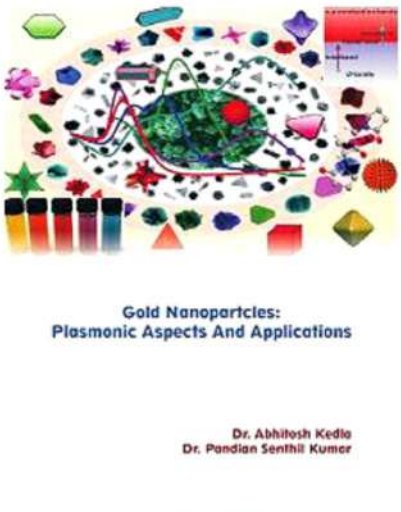
Dr. Manas Ranjan Panda, Dr. Prabodha Kr. Hota
ISBN : 978-81-930928-4-2

Poornima University
ISBN : 978-8193-6264-74

Institute of Public Enterprise
ISBN : 978-8193-6264-4-3



Dr. Jyothi Naik, Prof. Dr. Syed Manazir Ali
Dr. Uzma Firdaus, Prof. Dr. Jamal Ahmed
ISBN : 978-81-936264-9-8



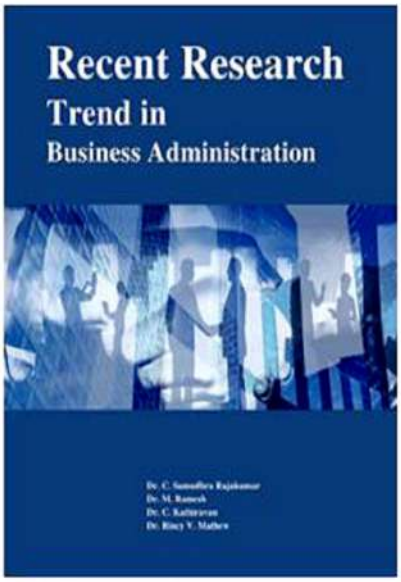
Dr. Abhitosh Kedia
Dr. Pandian Senthil Kumar
ISBN : 978-81-939070-0-9

Dr. Vinod
S. Chandwani
ISBN : 978-81-939070-2-3

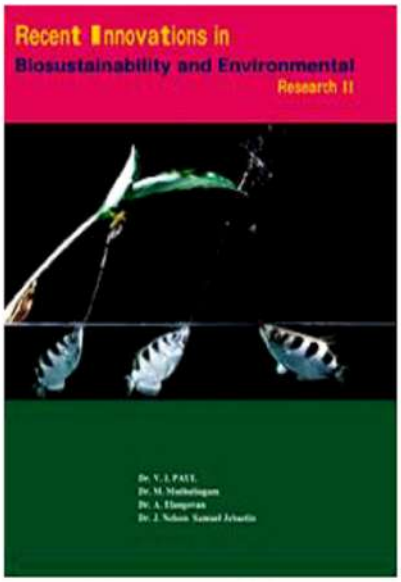
Prof. Dr. Dhananjay
Awasarikar
ISBN : 978-81-939070-1-6

Dr. C. Samudhra Rajakumar, Dr. M. Ramesh
Dr. C. Kathiravan, Dr. Rincy V. Mathew
ISBN : 978-81-939070-4-7

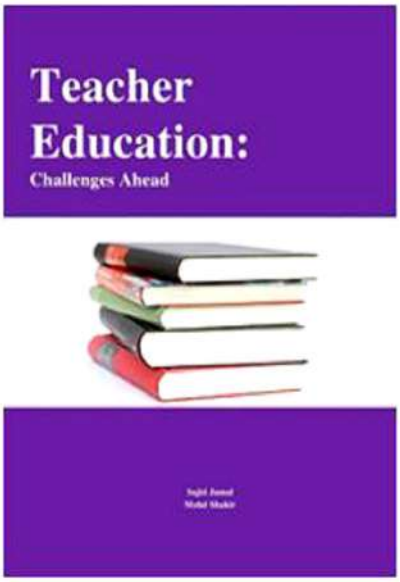
Dr. C. Samudhra Rajakumar, Dr. M. Ramesh
Dr. C. Kathiravan, Dr. Rincy V. Mathew
ISBN : 978-81-939070-6-1



Dr. C. Samudhra Rajakumar, Dr. M. Ramesh
Dr. C. Kathiravan, Dr. Rincy V. Mathew
ISBN : 978-81-939070-7-8

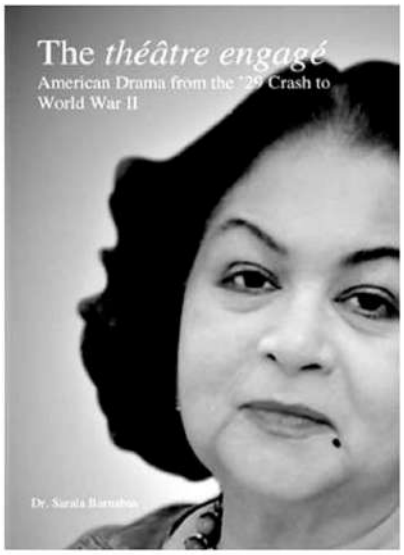


Dr. V. I. Paul, Dr. M. Muthulingam
Dr. A. Elangovan, Dr. J. Nelson Samuel Jebastin
ISBN : 978-81-939070-9-2



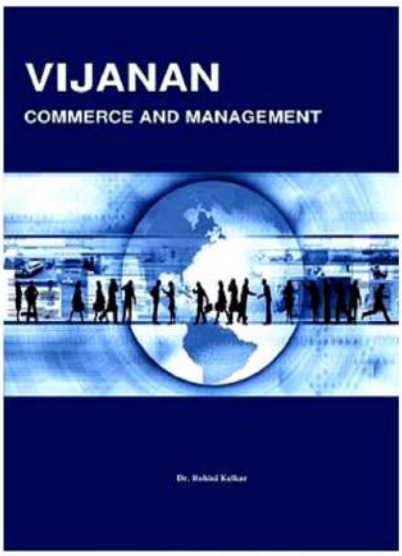
Sajid Jamal
Mohd Shakir
ISBN : 978-81-939070-8-5

Dr. R. Emmaniel
ISBN : 978-81-939070-3-0



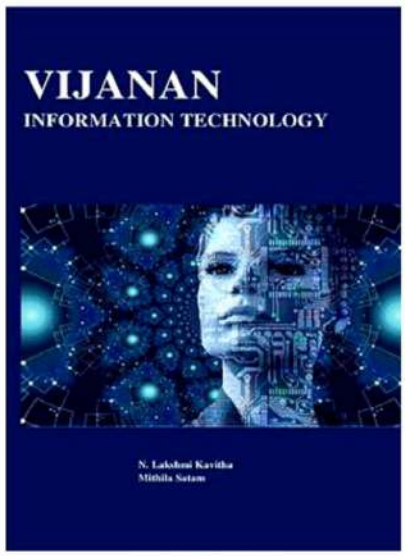
Dr. Sarala Barnabas
ISBN : 978-81-941253-3-4

Dr. M. Banumathi
Dr. C. Samudhra Rajakumar
ISBN : 978-81-939070-5-4

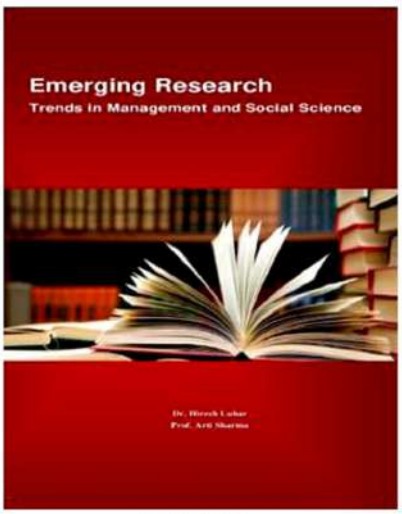


Dr. (Mrs.) Rohini Kelkar
ISBN : 978-81-941253-0-3

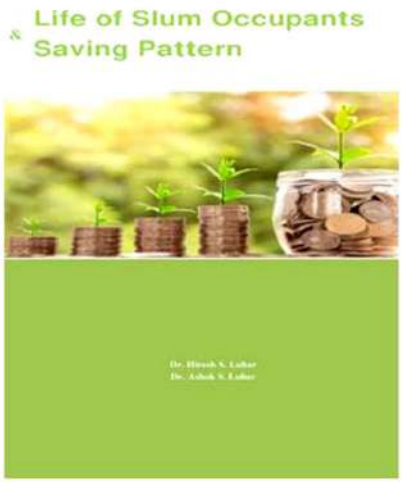
Dr. Tazyn Rahman
ISBN : 978-81-941253-2-7



Dr. N. Lakshmi Kavitha
Mithila Satam
ISBN : 978-81-941253-1-0

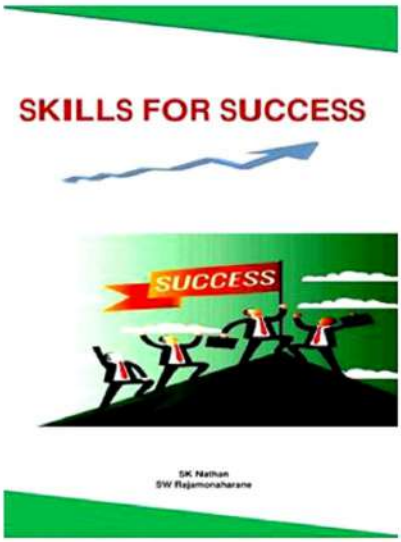


Dr. Hiresh Luhar
Prof. Arti Sharma
ISBN : 978-81-941253-4-1



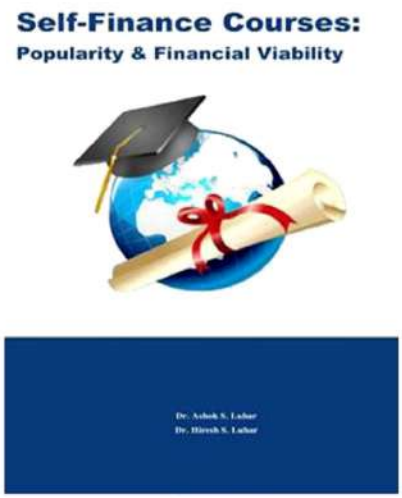
Dr. Hiresh S. Luhar
Dr. Ashok S. Luhar
ISBN : 978-81-941253-5-8

Dr. Babita Kanojia
Dr. Arvind S. Luhar
ISBN : 978-81-941253-7-2

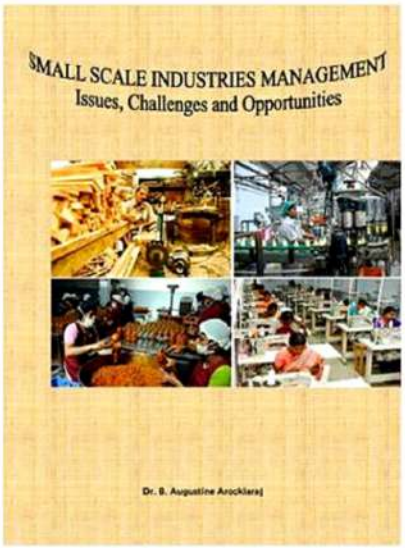


Dr. Sw Rajamonaharane
SK Nathan
ISBN : 978-81-942475-0-0

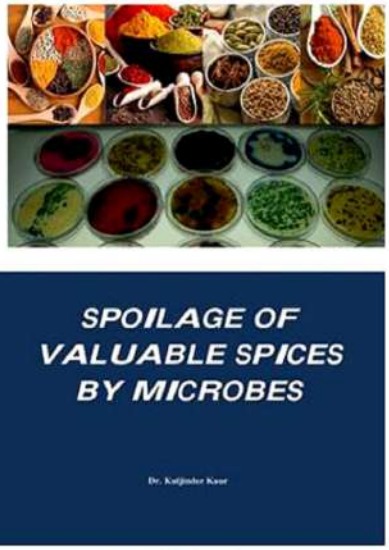
Aditi Sharma
ISBN : 978-81-941253-8-9



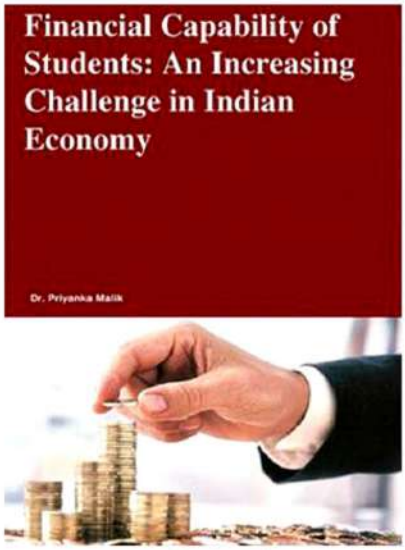
Dr. Ashok S. Luhar
Dr. Hiresh S. Luhar
ISBN : 978-81-941253-6-5



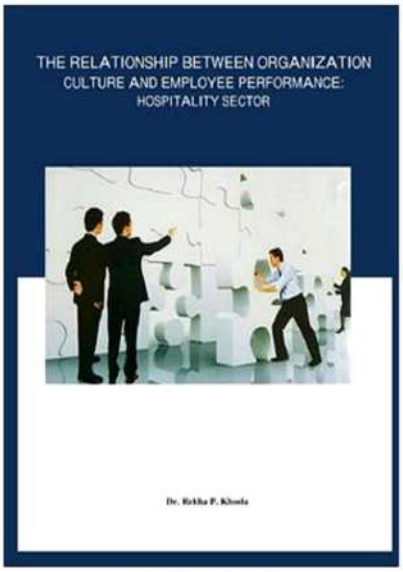
Dr. B. Augustine Arockiaraj
ISBN : 978-81-941253-9-6



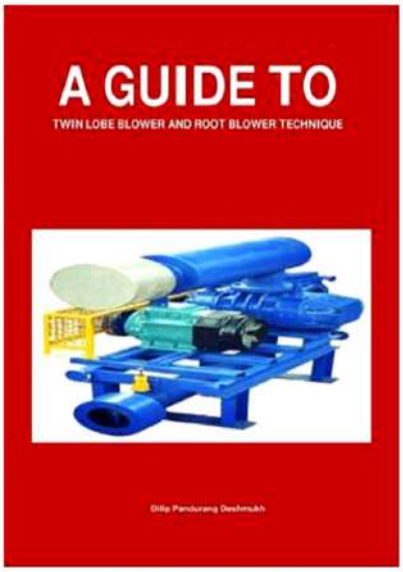
Dr. Kuljinder Kaur
ISBN : 978-81-942475-4-8



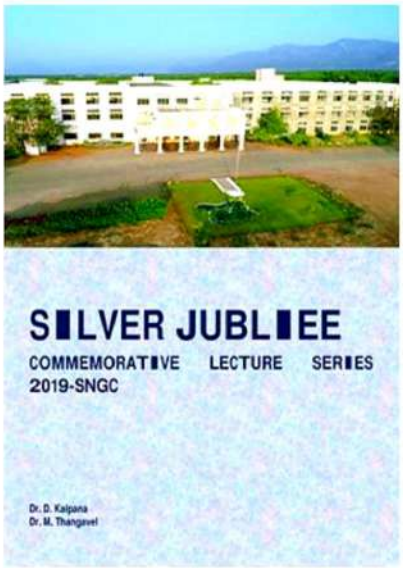
Dr. Priyanka Malik
ISBN : 978-81-942475-1-7



Dr. Rekha P. Khosla
ISBN : 978-81-942475-2-4

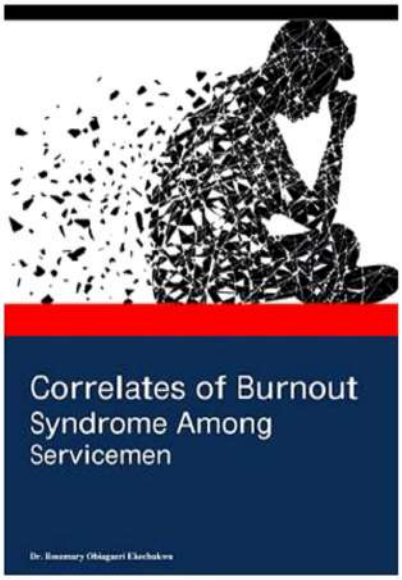


Dilip Pandurang Deshmukh
ISBN : 978-81-942475-3-1

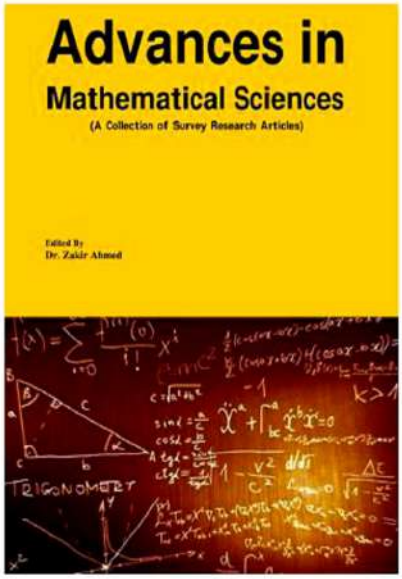


Dr. D. Kalpana, Dr. M. Thangavel
ISBN : 978-81-942475-5-5

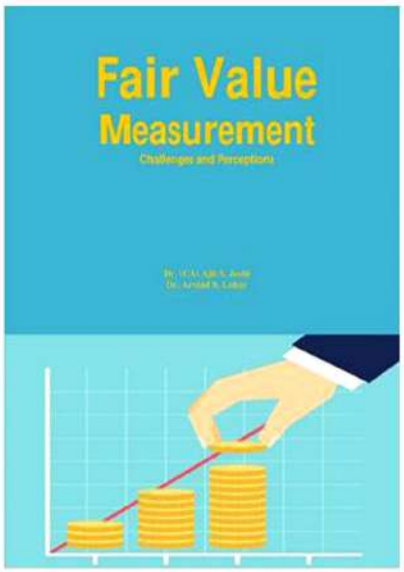
Dr. Aloysius Edward J.
ISBN : 978-81-942475-7-9



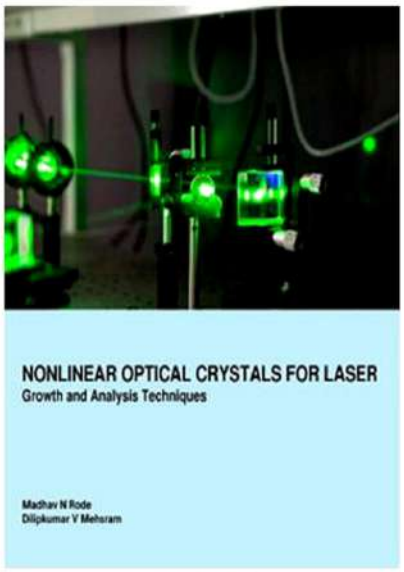
Dr. R. O. Ekechukwu
ISBN : 978-81-942475-8-6



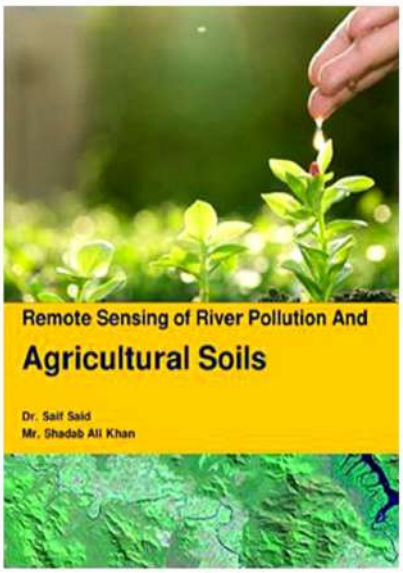
Dr. Zakir Ahmed
ISBN : 978-81-942475-9-3



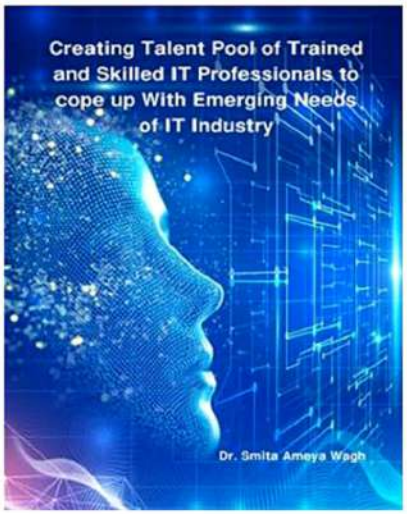
Dr. (CA) Ajit S. Joshi
Dr. Arvind S. Luhar
ISBN : 978-81-942475-6-2



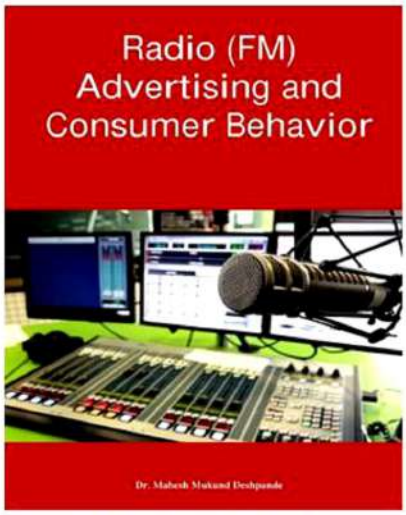
Madhav N Rode
Dilip Kumar V Mehsram
ISBN : 978-81-943209-6-8



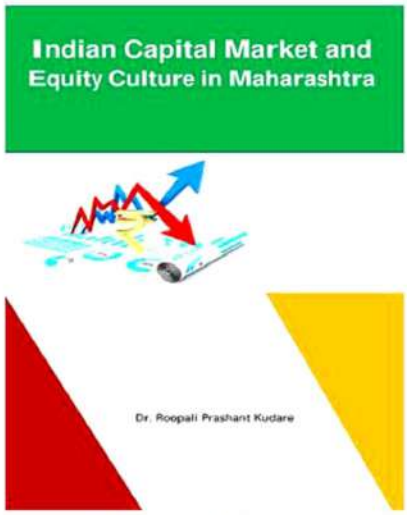
Dr. Saif Said
Shadab Ali Khan
ISBN : 978-81-943209-1-3



Dr. Smita Ameya Wagh
ISBN : 978-81-943209-9-9

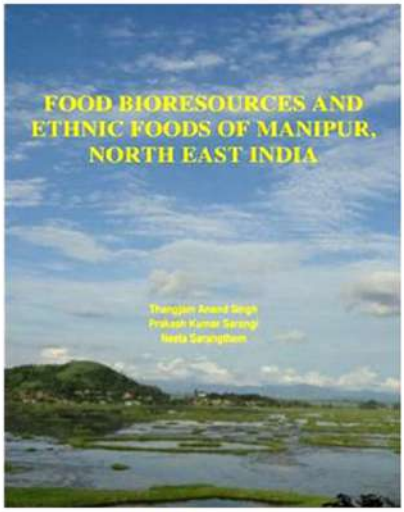


Dr. Mahesh Mukund Deshpande
ISBN : 978-81-943209-7-5

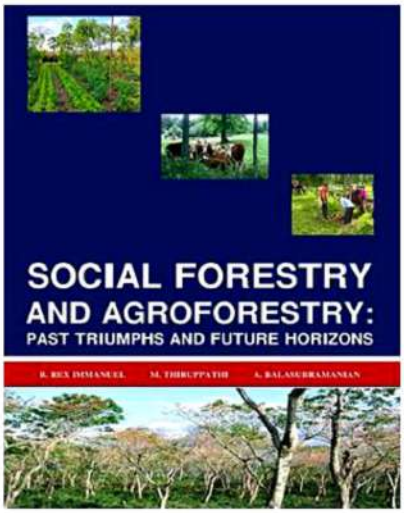


Dr. Roopali Prashant Kudare
ISBN : 978-81-943209-3-7

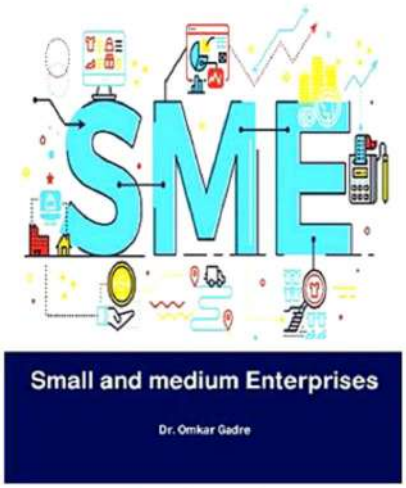
M. Thiruppathi
R. Rex Immanuel
K. Arivukkarasu
ISBN : 978-81-930928-9-7



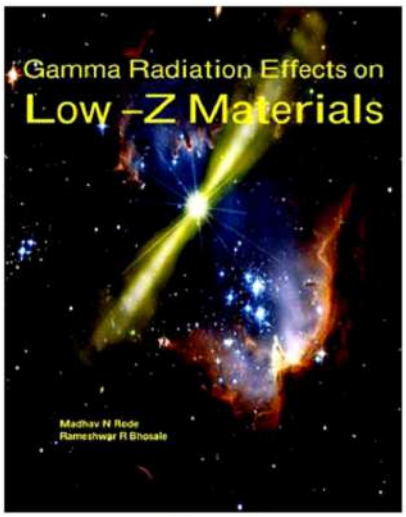
Dr. Th. Anand Singh
Dr. Prakash K. Sarangi
Dr. Neeta Sarangthem
ISBN : 978-81-944069-0-7



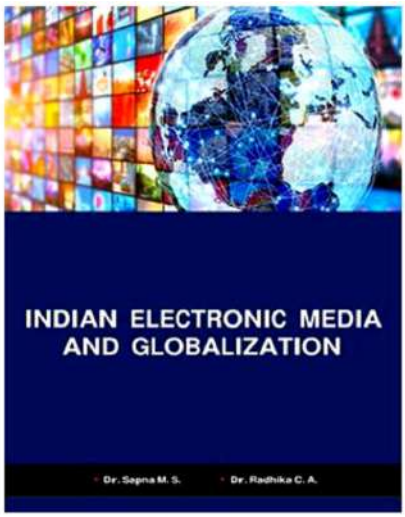
R. Rex Immanuel
M. Thiruppathi
A. Balasubramanian
ISBN : 978-81-943209-4-4



Dr. Omkar V. Gadre
ISBN : 978-81-943209-8-2

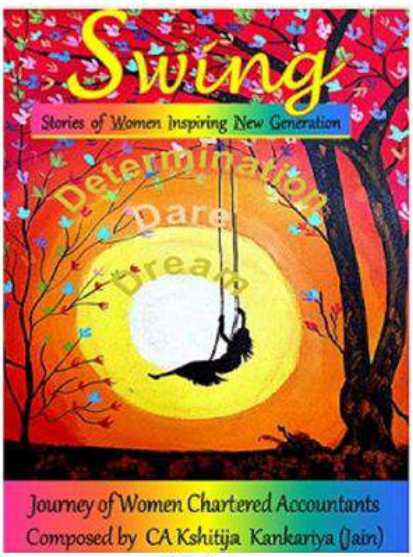


Madhav N Rode
Rameshwar R. Bhosale
ISBN : 978-81-943209-5-1

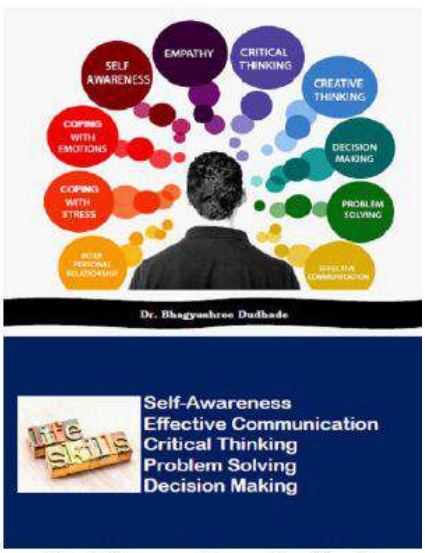


Dr. Sapna M S
Dr. Radhika C A
ISBN : 978-81-943209-0-6

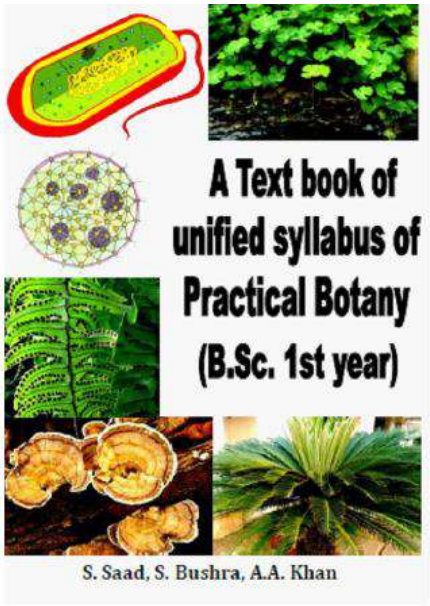
Hindusthan College
ISBN : 978-81-944813-8-6



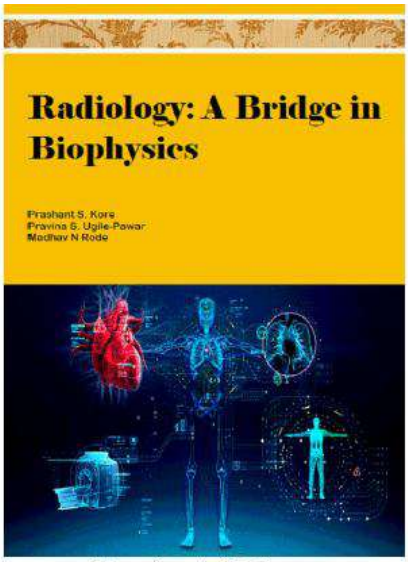
Swing
ISSN: 978-81-944813-9-3



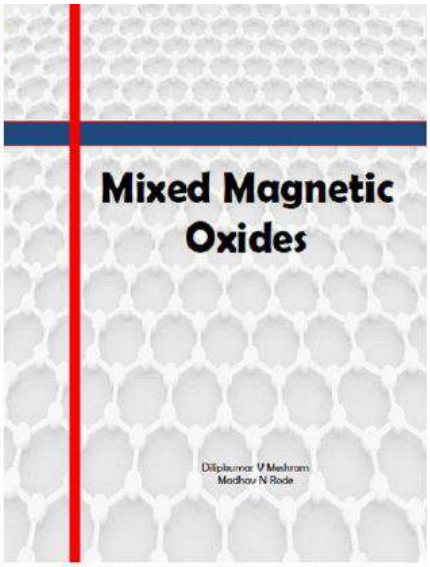
Dr. Bhagyashree Dudhade
ISBN : 978-81-944069-5-2



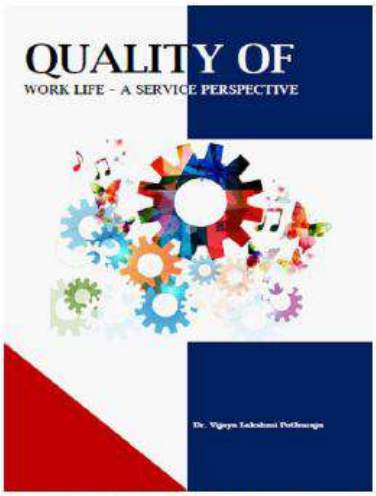
S. Saad, S. Bushra, A. A. Khan
ISBN: 978-81-944069-9-0



Prashant S. Kore
Pravina S. Ugile-Pawar
Madhav N Rode
ISSN: 978-81-944069-7-6



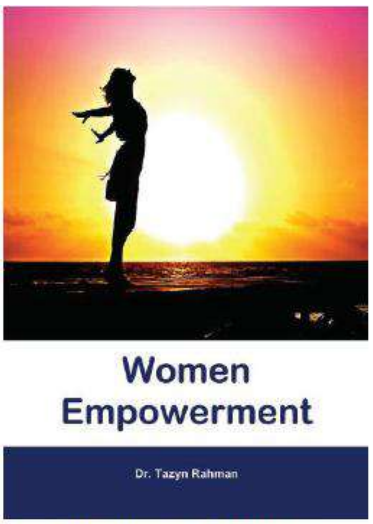
Dilipkumar V Meshram and
Madhav N Rode
ISSN: 978-81-944069-6-9



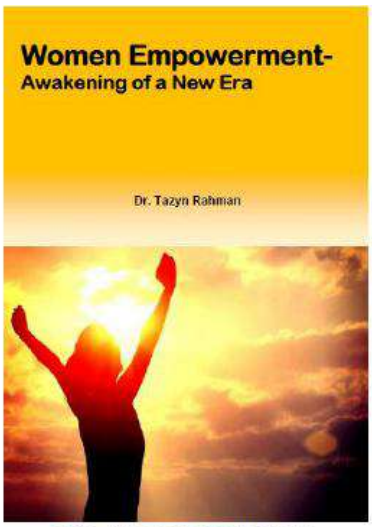
Dr. Vijaya Lakshmi Pothuraju
ISBN : 978-81-943209-2-0

Pratibha College
ISBN : 978-81-944813-2-4

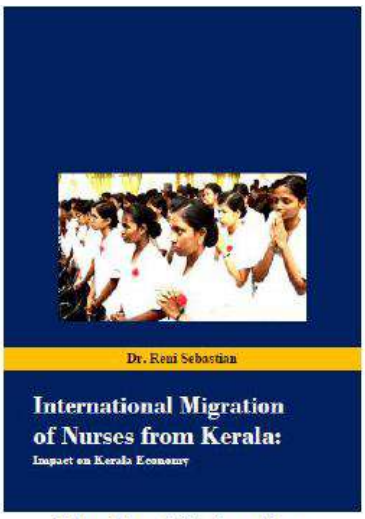
Pratibha College
ISBN : 978-81-944813-3-1



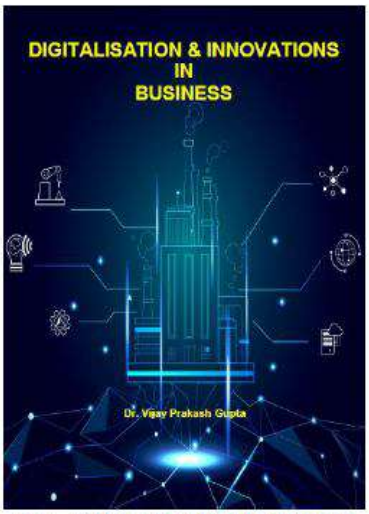
Dr. Tazyn Rahman
ISBN : 978-81-936264-1-2



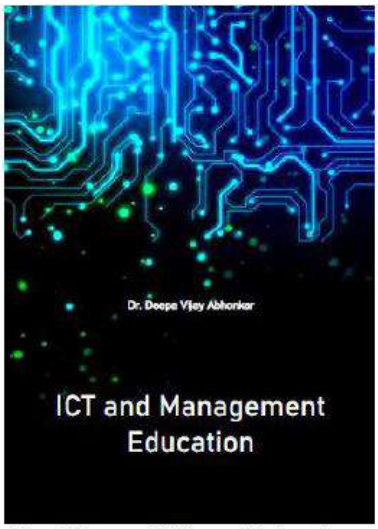
Dr. Tazyn Rahman
ISBN : 978-81-944813-5-5



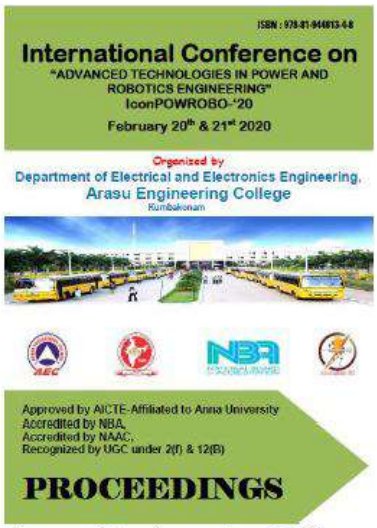
Dr. Reni Sebastian
ISBN : 978-81-944069-2-1



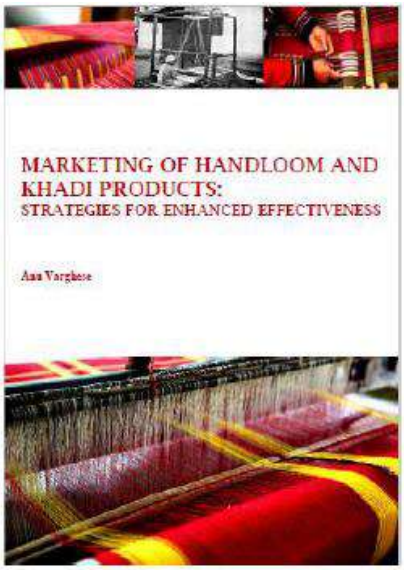
Dr. Vijay Prakash Gupta
ISBN : 978-81-944813-1-7



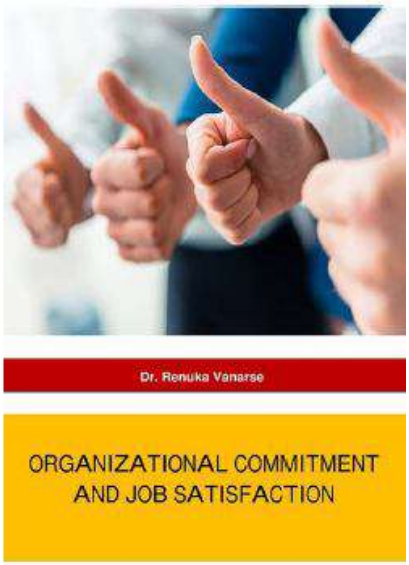
Dr. Deepa Vijay Abhonkar
ISBN : 978-81-944813-6-2



Arasu Engineering College
ISSN: 978-81-944813-4-8



Dr. Anu Varghese
ISBN : 978-81-944069-4-5



Dr. Renuka Vanarse
ISBN : 978-81-944069-1-4

Indian Academicians and Researchers Association
www.iaraedu.com