ISBN 978-1-7369640-5-7

Estd. in 2001

Conferred Autonomous Status by University Grants Commission (UGC) for 10 years w.e.f. AY 2019-20

ISO 9001:2015 Certified Institute	NBA Accredited Programs	NAAC Accredited Institute with 'A' Grade	AICTE-CII Survey rating in Platinum category for Industry linkages	Among Top 250 Colleges in NIRF Ranking	68th & 78th in All India Rank by Outlook survey published in June 2019 & May 2018 respectively

e-Conference on Data Science and Intelligent Computing (eC-DSIC) 2020:
Theory, Applications and Research Opportunities

November 27th & 28th 2020
Friday & Saturday

Organized by

Department of Computer Engineering

Editors

Dr. Harshali Patil

Dr. Anand Khandare

Jagdu Singh Charitable Trust's (Regd.)

THAKUR COLLEGE OF ENGINEERING & TECHNOLOGY

Autonomous College Affiliated to University of Mumbai

Approved by All India Council for Technical Education(AICTE) and Government of Maharashtra

A - Block, Thakur Educational Campus, Shyamnarayan Thakur Marg,
Thakur Village, Kandivali (East), Mumbai - 400 101

Tel.: 022-6730 8000 / 8106 / 8107 Telefax: 022-2846 1890 • Email: tcet@thakureducation.org

• Website: www.tcetmumbai.in www.thakureducation.org

e-Conference on Data Science and Intelligent Computing (eC-DSIC) 2020: Theory, Applications and Research Opportunities

Chief Patrons
Mr. V. K. Singh, Chairman

Patrons
Mrs. Karishmma V. Singh, Secretary
Mr. Karan Singh, CEO

Organizing Committee:
Organizing Program Chair
Dr. B. K. Mishra (Principal –TCET)

Organizing Program CO- Chair
Dr. Deven Shah (Vice-Principal-TCET)

Technical Program Chair
1. Dr. R. R. Sedamkar (Director-IQAC, HOD-Ph.D.)
2. Dr. Sheetal Rathi (HOD-PG)
3. Dr. Zahir Alam (TPO)

Convener
Dr. Harshali Patil (HOD-COMP)

Joint Convener
Dr. Anand Khandare (Dy.HOD-COMP)

Overall Coordinator
Mrs. Shiwani Gupta, Assistant Professor

Keynote Speakers:
1 **Data Engineering** by Mr. Sunil Pawar, (Assistant Vice-President, Big Data Architect, Barclays, Pune)
2 **Design of Computational Model** using ML by Dr. Vijay Bhaskar Semwal(Assistant Professor, NIT, Bhopal)

Technical Review Committee
Prof. Valentina Emilia Balas University of Arad, Romania
Dr. Selwyn Piramuthu, University of Florida
Dr. Dusko Lukac , University of Applied Sciences, Germany
Preggy Ready , Durban University of Technology, South Africa
Dr. Suneeta Agarwal, Professor, NIT, Allahabad
Dr. Mustaq Ahmed , Associate Professor , NIT Jaipur.
Dr. Padmaja Joshi, Director CDAC Mumbai, India.
Mr. Shitalkumar Dagde, Esamyak Software, PVT LTD, Mumbai
Dr. Sambit Kumar Mishra, GIET, Odisha

Estd. in 2001

e-Conference on Data Science and Intelligent Computing (eC-DSIC) 2020: Theory, Applications and Research Opportunities

PREFACE

"E-Conference on Data Science and Intelligent Computing (EC-DSIC 2020)" is a platform for conducting conferences with an objective of strengthening the research culture by bringing together academicians, scientists, researchers in the domain of data science and intelligent computing. The event is conducted online on 27th to 28th November 2020.

Data science is an inter-disciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from many structural and unstructured data. Data Science is a one of the thrust areas identified by AICTE. Data Sciences, will equip learners and researchers with the skills to lead the technological evolution rapidly changing the way business and policy is developed.

Intelligent Computing is heart of data science and consist of computational methodologies that mimic nature-inspired processes to address real world complex problems. It consists of various areas including artificial neural networks, evolutionary computation, genetic algorithms, artificial immune systems, fuzzy logic, swarm intelligence, artificial life, virtual worlds and hybrid methodologies.

EDSIC 2021 is not inculcating the research culture but also gained wide publicity through website, social media coverage as well as the efficacious promotion by the team of faculty members to the various colleges. The EDSIC 2020 has affiliation leading publication house and Conference proceeding with ISBN number.

TCET has strong belief in quality and relation building. A lot of care is taken for branding the event such as logistic support required for the event, compilation and printing of conference proceeding, souvenir etc. Total 30 papers received and 23 papers were presented in two days. Around 300 participants and delegates have attended the two days program. The delegates from national as well as International Researchers and Industrial personnel attended this event.

We also appreciate the efforts of all the members of the organizing and editorial committee for supporting the event and extending their cooperation to make it a grand successful event.

Team-EDSIC 2020
TCET, Mumbai

CONTENTS

PART-II COMPUTER ENGINEERIN

Cricket Match Winner Prediction Using Machine Learning

Sarvesh Kharche,Saranjeet Saluja,Yash Shah,Rohit Gupta
Department of Information Technology,Vidyalankar Institute of Technology Mumbai, India
sarvesh.kharche@vit.edu.in,saranjeet.saluja@vit.edu.in,sarvesh.kharche@vit.edu.in,rohit.gupta@vit.edu.in

Abstract— ***Cricket is one of the most famous sports activities in the whole world, and also one of the most popular sports in India. Cricketing occasions inclusive of the Indian Premier League (IPL) and One Day Internationals (ODIs) are thoroughly enjoyed by means of enthusiasts all throughout the country. Fans of the game love predicting the ongoing match results, and this is something that has ended up being a hobby for numerous people who observe the game. This is a sport with an abundant amount of statistics and using this information, we will make an evaluation on whether a team can win an ongoing IPL match or similarly, an ODI match. This prediction is implemented by the usage of system learning algorithms such as Multilayer Perception Classifier, Decision Tree Classifier, K-Nearest Neighbor and Random Forest. The required dataset is obtained via collecting the usage of an internet site and consolidated. As a result, the output is received which lists whether the home team has secured a win in the match or not.***

Keywords— Machine Learning, Data Mining, Cricket Match Winner, Score Prediction, Indian Premier League (IPL).

I. INTRODUCTION

Regardless of the type of format the sport is played in, cricket is a beloved and extremely popular sport in our country, having a massive fan-base. As fans, the people make their own predictions while watching a particular match based on the information given, they have and then, they make a call on who will win the match.

Cricket is essentially a bat and ball game that is played between 2 teams having 11 participants each. Each team comes to bat and has a sole inning in which it seeks to attain as many runs as possible, while the opposite team fields. The innings ends when the full quota of deliveries, which depends on the game format which is being played, or the ten batsmen have been dismissed, whichever comes first. The prime objective is to score more runs & therefore runs are the decisive factor.

There are 3 widely approved formats of cricket on the international level - T20, One Day Internationals and Test match. The scheduled length of the game is the prime distinction among these three formats, which without delay modifies the number of deliveries each team get to play of their respective innings. Test cricket format is the longest one and is acknowledged as the highest general of the sport. Match length is five days in which each squad gets to play 2 innings each. A regular day of a test match consists of three sessions, of two hours for every session.

One Day International i.e., ODI layout is of finite overs, where each team faces 300 deliveries (50 over's). Generally, ODI match falls in any of the two categories: Day or Day- Night match.

T20 is the shortest internationally identified format of this game, where each innings encompasses of 20 overs. This is extra of an "explosive" and greater "athletic" than the other two formats.

This research aims at predicting the result of an ongoing cricket match based totally on the information and data that is available from previous matches. The data that is available for each match includes the teams involved in the match, the venue, the winner of the match, the margin with which they won and the toss decision. We will be performing prediction for all of the matches that have taken place in the IPL. This is executed with the help of using machine learning algorithm for making the prediction of the result of the matches. This research is primarily based on predicting the winner of a cricket match based totally on the records available from previous matches.

II. RELATED WORK

Haghighat et al. [1] described the various prediction techniques including Nave Bayes which can be used to predict the best playing eleven for the team. Although this research work is aimed at specifically basketball, it can be extended to any sport, including cricket. This research work used basics of Python and the Scikit-learn API to implement the project. The different types of machine learning algorithms used in this research work included Support Vector Machine and Gaussian Naive Bayes.

D. Thenmozhi et al. [2] explains the comparison between four machine learning algorithms viz. Gaussian Naïve Bayes, Support Vector Machine, K-Nearest Neighbor and Random Forest applied on an IPL dataset.

Shubhra Singh et al. [3] The paper addresses the problem of predicting the outcome of an IPL cricket match. Factors such as luck and player strength were used as key functions in predicting the winner of a match. The novelty of the proposed method lies in addressing the trouble as a dynamic one and the usage of an appropriate non-relational database, HBase for scalability of application. Out of all the machine learning algorithms used, KNN has been located to be the maximum accurate.

Prakash et al. [4] proposed three variations of predictive models using Support Vector Machines to predict the winners of IPL matches.

Rory P. Bunker and Fadi Thabtah et al. [5] explains generating models and using machine learning algorithms such as Artificial Neural Networks to determine which team will win the match. This research

work was generalized to all sports, and so can be utilized for different sports including cricket.

Nazim Razali et al. [6] explains predicting the winner of a Football match in the English Premier League, which is a famous football tournament. This implemented by using Gaussian Naive Bayes, a machine learning algorithm. We have used similar to this in predicting the winner of the cricket match. This analysis makes use of a variety of models to identify and anticipate the winner of the match.

Jalaz Kumar et al. [7] shows data for ODIs was obtained from ESPNcricinfo [8] and scraped using a script by sending one request per second. Furthermore, the matches which finished in tie/draw or were disrupted due to rain were removed from the dataset as a part of data cleaning. Along with the above, the matches between special teams such as World XI or Asia XI were eliminated. The data for the matches reproduced by switching between two teams, i.e., a match between 1. Australia and 2. New Zealand was reproduced as 1. New Zealand and 2. Australia. The dataset thus, extracted and cleaned was converted into a categorical one from a continuous one by using counterfeit variables.

Factors included for analysis include:

- Previous performances of teams,
- The match being played on the home or the away turf
- Innings and
- Home turf upper hand.

The main aim of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by using certain tree based rules which is identified from prior data. In Decision Trees, for predicting a class label for a record we start comparing from the root of the tree. We compare the values of the root attribute with the record's attribute. Based on comparison, we follow the branch corresponding to that value and jump to the next node.

In KNN (K-Nearest Neighbors) algorithm the data points are classified based on the points that are most similar to it. It uses the prior data (training data) to estimate what an unclassified point should be classified as.

Random forest is a supervised machine learning algorithm that is used for both classification as well as regression. But however, it is mainly used for problems which involve classification. As we recognize that a forest is made up of trees and more trees means more robust forest.

Similarly, random forest algorithm creates decision trees on data samples and then receives the prediction from every tree and subsequently selects the best solution by using voting. It is an ensemble method that is better than a single decision tree as it reduces the over-fitting by using average of the result.

Thenmozhi et al. [2] In this research the author has predicted the outcome of an IPL cricket match using four different machine learning algorithms viz. Gaussian Naive Bayes, Support Vector Machine, K-Nearest Neighbor and Random Forest.

III. Proposed Methodology

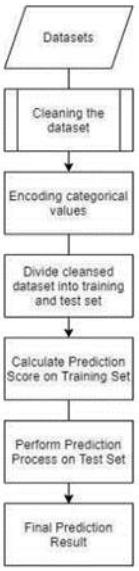


Figure 1. Flowchart of proposed Methodology

A. Dataset Selection

The data-set is obtained from Kaggle[www.kaggle.com/]. The data set consists of various attributes such as season(in years), the city and venue in which the match is being played, the teams involved in the match, the toss winner and decision(field or bat), the player of the match and the umpires for each match. The data set is cleaned by removing certain duplicate values, renaming certain values, and handling missing and null data.

B. Encoding Categorical Values

Certain categorical values such as the teams are replaced with numeric values, where an integer is assigned to each team. For example, all instances of the team 'Mumbai Indians' are allotted the integer value 1. Similarly, other categorical values such as toss decisions which consist of only two values (field/bat) are replaced by 0s and 1s, respectively. The remaining categorical values are dropped, as they have a lot of unique values which is not feasible.

C. Decision Trees

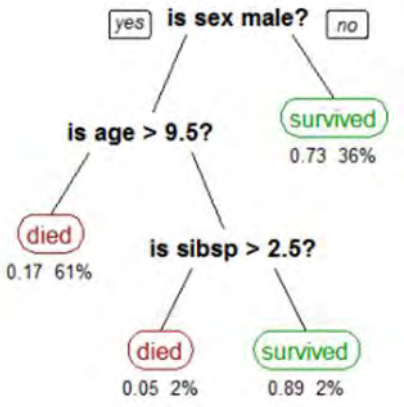


Figure 2. Example of Decision tree Classifier

A decision tree is drawn upside-down with its root on the top. In the figure above, the bold text in black represents a condition/inner node, based totally on which the tree splits into branches/ edges. The end of the branch that does not split anymore is the decision/leaf, in this case, whether or not the passenger died or survived, represented in red and green textual content, respectively.

Although an actual dataset will have a lot more features and this may simply be a branch in a much larger tree, however, you cannot ignore the simplicity of this algorithm. The feature importance is clear, and relationships can be easily viewed. This technique is more commonly known as a learning decision tree from data and the above tree is called Classification tree because the target is to classify if a passenger survived or died. Regression trees are represented in an identical manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are known as CART or Classification and Regression Trees. So, what actually happens in the background (growing a tree) involves identifying on which features to choose and what conditions to use for splitting, also knowing when to stop. As a tree typically grows arbitrarily, you will need to trim it down for it to be efficient.

D. Random Forest Classifier

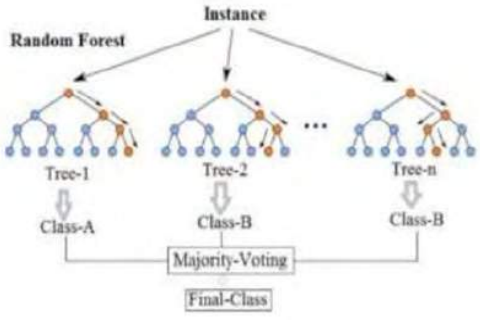


Figure 3. Random Forest Classifier

RF Classifier (Random Forest, RF-algorithm) is used here. RF classifier is an ensemble method which creates a myriad of decision trees during time of training. It finally takes the mode or average of the output classes by these trees.

Usually, as suggested[2],it is assumed that the objects of the data-set U which is used for the RF classifier development, are deleted into classes with the labels from the set $Y = \{1,2,...,\eta_1,...L\}$ (η_1 is the label of the l th class). A herewith, each object z_i ($i = \overline{1,s}$ where s is number of objects in data-set)can be described by the vector $z_i = (z_i^1, z_i^2, ..., z_i^n)$ of the numerical values in the n -dimensional space of features. The data-set U can be considered as the set $\{(z_1, y_1),...,(z_s, y_s)\}$, in which each object z_i has the class label y_i $(y_i \in Y = \{1,2,...,\eta_1,...L\})$.

Random Forest works in two-stages, first stage is to create the random forest by combining N decision trees, and second stage involves making predictions for each tree created in the first stage.

Steps involved in the working of Random Forest algorithm:

1. Select random number of data points from the training set.
2. Build decision tree with respect to each of the selected data points.
3. Choose a number N for the number of decision trees you want to build.
4. Repeat step 1 and 2.
5. For each new data point, find the predictions of the decision tree, and assign the new data points to the category that wins the majority votes.

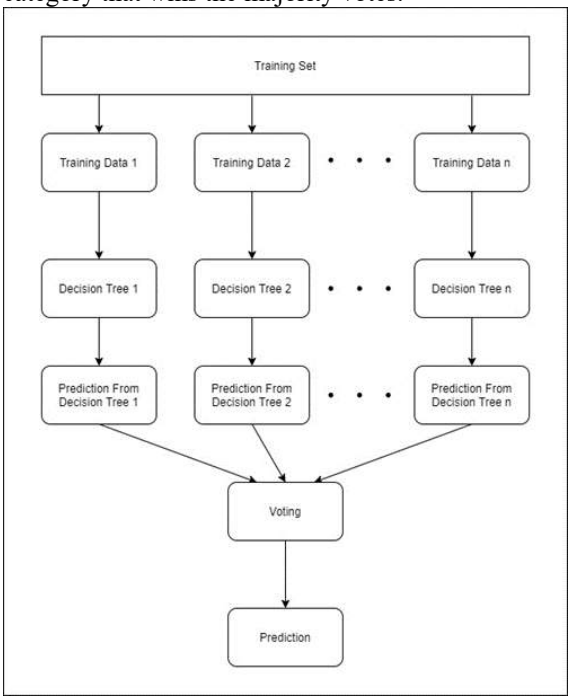


Figure 4. Random Forest Algorithm

Advantages of Random Forest Classifier:

- Highly accurate
- Estimation of attributes relevant for classification
- Generates an internal unbiased estimate of the generalization error as the forest building progresses.

Disadvantages:

- Prone to overfitting of data in a noisy classification.
- For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels

IV. RESULT

Figure 5. displays the final results after prediction. For each algorithm along the x-axis and accuracy along y-axis, the highest accuracy for each algorithm was depicted. Thus, the graph shows the highest accuracy for each algorithm.

Table 1 describes the highest accuracy obtained by each algorithm for a particular generated model. From the

results, it is observed that the most preferred algorithm which provides the highest accuracy is the Random Forest algorithm. This approach gives an overall accuracy of 87% on predicting the winners of the matches.

Figure 5. Accuracy achieved by different algorithms

TABLE I. OVERALL COMPARISON OF ACCURACIES

Algorithm	Accuracy
Decision Tree	80%
Random Forest	87%

V. CONCLUSION

Thus, the winner of the matches is predicted using various attributes using different algorithms such as Random Forest Classifier and Decision Tree. While the Decision Tree gives us about 80% accuracy, the Random Forest Classifier gives us better accuracy comparatively at 87%. Further, we aim to create a simulation system that will simulate the cricket match between two teams, selected by the user and accurately predict total runs scored by each team and thus predict the winner of the match along with the scorecard.

ACKNOWLEDGMENT (Heading 5)

We would like to show our deep appreciation to our project guide, Prof. Yash Shah, who helped us finalize our project, who gave us the opportunity to do the research and provided us invaluable guidance throughout this research.

REFERENCES

[1] Maral Haghighat, “A review of Data Mining Techniques for Result Prediction in Sports”, Advances in Computer Science : an International Journal (ACSIJ), vol. 2, no. 5, ISSN : 2322-5157, pp. 54–61, 2013.

[2] D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, Srivatsan Vasudevan, Veeramani Kannan V, Sagubar Sadiq S,” MoneyBall - Data Mining on Cricket Dataset”, Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019),2019.

[3] Shubhra Singh and Kaur, P., “IPL Visualization and Prediction Using HBase”, Procedia computer science, vol. 122, pp. 910–915, 2017.

[4] Prakash, C. D., Patvardhan, C., and Lakshmi, C. V., “Data Analytics based Deep Mayo Predictor for IPL-9”, International Journal of Computer Applications, vol. 152, no. 6, pp. 6-10, 2016.

[5] Rory P. Bunker, Fadi Thabtah, “A machine learning framework for sport result prediction”, Applied Computing and Informatics, vol. 15, no.1, pp. 27–33, 2017.

[6] Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim and Ruhaya Ab Aziz, “Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)”, International Conference on Material Science and Engineering, vol. 226, no. 1, pp. 012099:1–6, 2013.

[7] Jalaz Kumar, Rajeev Kumar, Pushpender Kumar, “Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks”,2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).

Detection and Performance Evaluation of Online-Fraud: A Review

Anam Khan Dr. Megharani Patil
Department of Computer Engineering Thakur College of Engineering & Technology Mumbai, India
anam.mail4u@gmail.com megharani.patil@thakureducation.org

Abstract— ***Fraud being committed by means of internet is termed as "Online-Fraud." Online fraud comes in many forms. It mainly involves financial fraud and identity theft. The perpetrators of online fraud are constantly using evolved techniques. In this paper, "Online Fraud" is used as an umbrella term to mainly focus on Click-bait Detection. Click-In social media, Click- baits are exaggerated headlines whose main motive is to mislead the reader to "click" on them. They create inconvenience in the online experience by creating a coax towards poor content. To get increased page views and thereby more ad revenue without providing the backing content online content creators are utilizing more of them. Click-baits are heavily present on social media platforms wasting the time of users.***

The currently build systems over LSTM-RNN show appreciable accuracy, precision, recall and support score, but as the research field is still active in trying to enhance the accuracy of the underlying systems, this paper aims to perform a comprehensive analysis by reviewing all the methods involved and techniques used till now for detection and performance enhancement of the models present.

Keywords— Machine Learning (ML), Deep Learning (DL), Neural Networks, On-line Fraud, Click-bait.

I. INTRODUCTION

Being a subfield of Artificial Intelligence (AI), the goal of Machine Learning (ML) generally is to know the structure of knowledge and fit that data into models which will be understood and utilized by people. In Machine Learning, tasks are generally classified wider. These categories are supported on how learning is received or how feedback on the training is given to the system developed.

There evolves a very young and limitless field i.e. Deep Learning (DL). It is a class of Machine Learning where theories of the subject aren't strongly established and views quickly change almost on a daily basis. Deep Learning is strengthening Artificial Intelligence and is playing a major role in all AI-based innovations through its successful execution. Also known as a subset of Machine Learning, Deep Learning is a specialist with an extremely complex skill set in order to achieve far more better results from the same data set. It relies completely on the basis of Natural Intelligence (NI) mechanics of the biological

Figure 1: Deep Learning Use-Cases

neuron system. It has a complex skill set because of its selected methods used for training i.e. Deep Learning models are based on "learning data representations" and not on "task-specific algorithms." which is the case for other methods.

"I think people need to understand that deep learning is making a lot of things, behind the scenes, much-better" – Sir Geoffrey Hinton

Thousands or even millions of simple processing nodes are densely interconnected in a neural network, being modeled loosely on the human brain. In today's scenario Neural Networks are organized into layers of nodes, and they're "feed-forward," implies that data moves through them in only one direction. An individual node can be connected to many nodes in the layer downwards, from which it receives data, and several nodes in the layer above it, to which it sends data. Neural Networks in Deep Learning are listed below:

- Feed-forward neural networks
- Recurrent neural network (RNN)
- Multi-layer perceptron (MLP)
- Convolutional neural network (CNN)
- Recursive neural networks
- Deep belief networks
- Convolutional deep belief networks
- Self-Organizing Maps
- Deep Boltzmann machines
- Stacked de-noising auto-encoders

As everything has their individual pros and cons, so as the continuously flourishing technology does have. It gave rise to Online Fraud. It is a way of using Internet services or software with Internet access to fraud victims or to otherwise take advantage of them. One of them is clickbait. Click-bait is all about teasing that predicament-putting forth false, exaggerated, mind-boggling claims or

pieces of data that are distant from reality, but peculiar enough for grabbing your attention. Being one of the lowest forms of journalism, authoring, and marketing, it has also become a pathway to Online fraud.

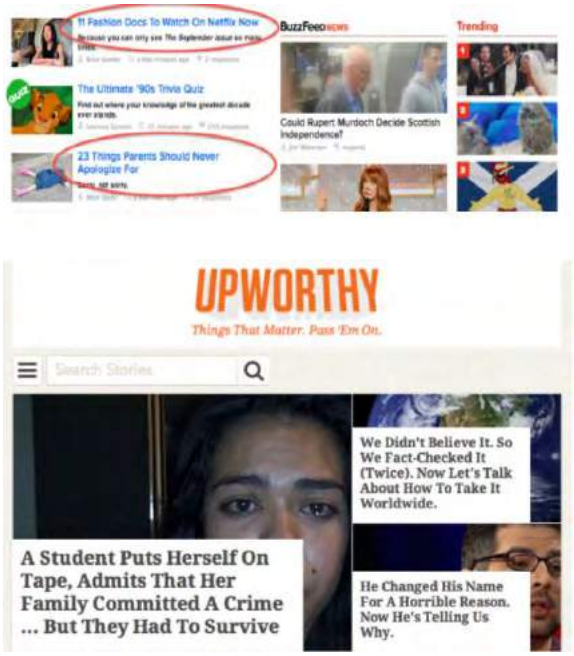


Figure 2: Click-Bait Headlines Collected from Various News Sites

II. LITERATURE SURVEY

The click-bait detection system developed here has one primary component. This component will eventually be combined with other features and further machine learning models to work on massive amounts of data to create a classification model based on specific attributes to help improve the accuracy of clickbait detection [1]. For data pre-processing and feature extraction, Beautiful soup and Pandas are used. For building the machine learning model, NumPy and various algorithms such as Gaussian Naive Bayes, Bernoulli Naive Bayes, Multinomial Naive Bayes and Multilayer Perceptron Classification algorithms from the Scikit Learn library were used.

In the following model distributed sub-word embeddings learned from a large corpus is used. The analysis also highlights how mainstream media is getting involved into clickbait practicing increasingly [2].
Close scrutiny of the social media posts also reveals that broadcast type media has higher percentage of usage of fraud practice than the print media and non-news type broadcast media mostly contributes to it.
In this work, use of variational autoencoders for tackling the clickbait problem on YouTube has been explored. The first proposed semi-supervised deep learning technique in the field of clickbait detection is practiced here [3].
Doing that, it enables more effective ways of automated detection of clickbait videos when large-scale labeled data is not available. The analysis indicates that YouTube recommendation engine does not take into account the clickbait problem in its recommendations.

The proposed system is used for stance detection of headlines with regard to their corresponding article bodies. The system is entirely based on simple, lemmatization-based n-gram matching for the binary classification of "related" vs. "unrelated" headline/article pairs [4].
Setup comprises of more fine-grained classification of the "related" pairs (into "agree", "disagree", "discuss") is carried out using a Logistic Regression classifier at first, then three binary classifiers with slightly different training procedures were used and the best result was obtained.

The model explores the utility of synthetically generated text in the context of click-baits, and concluded that synthetic click-baits can also be useful as additional labeled training data to train general ML models to detect click-baits better. It showed that VAE-based generative algorithms can generate high quality text that captures the most similar NLP feature distribution as the real ones among all synthetic sources [5].
Such an overlap in NLP feature distribution does not directly make synthetic click-baits as meaningful as real click-baits though, the outcomes resulted in a promising track using machines to generate realistic text in general. A novel direction toward solving the problem of insufficient training data in supervised learning can be solved using this framework.

III. TECHNIQUES USED FOR DETECTING ON-LINE FRAUD

A. CNN:

Convolutional Neural Networks (CNN) have been utilized for various deep learning tasks. Here, a simple CNN having one layer of convolution has been used.

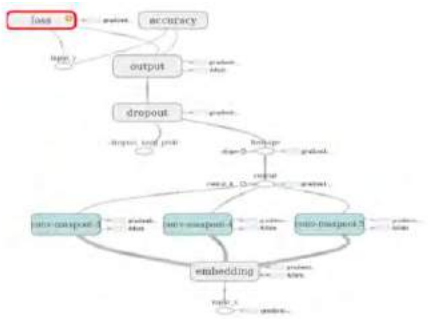


Figure 3: CNN Model

The first layer of the CNN [6] is used for embedding the words into vectors of low-dimensions. For word embeddings two variants have been used (1) word embeddings which are learnt from scratch, and (2) word embeddings which are learnt from an unsupervised neural language model which keep evolving as training occurs.

This technique of initializing word vectors from an unsupervised neural language model has been used to improve performance. The word vectors trained on 100

billion words of Google News have been used. These vectors are publicly available as word2vec.

B. Feature Engineering:

Here total 28 features are identified in the model. The more features are considered, the more time is needed for processing. Certainly, this might negatively affect the model's performance. It is not always true that adding more features will increase accuracy. Accuracy might decrease if a feature has a high correlation with another feature or is derived from another feature, the. More data is required to ensure there are enough samples for each combination of values if more features are added. Based on all these disadvantages of having a model with high dimensionality, recursive feature elimination to decrease the dimensionality of the model is been used. Recursive feature takes into account recursively considering smaller and smaller sets of features. Recursive feature elimination is been used because of the several factors affecting feature selection [7]. This is not the only factor to be considered but statistically, features with a correlation close to zero should be eliminated. The performance of the model can be decreased if a feature has high correlation with other feature. To ensure choosing the best discriminatory features, recursive method is adopted. After applying the algorithm, four features were removed.

C. Feature Extraction:

In order to capture the lexical differences between the two classes – clickbait and non-clickbait the following features
were selected:

• Sentence Composition – Number of stop-words, Length of the headline, interrogatory words.
• Word structure – punctuation patterns, use of numbers in headlines, personal pronouns, adverbs.
• Language Analysis-Number of slangs, sentiment of words used, attention seeking words.
• Lexical Nuances- N-grams and Part-of-speech tags

The need to capture the semantic and [8] syntactic dependencies prompted to the use of distributed word embeddings in addition to the lexical features. GloVe and Word2Vec is used to draw a comparative analysis between the two to see which yields the best results.

D. GloVe:

It is an algorithm for unsupervised learning that is used for representing words as vectors. Using GloVe [9] 6B, the tokens were embedded into the 300-dimensional word vectors space. This model works by first pre - computing the co-occurrence matrix, followed by factorizing it to reduce the dimensionality of the matrix by looking at the context in which the given word appears in the corpus. Thus, it is a count-based model that works on the principal of dimensionality reduction and is easier to parallelize.

E. Word2Vec:

It is used for mapping words in the sentences into multi-dimensional [10] vector space by sliding a window through the sentences and calculating the co-occurrence individually.

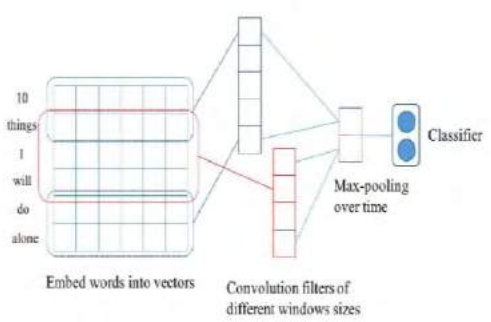


Figure 4: Word2Vec Embedding

F. Feature Selection:

To capture the structural properties of the headline as classifier features, the length of the headline is used, the average length of words, the ratio of the number of stop words to the number of content words and the longest separation between the syntactically dependent words of a headline totally constitute the sentence structure.
Word level structural features was included because in the beginning of the headline cardinal numbers were present, presence of unusual punctuation patterns and the number of contracted word forms employed in the headline.
To capture the variations of the language employed, especially in the clickbait headlines, features like, presence of hyperbolic words, common clickbait phrases, internet slangs and determiners are used. To model the popularity of the subject word in clickbait headlines as a feature, the score of a multinomial Naive Bayes classifier over sentence subjects is used [11].
Given the subject word of the sentence, ratio of the probability of assigning clickbait class label to the probability of assigning non-clickbait class label, is calculated and represented as score. Model parameters for the Naive Bayed classifier were estimated using both datasets.

G. N-gram Features:

Word N-grams, POS N-grams and Syntactic N-grams were used as features. N-gram feature space grows linearly with the size of the dataset. In order to limit the number of N-gram features used based on their frequency of occurrence, the feature space was pruned efficiently by using the sub-sequence property and an APRIORI-like algorithm [12]. Similar to the case with subject words, three multinomial Naive Bayes classifiers for the three sets of pruned features were built, i.e. the Word N-grams, POS N-grams and Syntactic N-grams.
The scores of these three auxiliary Naive Bayes classifiers were used as inputs (i.e., as features) to the main classifier.

H. Siamese Neural Network with Visual Embeddings:

The final component of the hybrid model is a Siamese net. However, it considers visual information available in the dataset, and sets the model apart from other approaches in this field. The relevance of the image attached to the post can be quantified by capturing its similarity with the target description. A 4096-dimensional vector for each image which is an output of VGG-19 architecture in turn, is fed as input into a dense layer converting each representation to a 300-dimensional vector. This further used as an input to the visual Siamese net. By passing it through the pre-trained Doc2Vec model, the target description is converted into its 300-dimensional vector representation which acts as the second input for the network. It is the rightmost part of the model [13].

The data extracted from the embedding layer is fed to the CNN module. Where the features are detected at different regions with the help of a sliding filter vector evolved in the convolution layer.

IV. EXISTING SYSTEM

The existing system here taken into consideration is a LSTM-RNN structured model.

Recurrent Neural Network Models:

Recurrent Neural Network (RNN) is a class of artificial neural networks [14] which utilizes sequential information and maintains history through its intermediate layers. The output at each time-step is dependent on that of the previous time-steps of an internal state of RNN.

Long Short-Term Memory (LSTM):

Standard RNNs have difficulty preserving long range dependencies due to the vanishing gradient problem [15]. This is related to interaction between words that are several steps apart. The LSTM using gating mechanism is able to diminish this problem.

Model	Accuracy
LSTM-RNN	90%

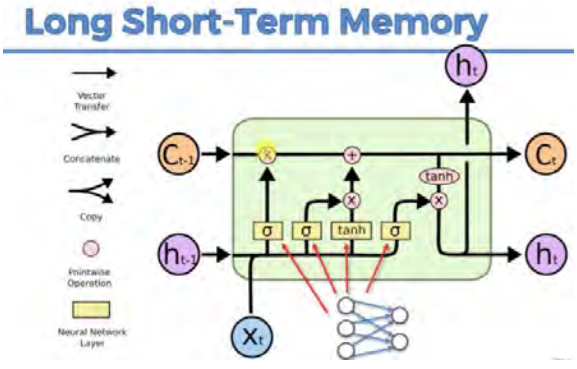


Figure 5: Detailed LSTM Architecture

The following iterative process are needed by each LSTM cell to compute its internal state:

- Cell State(C) is the key of LSTM which store information.
- LSTMs also has the ability to remove or add information to the cell state.
- The ask of removing or adding information to the cell state is regulated by structures called Gates.
- A sigmoid neural network layer and a pointwise multiplication operation constitute Gates.
- Numbers between 0 and 1 are generated by sigmoid layer as output.
- Here, 1 represents “completely keep this” and a 0 represents “completely get rid of this.”
- The three key components forget, update and output gate in total constitutes the LSTM unit.

Dataset:

We evaluate our method on a dataset of 20,800 clickbait and non-clickbait headlines dataset released on Kaggle website. The headlines are evenly distributed a 10,413 a nd 10,387 click-bait and non-click-bait resp. each. It has five columns and numerical labels of clickbait in which 1 repre-sents that it is clickbait and 0 represents that it is non-clickbait headline.

Word Encoding:

Another popular technique for treating categorical variables is One-Hot Encoding. Based on the number of unique values in the categorical feature it simply creates additional features. A new feature will be added for every unique value in the category.

Training setup: For training our model, we use the mini-batch gradient descent technique with a batch size of 64 and 10 epochs, the ADAM optimizer for parameter updates and Binary Cross Entropy Loss as our loss function. To prevent overfitting, we use the dropout technique with a rate of 0.3 for regularization. To learn effective representations for this specific task during training, the character embeddings are updated. Our implementation is based on the Keras library using a TensorFlow backend.

V. RESULT

Table 1: Accuracy value of the Existing System

Table 2: Various Evaluation Parameters of the Existing System

	Precision	Recall	F1-Score
0	0.93	0.90	0.92
1	0.88	0.91	0.89

VI. CONCLUSION

In short, Deep Learning goes much beyond machine learning and its algorithms that are either supervised or unsupervised prove computationally much efficient. Many layers of nonlinear processing units for feature extraction and transformation are used in Deep Learning.

It focuses on end-to-end learning supported raw feature where traditional machine learning focuses on feature engineering, Traditional deep learning creates/ train-test splits of the info where ever possible via cross-validation. The best part for anyone looking to find out deep learning only must have some working knowledge/background of calculus, fundamental algorithms, algebra and powerful applied mathematics only.

One thing needless to say about deep learning is, it teaches anyone a deep understanding of the math behind neural networks and the way deep learning libraries work.

REFERENCES

[1] Suraj Manjesh, Tushar Kanakagiri, Vaishak P, Vivek Chettiar, Shobha G, "Clickbait Pattern Detection and Classification of News Headlines using Natural Language Processing", 2nd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions, 2017.

[2] Md Main Uddin Rony, Naeemul Hassan, Mohammad Yousuf, "Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2017.

[3] Savvas Zannettou, Sotirios Chatzis, Kostantinos Papadamou, Michael Sirivianos, "The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube", IEEE Symposium on Security and Privacy Workshops, 2018.

[4] Peter Bourgonje, Julian Moreno Schneider, Georg Rehm, "From Clickbait to Fake News Detection:An Approach based on Detecting the Stance of Headlines to Articles", Proceedings of EMNLP Workshop on Natural Language Processing meets Journalism, 2017.

[5] Thai Le, Kai Shu, Maria D. Molina, Dongwon Lee, S. Shyam Sundar, Huan Liu, "5 Sources of Clickbaits You Should Know! Using Synthetic Clickbaits to Improve Prediction and Distinguish between Bot-Generated and Human-Written Headlines", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2019.

[6] Amol Agrawal, "Clickbait Detection using Deep Learning", 2nd International Conference on Next Generation Computing Technologies -NGCT, 2016.

[7] Daoud M. Daoud, M. Samir Abou El-Seoud , "An Effective Approach for Clickbait Detection Based on Supervised Machine Learning Technique", International Journal of Online and Biomedical Engineering (iJOE),Volume 15, Issue 3, 2019.

[8] Saumya Pandey, Gagandeep Kaur, "Curious to Click It?-Identifying Clickbait using Deep Learning and Evolutionary Algorithm", International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

[9] Peter Adelson , Sho Arora , and Jeff Hara, "Clickbait; Didn't Read: Clickbait Detection using Parallel Neural Networks", 2018.

[10] Hai-Tao Zheng , Jin-Yuan Chen, Xin Yao , Arun Kumar Sangaiah, Yong Jiang, Cong-Zhi Zhao," Clickbait Convolutional Neural Network", 2018.

[11] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, Niloy Ganguly, "Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016.

[12] Philogene Kyle Dimpas, Royce Vincent Po, Mary Jane Sabellano "Filipino and English Clickbait Detection Using a Long Short Term Memory Recurrent Neural Network", 2017.

[13] Vaibhav Kumar, Dhruv Khattar, Siddhartha Gairola "Identifying Clickbait: A Multi-Strategy Approach Using Neural Networks", 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, 2018.

[14] Kai Shu, Suhang Wang, Thai Le, Dongwon Lee,Huan Li, "Deep Headline Generation for Clickbait Detection", IEEE International Conference on Data Mining, 2018.

[15] Praphan Klairith, Sansiri Tanachutiwat , "Thai Clickbait Detection Algorithms using Natural Language Processing with Machine Learning Techniques", International Conference on Engineering, Applied Sciences, and Technology (ICEAST), 2018.

Air Quality During COVID 19 Lockdown

Tarunima Mukherjee, Dr.Harshali Patil
Computer Engineering Thakur College of Engineering & Technology Kandivli East Mumbai, India
tinaganguly24@gmail.com, harshali.patil@thakureducation.org

Abstract: - ***Air pollution is one of the major crisis which is omnipresent across the globe. The growing industrialization and modernization are further aggravating the problem. Various policies are being implemented by the governments all over the world to curb this problem. In India, the air quality drastically improved in 2020, due to total lockdown. The total lockdown was announced on 22 March 2020 to stop the spread of COVID-19 and the lockdown was extended in phases to curb the virus. COVID 19 has forced the government to impose the lockdown in the country which has affected the various strata of society adversely, however the lockdown has also shown some positive impact on the natural environment. Air Quality has considerably improved in this course of time. During the total lockdown, most of the sources for poor air quality like factories, automobile movements and industries were stopped in India which extensively reduced the air pollution. Severe Health Issues like respiratory and cardiovascular diseases are caused when the pollutants in the air exceeds the safe limit. In this paper we discuss how the air quality is impacted due to the lockdown and we also discuss the significance of air quality prediction using machine learning techniques to maintain the quality of the air.***

Keywords – Air Pollution, COVID 19, Lockdown, Air Quality, Health Issues, Prediction, Machine Learning

I. Introduction

Air Quality plays a vital role in healthy living of human beings. Air pollution is considered to be a severe crisis around the world, but a lockdown showed positive results in curbing the air pollution thereby improving the air quality. In 2020 the world faced a major global crisis which was of a rare type. COVID 19. It is caused by a virus named Corona. It incepted in December 2019 and hence the name COVID-19 It was first identified in WUHAN China and has resulted in an ongoing pandemic. The disease mainly spreads by physical proximity. It spreads contagiously and easily by the air, primarily through droplets and in aerosols, when an infected person coughs, talks, or sneezes. The worst part is that there is no vaccine available for this virus and that is the reason that the infection has increased exponentially.

The only way to contain this pandemic was social distancing and hence a lockdown was imposed by the Government of India on 23 March 2020. Lockdown means to be in isolation or restricted access implemented as a security measure. A lockdown requires people to stay where they are, because if they move freely they can be of specific risks to themselves or to others . Due to the COVID-19 pandemic widespread lockdowns were issued.

The lockdown has affected different strata of society adversely especially the migrant workers and daily wage earners. It has also affected people mentally and emotionally. It has created a turmoil in the lives of the people. However, the only positive aspect which emerged from this situation was that there was a tremendous decline in the air pollution and the air quality has improved drastically. Due to this nationwide lockdown most of the mass transportation and industrial activities were restricted. As a result, the pollution level in 88 cities across the country drastically reduced down just after four days of commencing lockdown according to the official data from the CPCB.[1]. Air quality is a measure used to identify how clean or polluted the air is.

Air quality monitoring is important as polluted air can be hazardous for our health and the environment. Air quality is measured by calculating Air Quality Index, or AQI. The AQI is a way of showing changes in the amount of pollution in the air. Of late technological advancement along with machine learning techniques and information technology, collection and compilation of real-time site-specific air pollution data is in practice throughout the world.

The main objective here is to study the various pollutants present in the air which causes air pollution and health problems followed by as to how they are monitored and predicted using machine learning algorithms. The basic purpose is to study the impact of lockdown due to COVID 19 on the air quality across India .

II Air Quality Index

. The earth's atmosphere is designed in such a way that it can sustain life. The atmospheric air is made up of mainly two gases that are essential for life: nitrogen and oxygen. However, the air also comprises of smaller amounts of other gases and particles like carbon dioxide, argon with very small amounts of helium, neon, krypton, methane and hydrogen. There are also many pollutants which are present in the air. AQI tracks six major air pollutants which are listed as below:

1. PM 10
2. PM 2.5
3. Ground level ozone
4. Nitrogen dioxide
5. Sulfur dioxide
6. Carbon monoxide

Particulate Matter 10 (PM 10) are the coarse particle with a diameter of 10 micrometers or less. Upon inhalation they penetrate deep into the lungs and when humans are subjected to it for long term and high concentrations of

PM10, it can cause several health issues ranging from wheezing, coughing to asthma attacks, bronchitis, heart problems and even strokes. PM10 is generated in the environment from the dust generated at construction sites, agriculture, forest fires, burning of wastes, various industrial sources, and even dust from open lands.

Particulate matter 2.5 have more unfavorable effect on the human health as compared to the other pollutants. These are solid and liquid droplets suspended in the air. Being extremely small and light, these particles tend to stay longer in the air thereby increasing the chances of getting inhaled by human beings. These particles can easily enter into the respiratory system through inhalation process and can affect the lungs and the breathing phenomenon. Also, it has the potential to cause cardiovascular diseases in people of any age group. Construction sites, smokestacks, car exhaust pipes, wildfires etc. are the sources of generation of these airborne particles. These particles are also formed due to chemical reactions in the atmosphere.

Ground level ozone is harmful for human health. It gets formed when the sunlight reacts with some specific chemical emissions like nitrogen dioxide, carbon monoxide and methane. These chemicals are generated from various industrial establishments, car exhausts, gasoline vapors etc. Short term exposure to high levels of ozone causes eye and lung infections. [2]

The important nitrogen containing gases which contribute to air pollution are nitric oxide (NO) and Nitrogen di oxide (NO2). NO2 causes inflammation of air passages of the human body which in turn deteriorates the functioning of the lungs. It can also increase the asthma attacks and worsen the cough and cold. NO is emitted by natural as well as anthropogenic sources. The main anthropogenic sources of nitrogen oxides are fuel combustion in domestic, industrial, power generation or transportation activities. Fuel combustion generates NO to the extent of 0.1% to 0.5% at flame temperatures, along with much smaller amounts of NO2. NO2 is also formed by oxidation of NO and with high concentration NO2 can cause lung and heart problems. [3]

The oxides of sulfur (SO2 and SO3) in form of Hydrogen sulfide, sulfuric acid mist & sulfate salts are the major sulfur containing pollutants of the atmosphere. The sources of atmospheric sulfur compounds are combustion of fossil fuels, decomposition of organic matter, sea spray over oceans and volcanoes. Long term exposures to high concentrations of SO2 give place to respiratory illness and aggravation of existing pulmonary and cardiovascular disease.SO2 also causes irritation to the eyes, skin and the mucous membranes. It can also cause bronchospasm and pulmonary edema.

Carbon Monoxide is one of the most abundant air pollutants in the atmosphere. It is generated due to incomplete combustion of fossil fuels specially from the engines of automobile. Emission of carbon monoxide exceeds in quantity to the emission of all other air pollutants from anthropogenic activities. Coal and wood burning, refining of petroleum, usage of solvent etc. are the anthropogenic sources. Carbon monoxide enters through the lungs and gets mixed into the bloodstream subsequently reducing oxygen supply to the organs and tissues in the body. Carbon Monoxide in small amounts, can causes lethargy, slows down the reflexes and impairs judgement. If the concentration is high it can cause death.

Due to its above-mentioned unfavorable effects, the pollutants concentration is monitored actively by almost all the civic bodies across the globe and is used as a basis for calculating the air quality index (AQI).

Air quality index is a resultant of pollutant concentration and is a dimensionless number. Different values of AQI represent different quantities of air pollution. If PM2.5 concentration is lower, it means lower value of AQI and hence a healthy air while on other hand higher concentration refers to a higher AQI value and an unhealthy air. The below method is used to calculate the value of AQI from the concentration of pollutant.

$$AQI = \frac{(AQI_{Hi}) - (AQI_{Lo})}{(Conc_{Hi}) - (Conc_{Lo})} \times ((\mathbf{Conc_i}) - (Conc_{Lo})) + (AQI_{Lo})$$

Where,

Conci = Input concentration for a given pollutant

ConcLo = The concentration breakpoint that is less than or equal to Conci

ConcHi = The concentration breakpoint that is greater than or equal to Conci

AQILo = The AQI value corresponding to Conci

AQIHi = The AQI value corresponding to Conci

When the air quality is bad enough people should stay inside. Air quality is good when AQI is under 50 and at this level, a person can spend time outdoors and air pollution will cause very little risk to their health. As the AQI level increases, the risk to human health also increases. The table gives the summary of the AQI levels of health concern.[2]

Table I. Summary of AQI Levels

Level of Health Concern	AQI Value	Meaning
Good	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Unhealthy	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alert: The risk of health effects is increased for everyone.

Hazardous	301 to 500	Health warning of emergency conditions: everyone is more likely to be affected.

II. Methodology

In order to assess the condition of air quality in India during the lockdown period, data from various air quality monitoring stations covering different regions has been taken into consideration. The monitoring organization for these air quality monitoring stations includes CPCB – Central Pollution Control Board, MAAQM - Manual Ambient Air Quality Monitoring, CAAQM - Continuous Ambient Air Quality Monitoring, IITM - Indian Institute of Tropical Meteorology and SAFAR - System of Air Quality and Weather Forecasting and Research. The daily or hourly concentration of air pollutants including Particulate Matters (PM2.5 and PM10), Sulphur Dioxide (SO2), Nitrogen Dioxide (NO2), Carbon Monoxide (CO), Ozone (O3). CPCB has defined rigorous protocols for the sampling, analysis and calibration to ensure the quality of air pollution data.

The AQI is usually based on pollutants criteria where the deliberation of an individual pollutant is transformed into a sole index using appropriate aggregation method [5] Calculation of the AQI is based on the maximum sub-index approach using the five kinds of air pollutants that is O3, PM2.5, SO2, NO2 and CO.

In this technological era, where all objects are connected has given rise to the concept of Internet of Things (IoT). Newly proposed systems are based on IoT sensors. Data gathered from sensors play a vital role in helping the cities manage and measure air quality. With the help of sensors generating data, the smart city decisions have been made much faster and easier, but the processing of data brings its own challenges.

A big challenge in handling a smart city's information is to make it more efficient and reliable in addition to being rapidly analyzed. False alarms are dangerous as it can lead to wrong decisions. Prediction using artificial intelligence and use of right data is very important to take appropriate decisions in order to ensure the robustness of speed and processing of communication sources. Furthermore, to analyze the pollution data, machine learning predictive algorithms are being leveraged to provide the right information at the right place with minimum errors at the right time.

There are various machine learning algorithms used for air quality prediction and are explained as below.

There are many more machine learning techniques which can be used for the prediction of air quality according to the requirement and convenience. Factors like amount of data, number of pollutants and the location plays a very important factor in deciding the technique which is to be used for air quality prediction. The aim is to achieve maximum accuracy so that the air quality can be improved and the health risks due to air pollution can be minimized.

1) Neural Network Techniques: -

The extensively used statistical method uses artificial intelligence (AI) for prediction. The accuracy of neural network (NN) forecasting models is higher than that of other statistical models, but they should be improved. Grivas developed an artificial neural network (ANN) which combined the time-scale input and meteorological variables. An ANN based forecast tool uses meteorological parameters and the emission pattern of sources. These ANN models are found to be more effective when used with same input parameters. Recurrent Neural Network, Long Short-Term Memory (LSTM), Convolutional Neural Network LSTM (CNN- LSTM) and BLSTM are the extensively used methods for air quality prediction and have given better performance when compared to other methods. Multilayer Perceptron is also widely used.

2) Regression Techniques: -

Random Forest in a simple way is a collection of decision trees by which it can analyze set of variables and reduce the high variance problem. By which it can be said that random forest is an extension of bagging. Bagging is a concept where decision trees are built based on the multiple samples of the training data. In the Random Forest approach, first the data is loaded then converted into the required string attributes. In the next step cross validation split is performed which helps in estimating the performance of all the pollutants and AQI. Accuracy metrics are used to evaluate each model. Then by using get _split method particular number of input attributes can be selected which is considered as sample for actual training dataset. Then Gini_index is used which is the attribute selection measure used for calculating the cost function, the minimum the value of cost function the more pure is the data. Next the data is split based on split points and Gini index. In order to build and predict the single decision tree having six labels, To_terminal, split and build_tree are used. Bagging prediction method is used to predict the group of decision trees, which gives better accuracy rate as the number of trees increases the accuracy rate also increases. [6]

Decision Tree is a supervised learning algorithm. A decision tree has an internal node which signifies an attribute, the branch signifies a decision tree rule, and the leaf node signifies the result. The topmost node in a decision tree is recognized as the root node. Decision Tree working for the Indian air quality dataset In the Decision tree approach first we check for the purity that is unique columns like so2, no2, spm, type and then classify the data accordingly by excluding the output column i.e. class labels column i.e. AQI level, next split the data as continuous and categorical, after dividing the data the entropy of the data frame is calculated, then after the overall entropy of the data frame is calculated. In decision tree algorithm counter

values and max depth of the tree are considered as key points. And then the accuracy of the AQI level is calculated, for finding the accuracy the depth of the tree and the test data frame is considered. [6].

SVMs are supervised learning methods used for classification and regression and have the ability of being a universal approximators for any multivariate function. The SVM was originally developed for classification and was later generalized to solve regression problems This method is called support vector regression (SVR). The support vector classification model depends only on a subset of the training data, because the cost function does not take the training points into account that lie beyond the margin. On the other hand, the model produced by SVR the cost function for building the model ignores any training data that are close that is within a threshold ε. The basic idea of SVR is instead of attempting to classify new unseen variables into one of two categories we now wish to predict a real-valued output for y. In case of regression SVM, if the predicted value is less than a distance ε away from the actual value then no penalty is allocated as a more sophisticated penalty function is being used in the regression SVM. [7]

3) Ensemble Techniques: -

Various ensemble approach is used for real-time air quality forecasting. The ensemble model integrates two or more models. The extensively used models for ensemble techniques are Adaboost, GBDT and XGboost.

Adaboost: - Ada Boosting was the first successful boosting algorithm developed for binary classification. Ada Boosting is basically used to boost the decision trees. performance. In this, the tree is first created, the performance of this tree on is used as training instance. It is also used to weight the amount of attention provided to the next tree. So higher weight is provided to the training data which is hard to predict and less weight is provided to instances which are easily predictable. Weighted average of the weak classifiers is calculated for prediction.

GBDT: - Gradient boosting algorithm is used to combine weak learners into a single strong learner in an iterative fashion. At the end of every step, the weight of the wrong classification points increases, and the weight of the correct classification points reduces. This enables resolving some points which were getting misclassified previously; that is, high weight is given. The smaller the base classifier weight value is, the smaller the error rate is. The smaller the error rate is, the larger the weight value of the base classifier is. After number of iterations, we obtain a simple basic learner. We continue to merge the outputs together and choose the one that has the greatest votes.

XGBOOST :- XGBoost model is one of the boosting ensemble algorithms, which is based on the lifting tree model, so it ensembles many tree models together to form a strong classifier. At the same time, XGBoost model is improved based on Gradient Boosting Decision Tree (GBDT), making it more powerful and applicable to a wider range. Therefore, XGBoost model has the advantages of fast computing speed, strong model generalization ability, and significant model improvement effect.

4) Hybrid Techniques: -

In Hybrid techniques different variations of machine learning algorithms are combined to enhance the performance and increase the accuracy with reduced error rates.

IV Conclusion

The most significant aspect which has emerged from this lockdown is the fact that we can have “clean air” and achieve the national ambient standards. The lockdowns period of two months, with complete restriction in the first phase and with some relaxations in the fourth phase demonstrated that the only way to achieve “clean air” is by reducing emissions. There are not many alternatives or short-cuts to achieve this goal.

Secondly it is observed that a city-centric or sector-centric approach is not sufficient. Even after these COVID times, we must plan to cut emissions at all the sources which are inside or outside the city limits, in order to maintain the better air quality. We should avoid temporary solutions like smog-towers, mist-cannons etc. which can only benefit small areas of the city.

These reductions demonstrated to the public that the only way to achieve “clean air” is by cutting down the emissions at the source, and not at the cost of social and economic curbs forced during the lockdown. Air quality changes observed during the lockdown also show that we must adopt an airshed approach where the reductions in the city are complemented with similar reductions across the region to achieve and sustain the national ambient standards. In order to reduce the health risks caused due to pollution we must majorly emphasize on two factors which includes Air Quality Monitoring and Forecasting. Monitoring needs to be done effectively in relevance to population exposures, measuring key contaminants to standards that can be compared across locations. Monitoring of PM2.5 should have the highest priority, followed by ozone, carbon, and NO2. Forecasting needs to be done accurately by using robust machine learning algorithms to anticipate specific hazardous conditions in order to take action to improve air quality, to advise the public by providing Air Quality Indexes and to forecast future trends and problem areas in order to formulate environmental policies and multi-

sectoral interventions to protect health. In absence of robust monitoring & forecasting poor air quality imposes great amount of risk to human health and deteriorates the environment. It not only causes severe lungs, cardiovascular, skin damage and chemical sensitivity diseases in humans also it can have catastrophic effects on the environment like acid rain, global warming and climate change. Due to this it can also lead to extinction of animal species and deterioration of fields.
The combinations depend on the requirements and the quality of the data set. Many hybrid models are used such as various combination of Arima, SVM, ANN and many more for prediction of air quality.

There are many more machine learning techniques which can be used for the prediction of air quality according to the requirement and convenience. Factors like amount of data, number of pollutants and the location plays a very important factor in deciding the technique which is to be used for air quality prediction. The aim is to achieve maximum accuracy so that the air quality can be improved and the health risks due to air pollution can be minimized.

In order to study the possible impact of this unconventional policy intervention in form of lockdown on air pollution, pollutant parameters like PM10, PM2.5, SO2, NO2, O3 & CO have been analyzed individually during the lockdown period and then compared with the result of the same for the pre-lockdown period for India. The below table represent [8] the concentrations of the various pollutants in the pre and post lockdown phases.

Table 2. Impact of Lockdown on Various Pollutants

Pollutant	Pre-Lockdown	Post Lockdown
PM 2.5	Before the lockdown the concentration of PM 2.5 was high in many regions across India and was greater than 60 ug/m3	PM 2.5 Concentrations reduced to the lower ranges of 0.1 to 40 ug/m3
NO2	The concentration of NO2 was in the range of 40 to 50 ug/m3.	The concentration has reduced to 0.1 to 10 ug/m3 after lockdown
SO2	Concentration of SO2 before lockdown in many regions was above the range of 25ug/m3	After lockdown the concentration came down to the range of 10 to 15 ug/m3 and below.
CO	Concentration of CO was in the range of 1000 to 1500 ug/m3	Concentration of CO after lockdown came down to 300- 500 ug/m3 & below.
Ozone	Concentration of O3 in the range of 40 to 50 ug/m3.	Concentration increased to 80 to 100 ug/m3. This increase is the result of atmospheric chemistry. NOx, the general term for the nitrogen oxides comprises of nitric oxide (NO) and nitrogen dioxide (NO2). NO2 produces ozone and NO depletes ozone. During the lockdowns, with little NO in the system to support the photo-stationary reactions that destroy ozone molecules, the overall ozone concentration went up.

References

[1] Susanta Mahato, Swades Pal, Krishna Gopal Ghosh, "Effect of lockdown amid COVID-19 pandemic on air quality of India", Science of the Total Environment 730- 139086 in 2020.

[2] SciJinks, "It's all about weather" [Online] Available: https: // scijinks.gov/air-quality

[3] John H. Seinfeld, "Air Pollution : Physical and Chemical Fundamentals" by McGraw Hill Book Company.

[4] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif ul Islam, Muhammad Nabeel Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities", IEEE Access.2925082, Volume XX, 2017.

[5] W.R. Ott, "Environmental Indices: Theory and Practice", Ann Arbor Science, Ann Arbor, MI (1978).

[6] Krishna Chaitanya Atmakuri, Dr. K V Prasad, " A Comparative Study on Prediction of Indian Air Quality Index Using Machine Learning Algorithms" ISSN- 2394-5125 Vol 7, Issue 13, 2020.

[7] A. Suárez Sánchez, P.J. García Nieto, P. Riesgo Fernández, J.J. del Coz Díaz, F.J. Iglesias-Rodríguez, " Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain)", Mathematical and Computer Modelling 54 (2011) 1453–1466, 2011.

[8] "Air Quality in India during COVID 19" [Online], Available: https://urbanemissions.info/blog-pieces/india-airquality-covid19

A Review on Deep Learning Approaches Used for Fruit Disease Prediction

Ms. Namrata Dattatray Deshmukh Dr. Sheetal Rathi
Department of Computer Engineering Thakur College of Engineering and Technology Mumbai, India
ndeshmukh7777@gmail.com sheetal.rathi@thakureducation.org

Abstract—***Agriculture plays a major role in the economy of developing countries and an important source of energy. Plant diseases are very compelling as they destructively effects quality as well as quantity of crops in agriculture production. Plant protection is the essential and challenging job in agriculture. Horticultural crops i.e. fruit and vegetable acquire a place of supreme as protective food. Cooperation and Farmers Welfare, Department of Agriculture has released the first Advance Estimates of 2019-20 of Area and Production of varied Horticulture Crops. The Fruits production is predicted to be lower by 2.27% in 2019-20 over 2018-19. it's mainly because of loss in production of Grapes, Banana, Mango, Citrus, Papaya and Pomegranate. In India, Agriculture is "only" ~16 % of GDP but the largest sector for employment, GDP growth rate slowing to 5% and 4.5 % in Q1 and Q2 of 2019-20 depicts an alarming situation. A survey showed that pests and diseases led to 26%-38% loss in yield production. Disease diagnosis is very crucial in early stage in order to control and cure them which will help farmers to increase the production. In recent years, Deep Learning techniques are widely studied in computer vision and has been used in agriculture field. This work aims to review various Deep Learning Approaches for prediction of fruit diseases by highlighting the contributions and challenges from recent research papers.***

Keywords—Agriculture, Prediction, Deep Learning

I. INTRODUCTION

Agriculture plays a vital role for the global economy. The sector employs around 50% workforce of India. Agriculture and its related activities have always meant a significant share in our national income. In recent years, the share of contribution of agriculture has declined gradually with the growth of other industrialized sectors in the country. In 1950-51, agriculture and related activities contributed about 59 %of the total national income. This number decreased to 40 % in 1980-81 and then to 18 % in 2008-09. But as compared to many developed countries of the world, the agriculture share in India still remains very high. For example, in U.K. and U.S.A agriculture contributes only 3 % to the national income [1].

India is one of the important fruits and vegetables producing countries in the world. Cultivation of fruit crops plays a significant role in the prosperity of any nation. It is generally considered that the standard of living of the people can be judged by per capital production and consumption of fruits. Fruits are rich source of vitamins and minerals. Fruits crops are capable of giving higher yield per unit area than other field crops. Fruit diseases are very compelling as they destructively effects quality as well as quantity of fruit crops production. Hence, fruit plant protection is very essential and challenging job in agriculture. Disease diagnosis is very crucial in early stage in order to control and cure them which will help farmers to increase the production.

This sector incorporates various challenges of agricultural production in terms of food security, environmental impact, productivity and sustainability. Smart farming is very important to face all these challenges. To address these challenges, it is necessary to understand and analyse the agriculture ecosystems which indirect constant monitoring of diverse variables. During this process, huge amount of data will be created which needs to be stored and processed in real time for several operations. The data can be comprised of images and which can be processed with various image analysis techniques to identify plant diseases, plants, etc. in different agriculture contexts [2].

Over the years, many approaches have been implemented so far to predict fruit diseases. Major type of approaches consists of Morphology, Colour Coherence Vector, K-means Clustering, SVM, Random Forest Algorithm. Morphology and Colour Coherence-based approaches suffers loss in accuracy. On the other hand, learning-based approaches pledge to provide more precision. From learning-based approaches, two best approaches are vector machines and deep learning which stand out due to high accuracy. Compared to all other techniques, Deep Learning has placed itself in the first position by delivering maximum accuracy.

Normally, symptoms of disease on fruit are observed by farmers manually. Experts may easily diagnose the symptoms or may depend on lab diagnostic test. Currently, most of the practices for fruit disease detection in India are performed by naked eye observation by an agriculture domain expert. It is not always possible to get it on time at the remote location of farmer and the consultation charges of professional experts are also very high. Hence, there is need of an automatic fruit disease detection system which can identify disease symptoms in early stages of disease.

In this review paper, deep learning-based approaches are discussed, which are deployed previously in the topics of fruit disease prediction.

II. SMART AGRICULTURE AND DEEP LEARNING

Great improvements have been achieved in recent years with massive enthusiasm adding into the Deep Learning field. Deep Learning has drawn a lot of attention in agriculture domain. One of its applications in agriculture is image recognition, which has defeated a lot of barriers that limit fast development in robotic and mechanized agro-industry and agriculture. These improvements can be seen in many aspects of agriculture, such as plant disease detection,

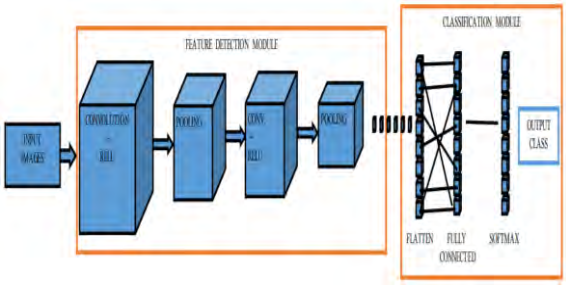


Fig. 1. Basic Deep Learning Structure

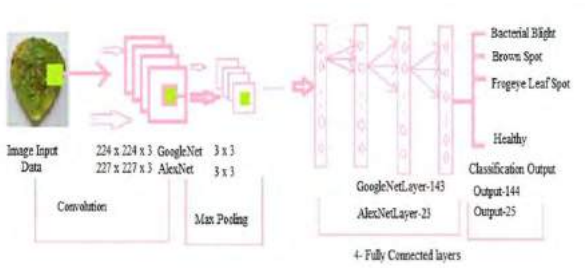


Fig. 2. AlexNet and GoogleNet CNN general architecture Proposed Jadhav et al. [5]

plant counting and weed control. An understanding of Deep Learning algorithms can relieve data analysis and thus enhance research in agriculture.

Right now, various industries are trying to incorporate artificial intelligence into their day-to-day operational needs. For example, manufacturing and automotive industries are also using AI for certain tasks and it is expected that the use of AI will increase as we move ahead. Agriculture is one of the industries, which has started applying AI in order to achieve a more effective and faster way of performing tedious tasks. There is no doubt that agriculture is one of the most important part of world's economy. The Environmental Protection Agency (EPA) claims that agriculture is responsible for about 330 billion dollars in worldwide annual profit.

The Deep Learning based drone technology is also highly beneficial for farming because of its ease to monitor, analyse and scan the crops by providing high-quality images. This technology is useful in identifying assessing their health and the progress of the crops. Farmers can determine whether the crops are ready for harvest or not, based on the images provided by this technology. Deep Learning and other machine learning technologies assist the farmers in determining the state of their soil and also used in order to determine the best time for planting and harvesting and understand how water and nutrients need to be managed. This results in higher efficiency of farming and even ROI (Return on Investment) from certain crops can be predicted taking into account their price and margin within the market [3].

Deep learning is subfield of machine learning methods based upon artificial neural networks. The types of learning can be differed as supervised, semi-supervised and unsupervised. Deep learning consists of various architectures as Convolutional Neural Networks, Deep Neural Networks, Deep Belief Network, Recurrent Neural Networks, etc. Deep learning uses a cascaded structure of layers to extract features which can be implemented for classification or pattern analysis.

The advantage of Deep Learning over all other machine learning algorithms is it can learn features on its own which lead to eliminating need to computation and manual feature extraction. Deep Learning has disadvantage of longer training time but when compared with other techniques for Learning, testing time is less compared to other methods based on machine learning. Deep Learning works best with large input datasets. Deep Learning is computationally expensive. Deep learning layers are given in Table 1. Fig.1 The basic structure of Deep Learning [4]. Deep learning architectures like and Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), Fully Convolutional Networks (FCN), Deep Neural Networks (DNN) have been studied and successfully applied to various research domains which also includes agriculture.

Table 1. Various Layers in Deep Learning Models

Layers	***Descriptions***
Image Input Layer	This layer inputs 2-D images to a network and normalize the image data.
Convolution layer	Here, convolutional filters are applied to the input.
Rectified linear unit (ReLU) layer	This layer Performs a threshold operation to each element and values less than zero is set to zero.
Leaky ReLU layer	Threshold operation is Performed where values are less than zero is multiplied by a fixed scalar.
Tanh layer	tan hyperbolic activation function is applied.
Average pooling layer	Down-sampling is performed where average values are computed by dividing input into rectangular pooling regions.
Max Pooling layer	Performs down-sampling and maximum values are computed by dividing input into rectangular pooling regions.
Fully Connected layers	This layer Multiplies the input by a weight matrix and then adds a bias vector.
SoftMax layer	SoftMax layer applies a SoftMax function.
Classification layer	Computes the cross-entropy loss for multi-class classification problems with mutually exclusive classes.

III. Literature Review

A. Identification of plant diseases using convolutional neural networks

Jadhav et al. [5] developed an efficient soybean diseases identification method based on transfer learning approach where pretrained AlexNet and GoogleNet convolutional neural networks (CNNs) were used. The

AlexNet and GoogleNet CNN architectures which are proposed by authors were trained using 649 and 550 image consists of diseased and healthy soybean images respectively. Testing is done with 80 images for both the networks. Four classes of diseases were considered for soyabean plant disease diagnosis were Frogeye leaf spot disease, bacterial blight disease, Brown spot disease and healthy. Data Collection process of soybean images was done at soybean fields in Kolhapur district, Maharashtra, India.

Implementation of the transfer learning technique is performed by using the already trained AlexNet and GoogleNet CNN models on a large data set, which could identify the accurate disease symptoms in the soybean infected leaves by using the proposed AlexNet and GoogleNet CNN models which could further assist plant pathologists in diagnosing diseases.

The last three layers of the GoogleNet model were modified in order to increase the performance of the proposed models. Some parameters of the CNNs were modified as well by setting the learning rate of the models as 0.0001 and the bias learning rate as 20 for the four fully connecter layers. The minibatch size was set to 64 ,the epochs was fixed to 30, and the number of iterations was set to 150. Fig. 2. Illustrates reposed AlexNet and GoogleNet CNN general architecture. This resulted in an overall classification accuracy of 96.25% with the GoogleNet deep neural network and 98.75% with AlexNet. Fig. 3 and Fig. 4 depicts the classification results of Alexnet CNN and GoogleNet CNN respectively.

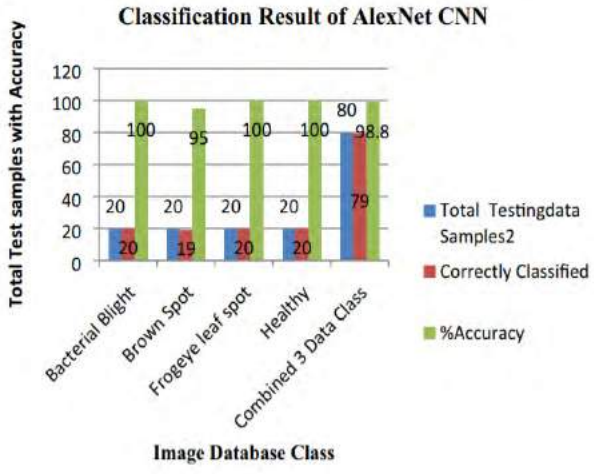


Fig. 3. Classification result of Alexnet CNN proposed by Jadhav et al. [5]

In this work, authors concluded that the accuracy of their proposed deep learning CNN model was considerably higher than the conventional recognition techniques and outperforms the machine learning model. They have achieved highest efficiency in the experiments performed by them on their proposed model for identification of soybean diseases. As future study they mentioned that there could be an attempt to increase the performance of the proposed deep learning model by changing the minibatch size, weight and bias learning rate.

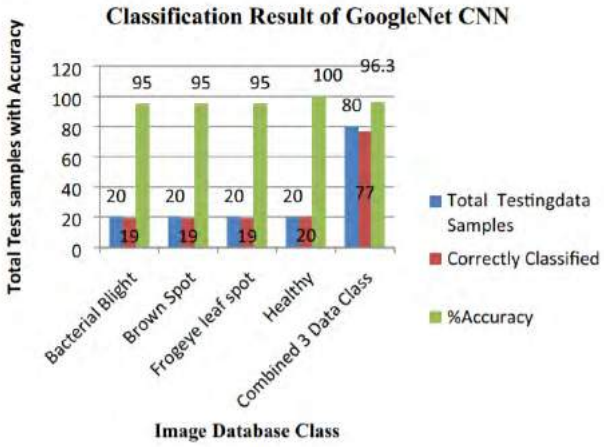


Fig. 4. Classification result of GoogleNet CNN proposed by Jadhav et al. [5]

B. Fruit and Leaves Disease Prediction Using Deep Learning Algorithm

V et al. [6] proposed a convolutional neural network to classify the diseases in leaves and fruits of apple, grapes and pomegranate. They have included image segmentation and image classification approach which can predict various types of diseases using Otsu thresholding method and convolutional neural network method.

First step was to perform Image Acquisition which includes process to upload the leaf and fruit images from the datasets. The next step was Data Pre-processing to remove noise from images. Image segmentation is performed to detect foreground objects in images with stationary background and after that disease prediction was done by implementing convolutional neural network classification algorithm. Fig. 5. Depicts the proposed model architecture.

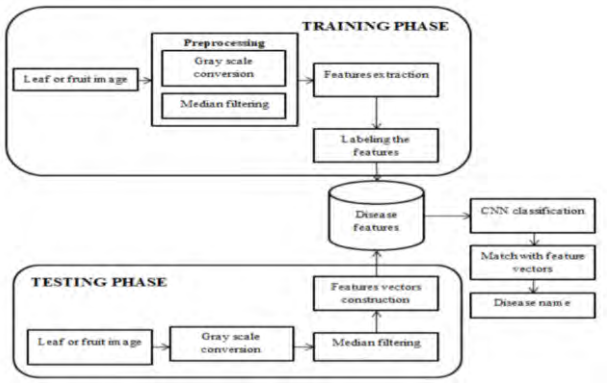


Fig. 5. Model architecture proposed by V et al. [6]

Authors overviewed various techniques and algorithms which are proposed for segmentation and classification to improve quality of image segmentation and also implemented the classification algorithm based on CNN to classify diseases of various leaves and fruits such as pomegranate, grapes and apple. They also planned to extend the framework in the future to classify diseases of various vegetables along with fruits and leaves with

improved accuracy using various deep learning algorithms.

C. Deep learning models for plant disease detection and diagnosis

Ferentinos et al. [7] implemented deep learning method for detection and diagnosis of plant disease using leaves images of healthy and diseased plants. The dataset used is an open database of 87,848 images which consists of 25 different plants with 58 distinct classes as set of combinations [plant, disease] combinations including healthy plants.

In this work, the five basic CNN architectures that were tested in the problem investigated focusing on the identification of plant diseases from leaves images, are following: (1) AlexNet (2) GoogLeNet (3) Overfeat and (4) VGG. Torch71 machine learning computational framework is used to implement, train and test these models. GPU of an NVIDIA® GTX1080 card and the CUDA® parallel programming platform on Ubuntu 16.04 LTS operating system used to implement the training algorithms.

The proposed approach showed the potential and the future scope is to make it wider in terms of various plant species and different diseases which can be identified to make the system more robust in real time cultivation conditions. An improvement can be done towards this direction by collecting variety of training data from various geographic areas, image capturing modes and cultivation conditions.

Finally, convolutional neural network was developed by authors for plant disease detection and diagnosis on leaves images of healthy and diseased plants. The proposed model obtained significantly high success rate which makes the model useful as an early warning tool or an advisory. Authors stated that the approach could be expanded as an integrated plant disease identification system to work in real time conditions.

D. Tomato crop disease classification using pre-trained deep learning algorithm

Rangarajan et al. [8] used two pre-trained models AlexNet and VGG16Net to perform disease diagnosis of tomato crop. The significance of hyperparameters namely minibatch size, weight and bias learning rate and number of images in the execution time and classification accuracy have been analysed.

Total 13,262 images were collected from PlantVillage and ImageNet dataset of 6 different diseases and healthy samples of tomato crops to train the models. The original image used for input was of dimension 256 x 256 and after applying image augmentation technique finally there were images of dimension 227 x227 and 224 x224 for AlexNet and VGG16Net respectively.

The model performance evaluated by modifying the number of images, varying the weight and bias learning rate and setting various minibatch sizes. When fine tuning of the minibatch size in AlexNet is performed it did not show a clear correlation to the accuracy of classification but accuracy of VGG16net decreased when minibatch size is increased. Classification accuracy of AlexNet obtained as 97.49% and VGG16Net as 97.23%. Fig. 6. shows proposed AlexNet architecture and Fig. 7. Shows proposed VGG16Net architecture.

The AlexNet model proposed by the authors contained several convolution layers which are followed by ReLU, maxpooling and normalization layers. ReLU activation function is a nonlinear and non-saturating which have been applied to the output of all convolution layers and last to fully connected layers as well. In the maxpooling layers, previous convolution layer output is reduced by finding and holding the maximum value in the receptive field. In the fully connected layers, there are 4096 neurons and all are connected to each other. Performance of the test phase improved by randomly avoiding the number of connections in a network.

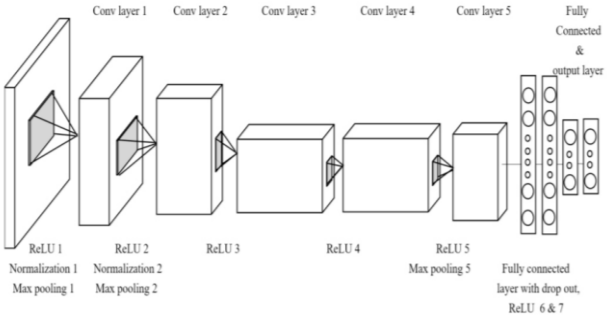


Fig. 6. AlexNet architecture proposed by Rangarajan et al. [8]

Stacked architecture of AlexNet based VGG16Net model was proposed which contains more number of convolution layers. Architecture was made up of 13 convolution layers and each layer was followed by ReLU layer. Similar to AlexNet, to reduce the dimension some of the convolution layers were followed by maxpooling. Filters used in VGG16Net model was of size 3 x 3 which were considerably smaller then AlexNet where large dimensional filters were used. Use of Smaller filters may reduce number of parameters and non-linearity was increased by adding ReLU layer after each convolution layer. This results in improvement of discrimination of each class compare to architecture which includes large filter.

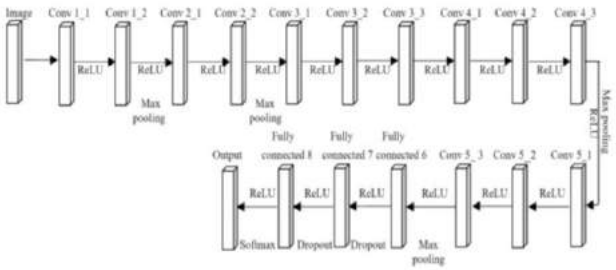


Fig. 7. VGG16Net architecture proposed by Rangarajan et al. [8]

E. Deep Learning for Tomato Diseases: Classification and Symptoms Visualization

Mohammed et al. [9] aimed to introduce deep learning as an approach for classifying plant diseases and focused on images of tomato leaves. Dataset for this work was created by extracting total 14,828 images of tomato leaves from PlantVillage dataset and divided into 9 classes of diseases.

Authors introduced deep learning as an approach to classify plant diseases using images of leaves. Authors presents two main contributions in the plant disease classification:

1) Improvements in the classification pipeline with deep learning models: According to results identified by authors, deep learning models shown good results in classification task and performed better than machine learning models. Addition to that, deep learning models can use raw data directly without feature engineering and also offers possibility of transfer learning using pre-trained models.

2) Symptoms detection on an infected leave: Localization of infected region on an infected leave helped the users by providing disease information and this biological information was extracted without any intervention of agriculture expert.

The proposed approach depicted in Fig. 8 and contains following four components (1) Pre-training phase: in this, deep architecture on a large dataset like ImageNet were trained using powerful machines with the objective of initialization of network weights for the next phase. (2) Training (fine-tuning): Resulted network from the first phase was fine-tuned and replaced the output layer of the pre-trained networks. (3) Disease classification: Here, the user can take a picture of a leaf and produced determine the disease that affects the tomato plant using the model. (4) Symptom detection and visualization: Here the user can visualize the regions of leaf image having the disease which gives the user a tool to estimate the spread of disease in the other tomato plants. Experiments are done using two pre-trained CNN models ALexNet and GoogleNet. Fine-tuning pre-trained models lead to improve the accuracy of GoogleNet from 97.711 to 99.185 and accuracy of AlexNet from 97.354 to 98.660.Summary of reviewed papers is tabulated in Table 2 at the end of the paper.

IV. Conclusion

In this paper, we have performed a survey of deep learning-based research efforts applied for fruit disease prediction in the agricultural domain. We have identified 5 latest and relevant papers and examined the particular area and research problem they focus on, techniques used, details of models proposed, sources of data used and overall performance according to performance metrics mentioned. Our findings indicate that deep learning offers better performance and outperforms other popular classification and image processing techniques. The benefits of Deep Learning are very encouraging to be used towards smarter, more secure food production and more feasible farming. For future work, we plan to apply the concepts and best practices of deep learning, as depicted through this survey, to other areas of agriculture where this modern technique has not yet been adequately used.

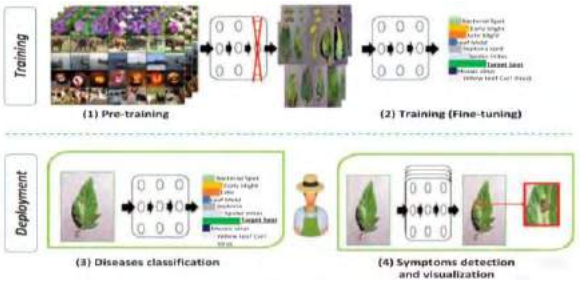


Fig 8. Deep Learning Methodology proposed by Mohammed et al. [9]

Table 2. Summary of reviewed papers

Summary of Papers					
Sr. No	***Paper***	***Deep Learning Technique***	***Targeted Crop***	***Dataset Strength***	***Accuracy***
1	Jadhav et al. [4]	Transfer learning using pretrained AlexNet and GoogleNet convolutional neural networks (CNNs).	Soyabean	649 images for AlexNet and 550 images for GoogleNet	AlexNet: 98.75 % GoogleNet: 96.25 %
2	V et al. [5]	Otsu thresholding method and convolutional neural network	apple, grapes and pomegranate.	Not Given	Not Given
3	Ferentinos et al. [6]	Pretrained CNN Architectures: AlexNet GoogLeNet, Overfeat, VGG	Generic	87,848 images	99.53 %
4	Rangarajan et al. [7]	Transfer learning using pretrained AlexNet and GoogleNet and VGGNet convolutional neural networks (CNNs).	Tomato	13,262 images	AlexNet and VGG16 net were 97.49 % and 97.23 % respectively
5	Mohammed et al. [8]	AlexNet and GoogleNet pretrained models	Tomato	14,828	99.18 %

REFERENCES

[1] https://yourstory.com/mystory/the-importance-of-the-agricultural-sector-in-india

[2] Santos, Luís & Neves Dos Santos, Filipe & Moura Oliveira, Paulo & Shinde, Pranjali. (2020). "Deep Learning Applications in Agriculture: A Short Review". 10.1007/978-3-030-35990-4_12.

[3] Magomadov, Viskhan. (2019). "Deep learning and its role in smart agriculture". Journal of Physics: Conference Series. 1399. 044109. 10.1088/1742-6596/1399/4/044109.)

[4] Moazzam, S. I., Khan, U. S., Tiwana, M. I., Iqbal, J., Qureshi, W. S., & Shah, S. I. (2019). "A Review of Application of Deep Learning for Weeds and Crops Classification in Agriculture". 2019 International Conference on Robotics and Automation in Industry (ICRAI). doi:10.1109/icrai47710.2019.8967350.

[5] Jadhav, S. B., Udupi, V. R., & Patil, S. B. (2020). "Identification of plant diseases using convolutional neural networks". International Journal of Information Technology. doi:10.1007/s41870-020-00437-5.

[6] V, S., H, R., & G, S. R. (2019). "Fruit and Leaves Disease Prediction Using Deep Learning Algorithm". International Research Journal of Multidisciplinary Technovation, 1(5), 8-16..

[7] Ferentinos, Konstantinos. (2018). "Deep learning models for plant disease detection and diagnosis". Computers and Electronics in Agriculture. 145. 311-318. 10.1016/j.compag.2018.01.009.

[8] Rangarajan, A. K., Purushothaman, R., & Ramesh, A. (2018). "Tomato crop disease classification using pre-trained deep learning algorithm". Procedia Computer Science, 133, 1040–1047. doi:10.1016/j.procs.2018.07.070.

[9] Mohammed Brahimi, Kamel Boukhalfa & Abdelouahab Moussaoui (2017) "Deep Learning for Tomato Diseases: Classification and Symptoms Visualization", Applied Artificial Intelligence, 31:4, 299-315, DOI: 10.1080/08839514.2017.1315516

Design and Development of Clustering Algorithm for Wireless Sensor Network

Pooja Ravindrakumar Sharma, Dr. Anand khandare
Department of Computer Engineering Thakur college of engineering & technology Mumbai university, Maharastra
pooja12sharma.1997@gmail.com,anand.khandare@thakureducation.org

Abstract— ***Clustering method are widely used for partitioned the data. In wireless sensor network it is very significant. The structure of cluster and how to improve it is a first challenge that faced the developers. The Wireless Sensor Network is worked of hubs from a couple to a few hundreds or even thousands, where every hub is associated with one another sensors. K-mean algorithm is one of most popular cluster algorithms that utilizing into organize sensor nodes. This algorithm is beneficial to construct the clusters for real world applications of WSN. K-mean algorithm has many drawbacks that hampering his work. In this paper we proposed a limitation of K-means and some suggestions are proposed. In light of these we can improve the exhibition of K-means which will be thought about sparing the energy for sensor hubs and therefore augment the lifetime of the wireless sensor networks. It is shown that how the modified k-mean algorithm will build the quality of clusters and mainly it focuses on the assignment of cluster which is centroid selection so as to improve the clustering performance by K-Means clustering.***

Keywords— Clustering, Algorithm, Wireless sensor network, Modified method in algorithm

I. INTRODUCTION

What is clustering?

Clustering is the task of dividing the number of data points into a number of groups. It is categorize the data points in the same groups are more similar compare to other data points in the same group and dissimilar to the data points in other groups. Clustering is organizing the data into a groups which it is formed a clusters and there is high intra-cluster similarity. [1] The aim of clustering is to find natural grouping among objects in which we can say that there is low inter-cluster similarity. There is a set of recorded historical transaction which tells various patterns on which customer bought what combination of items this can be done only through the clustering. Clustering is an unsupervised learning, by using the appropriate clustering technique, one can done a segmentation of ones customers. In data mining or machine learning person if question is asked. Then they will use the term supervised learning and unsupervised learning. [2]

II. EXAMPLE OF CLUSTERING AND CLASSIFICATION

What is the difference between clustering and classification?

Suppose that there is one basket and it is fill up of fresh fruits. Here the task is to arrange the same type of fruits at one place. So the task is how one will arrange the same type of fruits in a place. And this time one doesn't know anything about the fruits, one is seeing these fruits for the first time.

One will first take on a fruit and one will select any physical character of that particular fruit. Example color. Then will arrange them based on the color, then the groups will be something like this. Red color Group: apples & cherry fruits. Green color Group: bananas & grapes. So now one will take another physical character as size, the groups will be something like this. Red color and Big Size: apple. Red color and Small Size: cherry fruits. Green color and Big Size: bananas. Green color and Small Size: grapes. Hence work done. [3] For this situation one didn't master anything previously, which means there was no training data and no response variables available. This type of learning is known unsupervised learning. Clustering comes under unsupervised learning. [4]

III. WHY CLUSTERING?

A. Clustering helps in organizing voluminous, huge data into clusters which shows internal structure of the data. Example clustering of genes. The goal of clustering is partitioning of data. Example in Market segmentation. The data is ready to be used for other AI techniques after clustering. Example: For news Summarization wherein we group data into clusters and then find centroid. For knowledge discovery in data clustering Techniques are very useful. Example underlying rules, reoccurring patterns, topics, etc. From a few years back there is increased in the potential use of wireless sensor networks in applications such as Environmental management and various surveillance.[5]

B. Abbreviations and Acronyms

A WSN is a wireless sensor network consist of spatially distributed autonomous sensors to monitor physical or environmental conditions.[6]

These nodes are used to monitor real-time application to perform various tasks. [7]

There are many applications of WSN mainly includes health, military, environment, home and other commercial areas. [8]

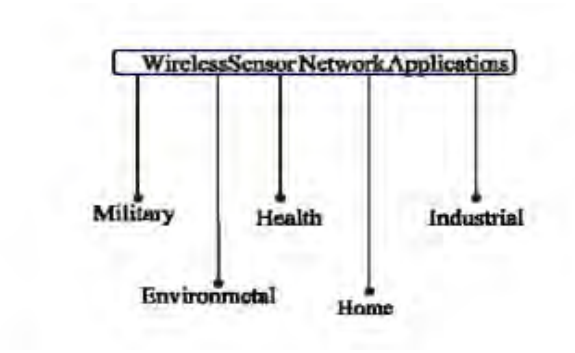


Fig.1 WSN applications

C. Existing algorithm for clustering

- There are various clustering algorithms exist like K-means, Clustering by Vertex Density in a Graph, Clustering by Ant Colony Optimization, A Dynamic Cluster Algorithm Based on L r Distances for Quantitative Data.
- Why k-mean? K-mean algorithm is given below and there are several ways to select the initial – points which represent the clusters. For the algorithm the heart is the for-loop, in which we consider each data point other than the k selected points and assign the closest cluster, where "closest" means closest to the cluster centroid. However, since only points near the cluster are likely to be assigned, the centroid tends not move too much.[9]

D. Algorithm

K-mean is most widely-applied clustering algorithm in real-world. [10]
The k-mean algorithm will categorize the items into k groups of similarity. To calculate that similarity, we will use the Euclidean distance as a measurement.

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

Where,

'c' is the number of cluster centers.

'$||x_i - v_j||$' find distance between x_i and v_j through Euclidean

Formula.
'c_i' is the number of data points in i^{th} cluster. [11]

K-mean algorithm

- From the items first we initialize k points, called means, which is selected to be randomly.
- Then we categorize each item to its closest mean. And then we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
- Repeat the process for a number of iteration and at the end we get our clusters. [12]

Modify Initialization in K-mean algorithm:

One of the widely used clustering algorithm is k-mean. However, it still has some drawbacks, and one of them is in its initialization step where it select data points and performed randomly. [13]

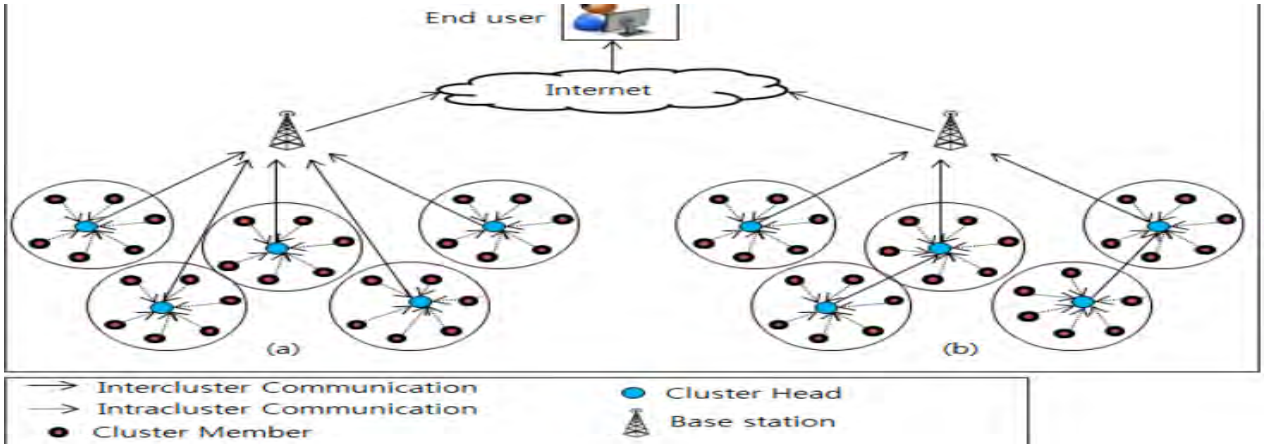
E. What is Wireless Sensor Network?

- Wireless sensor network (WSN) is consist of finite set of sensor devices geographically distributed all over the world. A WSN aims to gather the environmental data. Their Network nodes can have actual or logical communication with all devices. For communication it defines as a topology according to application.[14]
- Wireless sensor networks (WSNs) are highly resource constrained with limited power, bandwidth, processing and computational capabilities and storage. Along these lines, sensor hubs are generally inoperable and indispensable when failure occurs. The energy exhaustion of radio correspondence is straightforwardly identified with any transmission in the network. Clustering technique increases sensor network lifetime and various sensor applications.[15]
- Achieve small sensor network features in a large sensor network, various solutions have been proposed to break sensors into smaller groups. For that clustering is the one that demonstrates scalable results. Clustering provides logical organization of small units and hence it's easy to manage.
- There are many sensor application cluster the sensor nodes to achieve robustness, scalability and reduce traffic of network.[16]

IV. WSN IN CLUSTERING

Here, Clusters are provided with Cluster head. It wilL aggregate data to the base station sink. [17]

- In order to gather data more efficiently, a clustering algorithm is used for data communication in wireless sensor networks (WSNs). For that cluster analysis there is a significant procedure for ordering a "mountain" of data into sensible important documents.
- In real-time problem, it is often found that the estimation of the exact values of all the criteria is difficult and in WSN it involves grouping of sensor nodes into clusters and also in electing cluster heads (CHs) for all the clusters. The working of CHs is to collect the data from particular clusters nodes and forward the aggregated data to base station, so in order to energy efficient. Clustering is well known optimization problem. It will also calculated widely to extend lifetime of wireless sensor networks (WSNs).[18]
- Here a simple figure explain through a scenario. In Fig. 2 how clusters formed in sensors.



- A Wireless Sensor Network can be characterize as a network of wireless devices that can gather and communicate the information through the wireless links. It is a self-sorted out organization formed by an enormous number of miniature sensors that are haphazardly conveyed in monitoring regional through wireless communication.
- Clustering is one of the important techniques for prolonging the network lifetime in wireless sensor networks (WSNs). It includes gathering of sensor hubs into clusters and electing cluster heads (CHs) for all the clusters.
- Wireless sensor networks are mightily limited by energy, capacity and computing power. [19]

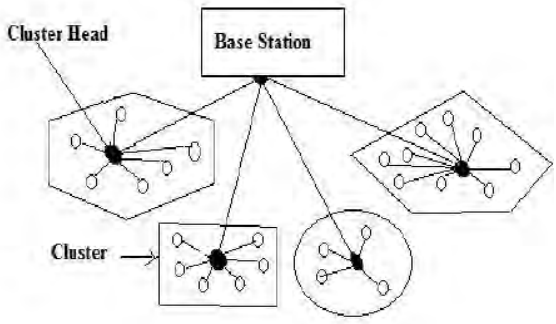


Fig.3 Cluster based mechanism of algorithm in WSN

V. RELATED WORK

- Mervat Mustafa Raouf – "Clustering in Wireless Sensor Networks (WSNs)" - 15 Jan 2019. In this paper they proposed Wireless sensor organizations (WSN) are spatially conveyed separate sensors to screen physical or ecological conditions, similar to temperature, sound, pressure, and so on just as to agreeably push through their information through the organization to a base station. The WSN is worked of hubs from a couple to a few hundreds or even thousands, where every hub is associated with one another sensors. Grouping is one of the significant strategies for expand the organization lifetime in remote sensor organizations (WSNs). It includes gathering of sensor hubs into groups and choosing bunch heads (CHs) for all the groups. CHs gather the information from specific bunches hubs and forward the amassed information to base station, so as to energy proficient grouping is notable improvement issue which has been determined generally to expand lifetime of remote sensor organizations (WSNs) and there are many types of clusters.[20]

- Dr Gayatri Devi, Srutipragyan Swain and Mr Rajeeb Sankar Bal-"The K-Means Clustering used in Wireless Sensor Network"- April 2016. This paper is based on the Sensor nodes in these applications are expected to be remotely deployed in large numbers and to operate autonomously in unattended environments. According to versatility, the nodes are often grouped into disjoint and mostly non-overlapping clusters. They propose K-mean clustering used wireless sensor network. The strategy can divide a sensor network into a few clusters.[21]

- Ezmerina Kotobelli, Elma Zanaj, Mirjeta Alinci and Edra Bumçi, Mario Banushi – "A Modified Clustering Algorithm in WSN" – 2015. Wireless Sensor Networks (WSN) as their satisfy the reason for assortment of information from a specific marvel. Their information driven conduct just as brutal limitations on energy makes WSN not quite the same as numerous different organizations known. During this work the energy the executive's issue of WSN is examined, by utilizing our proposed changed calculation. It is a grouping calculation, where hubs are sorted out in bunches and send their information to a move chosen group head. It will give enhancement for energy utilization as far as part hubs, by making bunch heads static. LEACH already has a good energy saving strategy but our modification will provide an easier approach towards efficiency.[22]

- Mrs. S. Sujatha and Mrs. A. Shanthi Sona, - "New Fast K-Means Clustering Algorithm utilizing Modified Centroid Selection Method" - February-2013 – In this paper Cluster examination the key information bunching issue might be characterized as finding bunches in information or gathering comparable articles. The objective of bunching is to discover gatherings of comparative articles dependent on a comparability metric. Notwithstanding, a closeness metric is predominantly characterized by the client to guarantee it suits his needs. Up to this point, there is still no supreme measure that consistently fit all applications. A portion of the issues related with current grouping calculations are that they don't address all the necessities enough, and need high time intricacy when managing countless measurements and enormous informational collections. K-Means is one of the calculations that tackle the notable grouping issue. The calculation orders objects to a predefined number of bunches, which is given by the client (accept k groups). The thought is to pick irregular bunch communities, one for each group. These focuses are liked to be quite far from one another. Beginning stages influence the grouping cycle and results. Here the Centroid instatement assumes a significant function in deciding the group task in successful way. Additionally, the assembly conduct of grouping depends on the underlying centroid esteems alloted. This paper centers around the task of group centroid choice in order to improve the bunching execution by K-Means bunching calculation. This paper utilizes Initial Cluster Centers Derived from Data Partitioning along the Data Axis with the Highest Variance to allocate for bunch centroid. [23]

$$d(X,Y) = \sqrt{(x1-y1)^2 + (x2-y2)^2 + \cdots + (xn-yn)^2}$$

$$d(X,D) = \min\left(d(X,Y),\ where\ Y \in D\right)$$

VI. PROPOSED METHODOLOGY

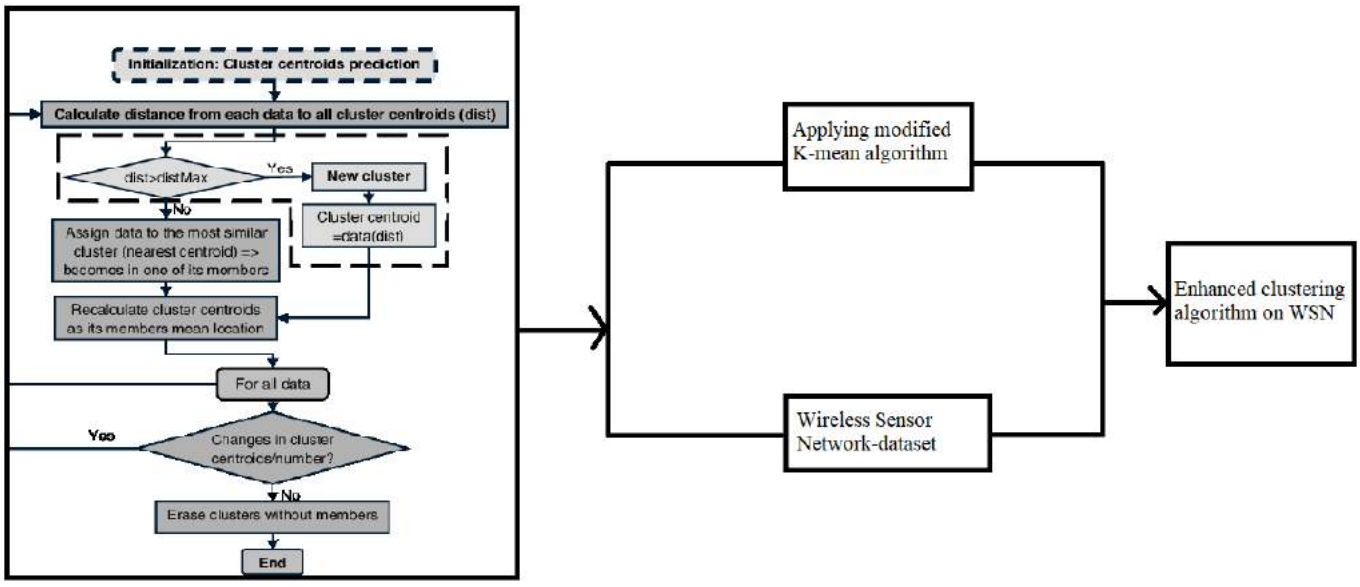


Fig.4 Modified clustering algorithm for WSN

To overcome the K-mean drawback we use K-means++. This K-mean algorithm ensures a smarter initialization of the centroids and improves the quality of the clustering. Apart from initialization, the rest of the algorithm is the same

as the standard K-means algorithm. That is K-means++ is the standard K-means algorithm coupled with a smarter initialization of the centroids.[24]

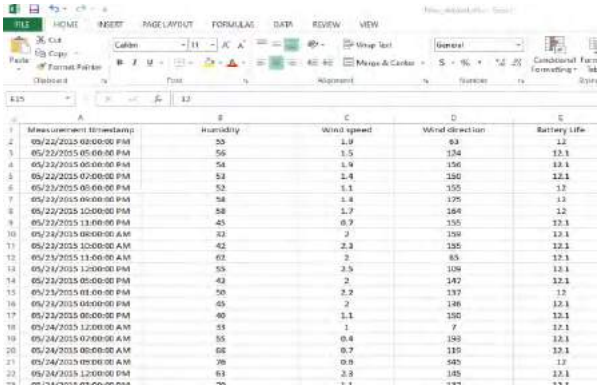
- Randomly select the first centroid from the data points.
- For every data point compute its distance from the closest, recently picked centroid.
- Select the next centroid from the data points such that the probability of choosing a point as centroid is directly proportional to its distance from the nearest, previously chosen centroid. (i.e. the point having maximum distance from

the nearest centroid is most likely to be selected next as a centroid).

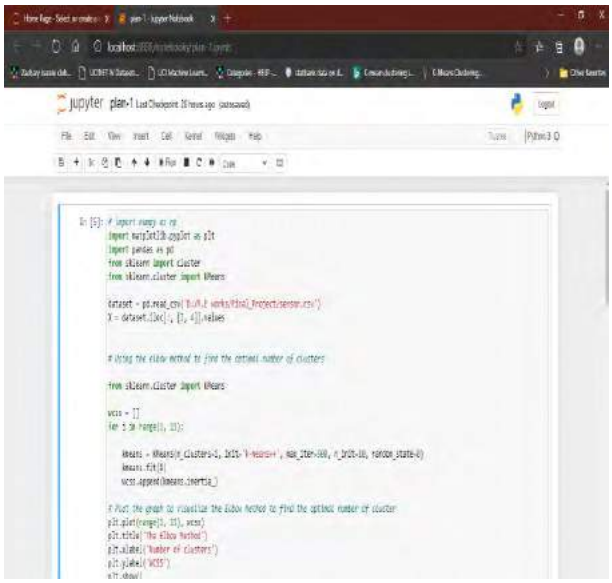
- Repeat steps 2 and 3 until k centroids have been examined.

VII. EXPERIMENTATION

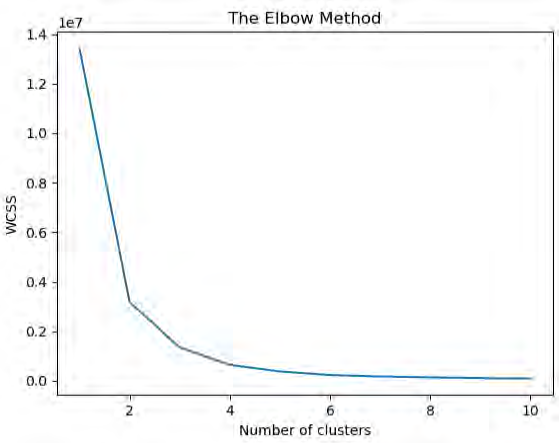
- Dataset is consist of several attributes and it is related with the environment of beach whether sensor dataset which is taken from data.world [25]



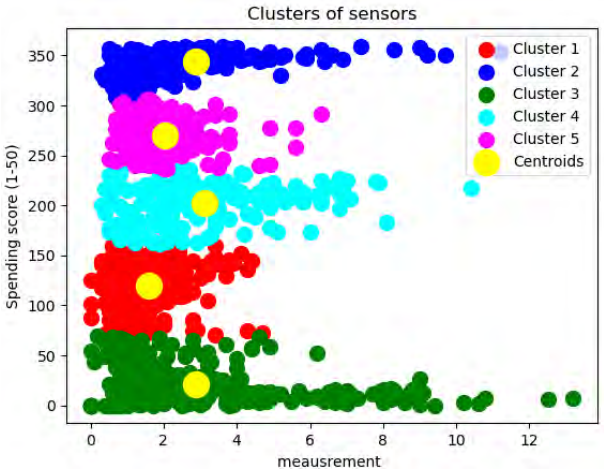
- Taking k-mean algorithm using beach whether sensor dataset.



VIII. RESULTS



Within-Cluster-Sum-of-Squares(WCSS) V/S No. of clusters



Output of clusters formed through beach whether sensor

IX. CONCLUSION

The k-means algorithm is generally utilized for clustering large sets of data. Yet, the standard calculation don't generally ensure great outcomes as the accuracy of the final clusters depend on the selection of initial centroids. Besides, the computational complexity of the standard algorithm is objectionably high owing to the need to reassign the data points a number of times, during every iteration of the loop. This presents an enhanced k-means algorithm which combines a systematic method for finding initial centroids and It is an efficient way to utilize on WSN application. There is still researching for a better result the result shown above is experimenting of different datasets.

References

1. MERVAT MUSTAFA RAOUF -" CLUSTERING IN WIRELESS SENSOR NETWORKS (WSNS) ", RESEARCHGATE PAPER, MARCH 2019.
2. Book:https://www.intechopen.com/books/wireless-sensor-networks-insights-and-innovations/modern-clustering-techniques-in-wireless-sensor-networks.
3. Ms. K.Thirupura Sundari , Ms. S.Durgadevi , Mr.S.Vairavan, "Maturity Detection of Fruits and Vegetables using K-Means Clustering Technique", IJETSR, NOV 2017.

4. "K-mean clustering introduction" - https://www.geeksforgeeks.org/k-means-clustering-introduction/.
5. http://www.ijltet.org/journal//147005795120.pdf
6. Abbreviation and applications- https://www.elprocus.com/architecture-of-wireless-sensor-network-and-applications/.
7. http://www.enggjournals.com/ijcse/doc/IJCSE16-08-04-024.pdf.
8. SIWEI WANG, MIAOMIAO LI, NING HU, EN ZHU , JINGTAO HU , XINWANG LIU , (Member, IEEE), AND JIANPING YIN , "K-Means Clustering With Incomplete Data", IEEE-2019.
9. "Why k-mean algorithm needed in present day" https://stats.stackexchange.com/questions/58855/why-do-we-use-k-means-instead-of-other-algorithms.
10. https://www.geeksforgeeks.org/ml-k-means-algorithm/.
11. "Euclidean distance"- https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm.
12. Ezmerina Kotobelli, Elma Zanaj, Mirjeta Alinci and Edra Bumçi, Mario Banushi, "A Modified Clustering Algorithm in WSN", IJACSA, 2015.
13. https://www.geeksforgeeks.org/ml-k-means-algorithm/.
14. HASSAN EL ALAMI AND ABDELLAH NAJID - "ECH: An Enhanced Clustering Hierarchy Approach to Maximize Lifetime of Wireless Sensor Networks", 2019.
15. https://www.intechopen.com/books/wireless-sensor-networks-insights-and-innovations/modern-clustering-techniques-in-wireless-sensor-networks.
16. Bangoria Bhoomi M (2014)- "Enhanced K-Means Clustering Algorithm to Reduce Time Complexity for Numeric Value", IEEE, 2014.
17. http://www.ijltet.org/journal//147005795120.pdf
18. https://ieeexplore.ieee.org/search/searchresult.jsp?queryText=clustering%20for%20incomplete%20data&highlight=true&returnType=SEARCH&matchPubs=true&pageNumber=5&ranges=2018_2020_Year&returnFacets=ALL.
19. Algorithm concept Available on: https://www.oreilly.com/library/view/data-algorithms/9781491906170/ch12.html.
20. Mervat Mustafa Raouf -" Clustering in Wireless Sensor Networks (WSNs) ", Researchgate paper, March 2019.
21. Asmita Yadav and Sandeep Kumar Singh, "An Improved K-Means Clustering Algorithm", IEEE, Nov 2016.
22. Bangoria Bhoomi M (2014)- "Enhanced K-Means Clustering Algorithm to Reduce Time Complexity for Numeric Value", IEEE, 2014.
23. Mrs. S. Sujatha and Mrs. A. Shanthi Sona, - "New Fast K-Means Clustering Algorithm using Modified Centroid Selection Method" - February- IJERT, 2013.
24. Divya Sindhu*, Surender Singh2 1M. Tech. Scholar ,"CLUSTERING ALGORITHMS: MEAN SHIFT AND K-MEANS ALGORITHM",IEEE,2014.
25. Dataset from data.world available on: https://data.world/datasets/beaches.

A Critical Review: Customer Segmentation Technique on E-Commerce

Lakshmi K. Jha
ME Scholar, Department of Computer Engineering Thakur College of Engineering and Technology,Mumbai University
lakshmijha19@gmail.com

Abstract—***Ecommerce business is no longer, a new thing. Many individuals shop with online business and many organizations use online business to promote and to sell their products. However, overloaded data appears on the customers' side. Customers today expect relevant information, personalized and offers driven by their preferences, recent interactions and latest product and support experiences. They are ready to leave their journeys with a single poor experience. Companies cannot risk to falter or even provide a below an average interaction at any step along the entire customer journey to affect their revenue. Enhance personalization is a solution to this problem. This technique is used to enhance sales by boosting potential customers. Customer segmentation will target potential customers. This paper aims to analysis customer segmentation by using data, methods and process from a customer segmentation research. The existing build systems over data segmentation algorithm shows measurable accuracy, precision, robustness and specificity, however, as the research field is as yet dynamic in attempting to improve the precision of the fundamental frameworks, this paper intends to perform a comprehensive analysis by reviewing all the methods involved and techniques used till now for detection and performance enhancement of all the models present.***

Keywords*— *Customer Segmentation, Personalization, Machine Learning, E-commerce Market Segmentation, Customer Lifetime Value.

I. INTRODUCTION

Online business gives a simple method to sell products to a large customer base. However, there is a ton of competition among multiple e-commerce sites. Customers no longer need to take an outing to physical stores to make their purchases. The J-curve digitalization has transformed the way companies operate. E-commerce companies still deal in goods and services, but now this takes place across multiple touchpoints within an online environment. The main intention of any e-commerce website is to help customers narrow down their broad ideas and enable them to finalize products they want to purchase [1]. To know your customers, segmenting them is the main alternative.

Segmenting customers will help to focus marketing efforts, so it can expand benefits and overall customer satisfaction. Customer segmentation allows e-commerce companies to explore, recognise and interact with visitors based on their browsing behaviour, customer journey, past conversations, referral page, Companies cannot risk to falter or even provide below an average interaction at any step geographical location, and much more. In turn, e-commerce companies can figure out customer needs, offering them a personalized service. Personalized services in ecommerce can maintain customer loyalty of existing customer, getting new customers by providing service to customers in accordance with their needs and characteristics. It will generate more profits for the company by creating an efficient stream of conversions. Before the personalization is executed, customer segmentation must be carried out because the outcome from customer segmentation process will be used as inputs to personalize ecommerce services, resulting in more efficient dynamic personalization ecommerce services based on customer current conditions.

Customer segmentation is presently performed by processing customer database, i.e., demographic information or purchase history. Several researchers examine customer segmentation method on their papers. Other researchers discuss the implementation section of customer segmentation. This paper will analyse customer segmentation methods based on data processing.

II. LITERATURE SURVEY

In the paper, "An Introduction to Customer Segmentation", Magneto [3] defines an approach for identifying most profitable customers to increase revenue for market. The paper also mentions that online retailers should focus their marketing strategy from average customers to best customers in order to attract as many visitors as possible to their stores. Magneto defines clear variable customer segmentation.

In 2019, "E-commerce Market Segmentation based on the antecedents of Customer Satisfaction and Customer Retention", Brian [4] studies descriptive and associative research. Effect of customer retention through customer segmentation is determine by path analysis. Demographic factors for market segmentation are identify using cluster analysis. The paper segments E-commerce market segmentation into three clusters namely functional shoppers, credibility matters shoppers and money dietary shoppers. However, the papers liquefy to have some limitations. Very few

variables are measured to fragment the customers, missing all other equally important parameters like considering psychographic factors and also if demographic factors combined with psychographic factors, the results could have led to more comprehensive customer profile segmentation in e-commerce firm.

The paper "The Research and Application of Customer Segmentation on E-Commerce Websites" proposes Customer lifetime value and its determination through RFM model. For Customer satisfaction evaluation, kano model approach has been taken which is followed by a questionnaire i.e., satisfied, should be, indifferent, tolerable and dissatisfied along with attribute with and without. The paper follows some statistical approach like probability density function using gamma and beta distribution. Xixi [5] research shows three-dimensional model combines with three variables in order to segment customers more accurately unlike one-dimensional or two-dimensional model which led to fuzzy results. Although, the research has limitation that is the paper does not focus on customer behaviour, which is a huge drawback and the research further faces problem of cold start.

The paper by Gang Fang and Xiongjian Fang [6] proposes a system CV-CL Index Determination which classifies segments into silver customer, gold customer, diamond customer as high value customer and silence customer, basic customer as low value customer which gives effective practice on e-commerce business operations for good customer segmentation.

III. CUSTOMER SEGMENTATION

Customer segmentation is the method of breaking a target market down into segments using specific variables and various modelling. As it helps identify the particular needs of each segment, discovering the best fit for a product, ultimately serving the segment better. To create a captivating product or service, one needs to know whom they are selling to. Through customer segmentation analysis, marketers and advertisers can come up with the right messages, utilising the right words, to promote their products. Meanwhile, continuous refinement of the market and its segments inform the design and plan of the product or service itself; by figuring out more about each segment and how they use the product, businesses can present some characteristics forward and drop others. Online businesses work on STP methodology: Segmentation-Targeting-Positioning [2]. Presently, customer segmentation is mould with upselling and cross-selling methodology. Upsell is the point at which you urge your customer to purchase a more expensive option of the current consideration.
Cross-sell is when you complement a customer's existing purchase with a different category product. Researchers discuss different methods for customer segmentation like Magneto [3] has used several variables like demographic, psychographic, behaviour, profit potential, past purchase and many more. Another application discusses customer segmentation model by considering customer lifetime value and customer satisfaction including RFM model and customer lifetime value determination. Some researchers have discussed customer segmentation method considering Business rule, Customer profiling, Supervised & Unsupervised clustering, Customer likeliness clustering and Purchase affinity clustering. Some methodology has similarity in between them.

A. Data Collection for Customer Segmentation

Data is categorised into external and internal data. Magneto has collected customer registration, purchase history and customer profile as internal data from ecommerce database. Media browsing, cache, cookies, market research and surveys, web and social media analysis as external data. Market research and surveys give information about customer lifestyle, activity, attitude and shopping preferences. Brian [4] has shown his research, Customer Segmentation Intelligence, by considering demographic as internal data from customer profiling and purchase history. Likewise, Xixi [5] uses customer database to increase customer lifetime value and customer satisfaction.

B. Methods for Customer Segmentation

Customer segmentation can be performed on various factors: The most basic form of segmentation is Demographics which consists of age, gender, race, ethnicity & education. Behavioural segmentation isolates customer based on the way they react and respond to ecommerce platform. Customer pattern of purchasing is determined by considering purchase occasion, sought out benefits and rate of product usage. Geographic segmentation is breaking down the market based on location, target approach [10]. Psychographic segmentation is based on customer interest and opinions. Lifestyle, social class, opinions, hobbies and interests are common factors of psychography. Browser are visitors that simply peruse a site, buyers are visitors that produce a purchase, and shoppers are customers as they buy, but want to read product reviews and feature list before proceeding for call to action.

Magneto highlighted various factors in his research:

1. Profit Potential: measuring frequent shoppers, high average order value, few returns, provide reviews & responsive customers.
2. Past Purchase: is a factor measure upon product attributes, product pricing, shipping method being used, product satisfaction & product benefits sought like price and quality.
3. Behaviour Pattern: tracking customer web-pattern page viewed, response to offers and promotions, reward-based program participation record, channel of engagement like social, mobile, online or in-store.
4. Demographics: considering physical location, age, gender, income, occupation, browsing activity through

device, source of traffic i.e. referral site, organic search or banner link.

5. Psychographics: personality traits like modern vs traditional, social vs private or spontaneous vs cautious, affiliation including political, cultural, professional, institutional, religious, recreational activities, hobbies and interests.

Magneto also have thrown some light on how to keep segmentation fresh by avoid adding consumers in more than one segment, make clear that segment is large enough to generate, and look for opportunities to tailored your services and products. Brian highlighted online business perception on customer segmentation in his research as:

1. Quantile Memberships: involves RFM which is a great practice of identifying a group of customers who needs special treatment.
2. Supervised Clustering: decision tree algorithm has been used to target their nodes that is customer attribute but it shows only one form of customer behaviour which is not sufficient in market segmentation.
3. Unsupervised Clustering: k-means clustering algorithm has been used to target any number of customer attribute which uses Euclidean distance and then volumized the similarity among them.

   Euclidean distance=$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$.

Brian has also worked upon Kano model which holds back characteristics of commodities demand and is divided into 5 categories: Attractive attribute, Reverse attribute, one dimensional attribute, Indifferent attribute & Must be attribute.

Xixi highlighted customer loyalty and satisfaction as a key role in his research:

1. Dynamic Customer Profiles: knowing how your customer evolve product usage over time, how migration take place among different products and which factors leads to change in behaviour are valuable in planning share of wallet effective strategies.
2. Smart Selling: focus on when not to target certain customers for an offer based on customer behavioural data, including service demanders, promotion maximisers, revenue reversers, & spending limiters.
3. Customer Lifetime Value: Amazon spends 85k on a single customer to generate CLV as:
   Average value of sale + Average retention time for a typical customer + Number of repeat transactions = Lifetime Customer Value.
4. Business Rule: Segmenting customers can help out in figuring which customers to target, and how to build a best communication channel with segmented customer. Online Retailers enhance their marketing strategies to appeal average customers and attracting as many customers as possible to their stores.

Xixi summarised that market segmentation criteria achieves benefits with various customer data, such as CLV, conversion rate, add to cart rate and many other ecommerce metrics of marketing.

The following table summarised methods of review papers as:

table I: **Methods of Customer Segmentation**

Paper	Method	Data	Result	Gap Finding
Magneto (2019)	Magneto	Transaction History, Demographic, Data Marketing, Data Product, Data Media, Server Log	Approach have clear variable customer segmentation	Data processing technique has not been used
Brian (2019)	Quantile Membership	Transaction History	Can handle little information, can be utilized with other information	Great outcome acquired while determining a good classification
	Purchase Affinity Clustering	Transaction History, Data Product	Identify the products that are in demand	Distinct to product segmentation
	Supervised Clustering with Decision Tree	Demographic, Transaction History	Classify customers according to target	Use only one variable to cluster
	Unsupervised Clustering	Transaction History	Any number of customer attributes is used	Speed of computation depends on k values
Xixi (2016)	Dynamic Customer Profiling	Demographic, Transaction History	Use database query if data is small	Not focus on behaviour
	Smart Selling	Customer Behavioural Data	Increase service demand and promotion is maximised.	Problems of cold start
	Customer Lifetime Value	Demographic, Transaction History, Data Product	Classify customers according to the target	Problem arises when there are different unit in record
	Business Rule	Demographic, Transaction History	Easy to apply, Use database query	Not focus on customer behaviour

Some researchers have examined various models like Gang [6] in his research, evaluated CV-CL matrix to find most potential valuable customer. Maria [7] compares three different approaches pre-filtering, post-filtering and profiling to integrate context into segment-based behaviour modelling. Balmeet [8] found out positive, negative and no correlation between customer features which, in turn can leads to profitable margin.

IV. BUSINESS ANALYTICS PROCESS OF CUSTOMER SEGMENTATION

Based on researchers and table above, customer segmentation is correlated with business objective. Business Analytics helps decision makers to improve insight about their business operations and help them make better, fact-based decisions [9]. The first step of customer segmentation is to define primary business objective. Defining scope is another important factor of business analytics. Formulating BA plan after leading from scope statement. The secondary business objective is to define detailed requirements in order to discover BA plan, supported by technical implementation. The procedure will help the business to implement the solution. As a result, we are able to fetch the access value created by the solution.

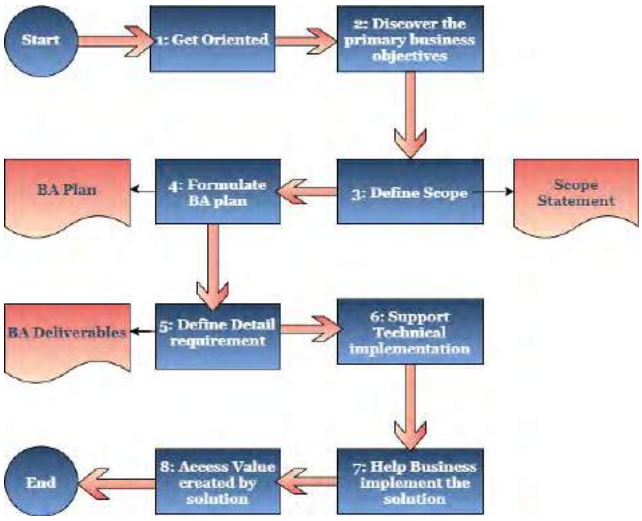


Fig. 1. Process of Business Analytics

V. RESEARCH METHODOLOGY DESIGN

The proposed research methodology comprises four primary steps. The basic phase is related to pre-analysis efforts aims to data cleaning and transformation. The main intention of research is to understand the market segmentation so that retailers have clear perspective of customer behaviour, store layout, Product complementary recommendation; purchase patterns which will lead to strive for greater loyalty by taking relevant actions for the defined segment. The second proposed research phase is defined to improve response rate of customers by performing RFM Modelling, identifying which customer to be target for cross-selling and up-selling opportunities. Since, every customer is unique, personalizing customer is next phase. The recommendation engine is proposed to fulfil customer expectation to be personalized, followed by uplifting customer modelling and predicting customer lifetime value to enhance the market strategy. Finally, to measure and monitor of result will be presented. The proposed research methodology process is presented in Fig. 2.

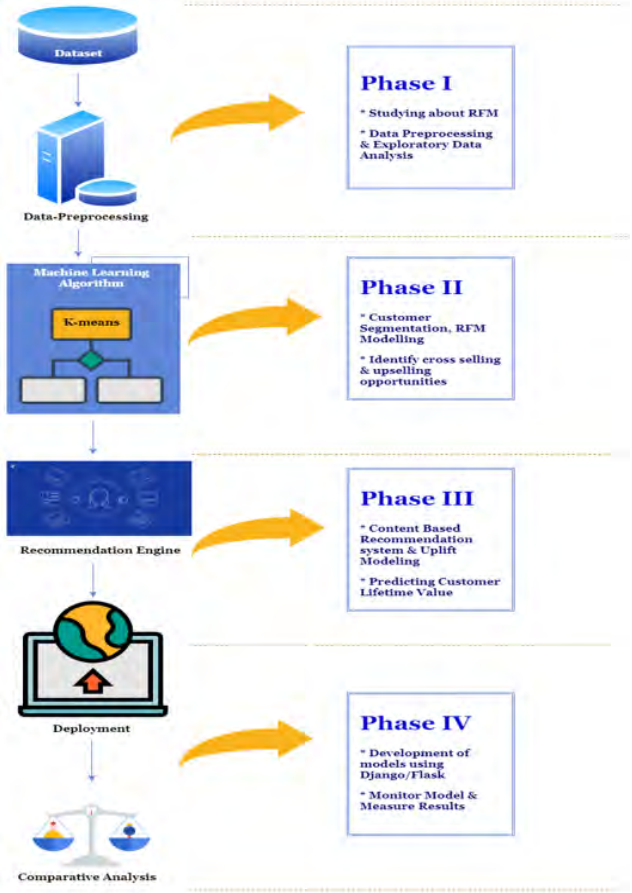


Fig. 2. Research Methodology for Customer Segmentation

VI. CONCLUSION

The paper reviews benefit of market or customer segmentation enhancing integral part of market research. If a business to be launch, looking for sustainable and profitable firm, customer segmentation must be taken into consideration. Integrating a small section of segment and tracking its performance over time. Depending upon the performance along the way, insertion or removal of segment takes place. Evolve around the segments either dynamically, or in real-time; Conversational Marketing, Website Tracking and Email Marketing offer professional result at minimal cost. Process of Customer Segmentation can niche any business.

VII. RESULTS

The expected result is recommendation based on RFM Modelling using their purchasing behaviour, Cross Selling, Loyal Customer Segment, Popular Recommendation (New Customer), Upselling Highest Rated (Big Spenders and VVIP) The proposed customer segmentation model aimed at suggesting relevant items to users. The research model is really critical in many industries as they can generate a huge amount of income when they are efficient or also be a way to stand out significantly from competitors.

ACKNOWLEDGEMENTS

I wish to record my deep sense of gratitude and profound thanks to my research supervisor Dr. R. R. Sedamkar, Professor, for his keen interest, inspiring guidance, constant encouragement with my work during all stages, to bring this research into fruition.

References

[1] Monireh Hosseini, Mostafa Shabina, "*New approach to customer segmentation based on changes in customer value*", Journal of Marketing Analytics.

[2] John Hymas, "*Online marketing: Segmentation and targeted customer strategies for the web*", Journal of Financial Services Marketing.

[3] Magneto, "*An Introduction to Customer Segmentation*", 2019.

[4] Brian Garda Muchardie, Annetta Gunawan, Billy Aditya.; "*E-Commerce market segmentation based on the antecedents of Customer Satisfaction and Customer Retention*", Institute of Electrical and Electronics Engineers, 2019

[5] Xixi He, Chen Li, "*The Research and Application of Customer Segmentation on E-Commerce Websites*", Institute of Electrical and Electronics Engineers, 2019

[6] Gang Fang, Xiongjian Liang, "*E-Commerce Marketing Strategy on the Basis of Customer Value and Customer Loyalty"*, 2017 Institute of Electrical and Electronics Engineers, ISSN: 1314-3395.

[7] Maria Francesca Faraone, Michele Gorgoglione, Cosimo Palmisano, "*Contextual Segmentation using context to improve behaviour predictive models in E-Commerce*", Institute of Electrical and Electronics Engineers, 2010

[8] Balmeet Kaur, Pankaj Kumar Sharma, "*Implementation of Customer Segmentation using Integrated Approach*", International Journal of Innovative Technology and Exploring Engineering, ISSN: 2278-3075, April 2019.

[9] Claudio Marcus, "*A practical yet meaningful approach to customer segmentation*", Journal of Consumer Marketing, June 2015.

[10] A. Vellido, P.J.G Lisboa, K. Meehan, "*Segmenting the E-Commerce Market using the Generative Topographic Mapping"*, Springer 2000.

[11] Onur Dogan, Ejder Aycin, Zeki Atil Bulut, "*Customer Segmentation by using RFM Model and Clustering Methods: A Case Study in Retail Industry"*, International Journal of Contemporary Economics and Administrative Sciences, ISSN: 1925-4423.

[12] Xiaojun Chen, Feiping Nie, Zhou Zhao,Min Yang, Yixiang Fang and Joshua Zhexue Huang, "*PurTreeClust: A Clustering Algorithm for Customer Segmentation from Massive Customer Transaction Data",* Institute of Electrical and Electronics Engineers, 2017

[13] Geo↵ Skinner, Ilung Pranata, *"Segmenting and targeting customers through clusters selection & analysis",* Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on. IEEE. 2015, pp. 303–308.

[14] Prashant R Makwana, Trupti M Kodinariya. *"Review on determining number of Cluster in K-Means Clustering",* International Journal, 2013, pp. 90–95.

[15] Minghua H*., "Customer segmentation model based on retail consumer behaviour analysis"*, Intelligent Information Technology Application Workshops, 2008, International Symposium on. IEEE. 2008, pp. 914–917.

Pneumonia Detection using Machine Learning Approach : A Case Study

Ms. Sukhada Raut Mrs. Veena Kulkarni
Thakur College of Engineering & Technology Mumbai India
sukhada795@gmail.com,veena.kulkarni@thakureducation.org

Abstract— ***Pneumonia is an infection of the lungs; it is one such kind of lungs disease wherein there is inflammation or infection in lungs caused by viruses or bacteria. In case of pneumonia the air sacs of the patient's lungs get filled with liquid substance which do not allow the lungs to function properly. Radiology is a branch of medical science where this disease is diagnosed by examining the x-ray images.***

Machine learning has been promoted as an effective way to automate the analysis and diagnosis of medical images. The commonly used method to detect Pneumonia is using chest X-ray, which requires careful examination of X-ray images by an expert. Human assisted diagnosis has its limitations like the unavailability of an expert, high cost, etc and hence there is a need of an efficiently automated system which will be invariant to many factors that can affect the radiologist's diagnosis, such as eyestrain, distraction, stress etc.

The aim of automated system is to improve the quality and productivity of radiologist's task by assisting for detection and classification of diseases more accurately. The proposed paper aims to perform a comparative study of the available methods for pneumonia detection in X- ray images and image processing.

Keywords— Machine learning Pneumonia, Classification, Machine Learning, Deep Learning, Chest X-Ray, Healthcare ;

I. INTRODUCTION

Pneumonia is one such kind of lungs disease wherein there is inflammation or infection in lungs caused by viruses or bacteria. In case of pneumonia the air sacs of the patient's lungs get filled with mucus or liquid substance which do not allow the lungs to function properly.

Diagnosis of pneumonia primarily includes examinations by a physician which will include blood tests, such as a complete blood count (CBC) to see whether patient's immune system is fighting an infection. Pulse oximetry to measure how much oxygen is in patient's blood. Pneumonia can keep patient's lungs from moving enough oxygen into your blood. This will then lead to examination of x ray to confirm whether the patient is suffering through pneumonia or not.
Pneumonia can be detected using chest x-ray image of the patient. When interpreting the x-ray, the radiologist will look for white spots in the lungs in the x-ray image that identify an infection. [1]
The main goal of an automated pneumonia classification system is to improve the quality and productivity of radiologist's task by providing a computer system for detecting and classifying diseases. Since even for a trained radiologist, it is a challenging task to examine chest X-rays. So the automated system could aid the radiologist's in their decision making process to increase the accuracy and reduce the time taken which is very important in healthcare field. The automated system can utilized to improve diagnostic accuracy, not as a means of replacing the specialist, but instead working like a second one, which is invariant to many factors that can affect the radiologist's diagnosis, such as eyestrain, distraction, stress and others. [2]

II. ARTIFICIAL INTELLIGENCE IN COMPUTER AIDED DIAGNOSTICS

In medical science, Radiology is a branch which uses imaging technology and radiation to diagnose and treat a disease. Chest X-Rays image classification in medical image analysis as well as computer-aided diagnosis for radiology is an active field of study. The main goal is to improve the quality and productivity of radiologists' task by providing a computer system for detecting and classifying diseases. [2]

A Computer-Aided Diagnosis (CAD) is an automated system which helps or assists the main radiologist. This kind of software is utilized to improve diagnostic accuracy, not as a means of replacing the specialist, but instead working like a second one, which is invariant to many factors that can affect the radiologist's diagnosis, such as eyestrain, distraction, stress and others. Currently pneumonia is detected using chest radiographs.

Machine learning has been promoted as an effective way to automate the analysis and diagnosis of medical images. Thus, ML contributes to the enhancement of CAD. Furthermore, deep learning has been investigated and proved as the most successful Machine Learning model for medical image analysis.

Machine learning is a subset of artificial intelligence that is able to learn complex relationships or patterns from empirical data and make accurate decisions. ML algorithms are mainly divided into three types, supervised learning, semi-supervised learning, and unsupervised learning. Examples of supervised learning include classification, regression, and reinforcement learning. [6]

In the context of radiology, Machine Learning provides an effective way to automate the analysis and diagnosis of medical images. Having an automated system can assist the radiologist.

There are a few Machine Learning algorithms which have been applied in medical image analysis. The most used methods are classified as deep learning. The basis of most deep learning methods is based on neural networks. A neural network consists of neurons with some activations

and parameters. Neural Network contains multiple layers, which refer to the input layer, output layer, and hidden layers (i.e., layers in between input and output). Meanwhile, the most popular architecture of deep learning in medical image analysis is Convolutional Neural Networks (CNNs). The main reason is because CNNs preserve feature relationships when filtering input images.[3]

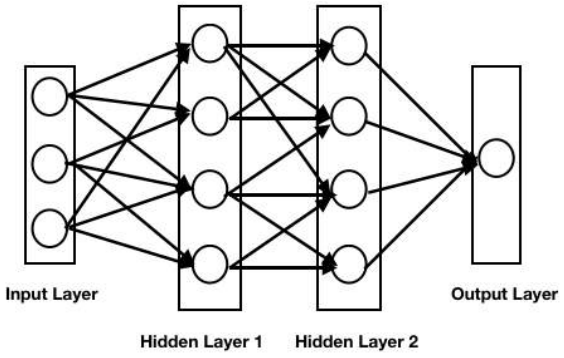


Fig 1. Architecture of Neural Network

CNNs take an input as a image and results in the assignment of class scores or probabilities. This means, the class with the highest probability is correct class for given input. The process of transforming the inputs into the chosen classes has similar layers as neural network as well as the implementation of backpropagation algorithm. [4]

III. LITERATURE REVIEW

Pneumonia is a lung disease which can be diagnosed using the chest x-ray of image. In this literature review, I studied various pneumonia detection system algorithms used for computer aided diagnostics.

Technique 1: X-ray image classification using SVM, KNN algorithms

In this proposed technique Support Vector Machine (SVM) and K-nearest neighbours (KNN) scheme is adopted to detect the pneumonia.

In this the x-ray images of both pneumonia and normal chest X-Rays of subjects is transformed into grayscale and noise removed by median filtering. After pre-processing steps, we apply wavelet transform for image decomposition so that we extract features from coefficients obtained using wavelet transform. Wavelet transform is any arbitrary function that is represented as a superposition of wavelets. These wavelets are functions generated from a mother wavelet by dilation and translation.[1]

SVM is a classification approach by which both kinds of problems, i.e., classification and regression (curve fitting) can be solved. It is a supervised machine learning technique. In this approach; data is plotted in space and dimension of space is taken equal to number of feature. We perform classification problem by finding the hyper plane that differentiate between classes very accurately. Depending upon the equation of hyper plane (kernel) SVM are classified such as linear, quadratic, cubic and fine Gaussian SVM etc.

KNN, the approach is instance-based learning, where approximation of function is calculated locally and computational process is continued till correct classification is done. It is a type of non-parametric method which is used for classification and regression problems. Depending upon number of neighbours K-NN are classified as Fine, Medium, Coarse, Cosine, Cubic and Weighted K-NN. [1]

After applying pre-processing step on chest X-Rays images, we decompose images by using wavelet transform technique. The feature matrix is formed by taking a total seven features which are mean, standard deviation, entropy, contrast, correlation, energy and homogeneity. The resultant feature matrix acts as input to a multiclass SVM and KNN classifiers to differentiate between the given classes.

Technique 2: X-ray image classification using CNN

In this technique deep convolution neural network's Convolution Neural Network is used architectures to extract features from images of chest X-ray and classify the images to detect if a person has pneumonia. CNN is a subclass of deep neural networks that has attained considerable success in computer vision domain for example, image segmentation, image classification, object detection etc. A CNN comprises of convolution layers, pooling layers and a fully-connected layer.

Currently CNN-based deep learning algorithms have become the standard choice for medical image classifications even if the CNN-based classification techniques pose similar fixated network architectures of the trial-and-error system which have been their designing principle. CNNs has an presides the DNNs by possessing a visual processing scheme that is similar to that of humans and extremely optimized structure for handling images, as well as ability to extract features through learning. A which are gradient based are employed in training CNNs models and they are less prone to diminishing gradient problem. CNN frameworks always require images of fixed sizes during training. [2]

In this technique a model is developed to detect and classify pneumonia from chest X-ray images taken from frontal views at high validation accuracy. The initial phase of algorithm includes transforming chest X-ray images into sizes smaller than the original. Then identification and classification of images by the convolutional neural network framework, which extracts features from the images and classifies them.

In this technique it is demonstrated to classify positive and negative pneumonia data from a collection of X-ray images. In this the model is built from scratch, which separates it from other methods that rely heavily on transfer learning approach. This work can be extended to detect and classify X-ray images consisting of lung cancer and pneumonia. Distinguishing X-ray images that contain lung cancer and pneumonia has been a big issue in recent times, and our next approach will tackle this problem. [3]

Technique 3: X-ray image classification using Deep Convolution Architecture

In this technique the two widely known deep convolutional architecture such as residual network and mask-RCNN in classifying and detecting pneumonia are studied.

Deep convolutional architecture has been led to a several breakthroughs for image classification and object detection since it can integrate the image information from lower and higher features. The information can be enriched with the addition of stacked layers. There are many different architectures which widely used for image classification. Faster R-CNN is used for detecting object, i.e. a particular region as the name suggests Regional Convolution Neural Network. These networks had been optimized to preside the performance of previous developed architecture.

Mask Regional CNN (mask-RCNN) is the further development of Faster Regional CNN algorithm. This algorithm usually used in object localization and recognition in an image by combining object detection and semantic segmentation [7].

The objective of object detection is to localize each object on the image using a bounding box. Meanwhile, semantic segmentation goal is to classify each pixel into a fixed set of categories using object delineation. Faster Regional CNN algorithm is involved in the object detection process. This algorithm consists of 2 stages, Region Proposal Network (RPN) that proposes region of interest (RoI) candidate. The second stage extracts the features using RoI-align and classifies the class of the object inside RoI.

The Residual network provides the output with bounding boxes if the predicted images contain pneumonia. mask-RCNN gives boundary boxes and semantic areas. Images above are examples of the testing data output from trained mask-RCNN. The prediction results consist of bounding boxes, semantic segmentation of the bounding box and the confidence level of the area. residual network shows better performance than mask-RCNN. The two networks also show the contrast gap between sensitivity and specificity which caused by unbalance dataset. In the future research, we can improve the performance of the two architectures by tuning the hyperparameters. Using more complex network structure and augmenting the unbalance dataset may also possible in the future so that we can get the best architecture for pneumonia CAD system. [4]

Technique 4: X-ray image classification using multiple machine learning algorithms

In this work we use the features and dataset employed in previous studies, which have resulted in a full CAD (Computer Aided Diagnostics) system for pneumonia detection called PneumoCAD, which has been applied to assist in radiologist, as well as to train and improve radiologists' expertise in childhood pneumonia detection using chest radiographs. PneumoCAD is currently in prototype stage.

In this study five different classifiers are applied, namely: The k-nearest neighbour classifier (kNN), which was used originally in our CAD system, Naïve Bayes probabilistic classifier, non- linear Support Vector Machine (SVM), neural network of Mult-layer Perceptron, and the decision tree.

In this three dimensionality reduction algorithms used are: Sequential Forward Selection (SFS), Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA). The five contemporary machine learning classifiers (Support Vector Machine, K-Nearest Neighbours, Naive Bayes, Multi-Layer Perceptron and Decision Tree) were tested to identify and classify radiographic images in order to detect and diagnose childhood pneumonia. The classifiers have been evaluated with a dataset taken from clinical routine. [5]

IV. DEEP LEARNING FOR PNEUMONIA DETECTION

Imaging plays an important role in the detection and diagnosis of pneumonia infected subject patient. The most available imaging is X- Ray which is more accurate, painless, more accurate and noninvasive type. The proposed work involves chest X-Ray images of normal and infected pneumonia subjects from kaggle data set which is open data set repository owned by google.

Below is the table containing various techniques and algorithms used for the classification of pneumonia using chest x-ray image previously.

Sr. No.	Title	Year of Publication	Technique Used
1	"Pneumonia Classification of Chest X- Ray Images"	2019 IEEE	Machine Learning Algorithms such as SVM and KNN are used for classification. Also, for feature extraction Discrete Wavelet Transform method is used.
2	"An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare"	2019	This study proposes a convolutional neural network model trained from scratch to classify and detect the presence of pneumonia using the chest x-ray image data given.

3	"Pneumonia Detection with Deep Convolutional Architecture"	2019, IEEE	This paper aims to know the performance of two widely known deep convolutional architecture such as residual network and mask-RCNN in classifying and detecting pneumonia.
4	"Transfer learning for image classification"	2018 IEEE	This paper is a comparative study of different image classification techniques.
5	"Evaluation of Classifiers to a Pediatric Pneumonia Computer-aided Diagnosis System"	2014 IEEE	In this paper the comparative study of various machine learning algorithms is performed for evaluation of classifier.

Image Processing:

On the available dataset, that is the chest x-ray images of normal and pneumonia, we can apply various image processing techniques. On the input x-ray image, the preprocessing techniques are applied like RGB to Gray scale conversion and used appropriate filtering techniques. While capturing the image some noise can be added to image like blur the image, some black dots on images or unwanted effects on the image. There is multiple noise during capturing the images but in X-Ray images mostly Gaussian noise and Salt and pepper noise presents.[7]

Edge detection operation reduced the pixels and then saves the Image for further processing. Edge detection is the method of identifying points where image brightness changes sharply, or blur more etc. Edge detection has two methods as gradient which uses first derivative of x-ray images and Laplacian which used second derivative of images to find edges. [8]

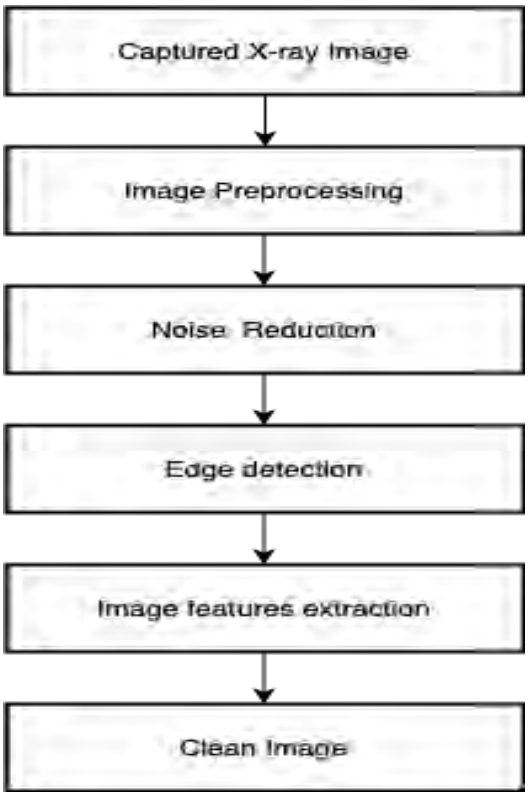


Fig 2. Process of Image Preprocessing

Applying Machine Learning Algorithms:

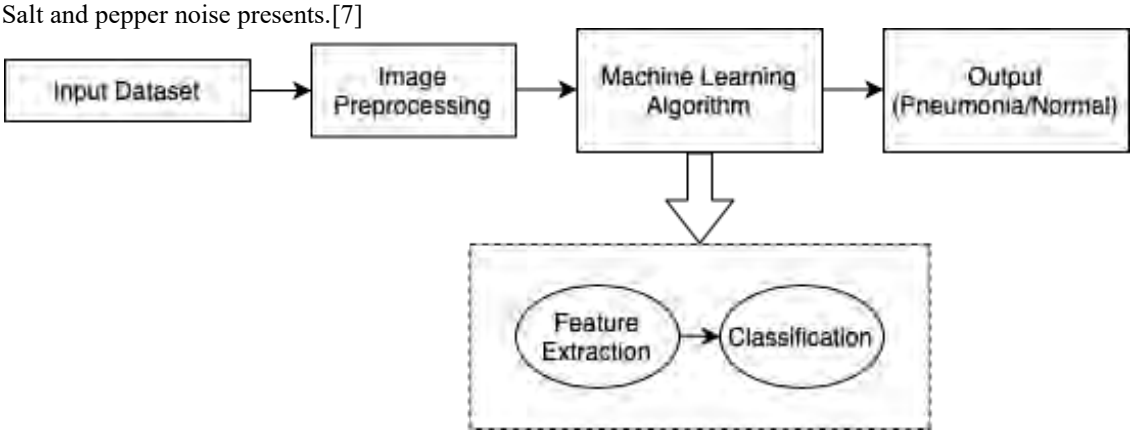


Fig 3. Process flow diagram

After preprocessing the machine learning algorithms can be applied on the clean image for further processing.

Feature extraction process captures more meaningful data, i.e., images are converted into numerical values that helps in further analysis of images. The process of feature extraction reduces the dimensionality of input data matrix and the feature matrix.

As shown in above diagram, the input dataset will be the x-ray images of normal and pneumonia, the input dataset will be preprocessed and the preprocessed image will be then given as a input to the machine learning algorithm. Feature extraction and classification will be performed by the machine learning algorithm. The output will be the classification result which will be the class whether the patient has pneumonia or it is normal.

Activation Function:

Activation functions are mathematical equations that determine the output of a neural network. The functions are attached to each neuron in the neural network, and they determines whether the neuron should be activated or not, based on whether each neuron's input is relevant for the model's prediction.

The predicted result may have high error rate and can require frequent backtracking between layers of the model to rectify the errors. For this, as a result ReLU i.e. Rectified Linear Unit can be used as activation function.[9] Rectified Linear Unit (ReLU) function is one of the mostly widely used functions in deep learning.

V. CONCLUSION

Pneumonia constitutes a significant cause of morbidity and mortality. It causes a considerable number of adult hospital admissions, and a significant number of those patients ultimately die. According to the WHO, pneumonia can be prevented with a simple intervention and early diagnosis and treatment.

As we have studied there are many machine learning algorithms which can be applied here for the detection of pneumonia in chest x-ray image, It can aid the radiologists in the decision making process; the final decision has to be made by an expert. But the computer aided system will be very helpful for time reduction in decision making process which is very effectual in healthcare sector. Usually, deep learning models are trained over thousands of images.

Training deep neural networks with limited data might lead to overfitting and restricts the models' generalization ability. This limitation can be overcome with data augmentation and transfer learning techniques.

References

[1] Harsh Sharma, Jai Sethia Jain, Priti Bansal " *Feature Extraction and Classification of Chest X-Ray Images Using CNN to Detect Pneumonia*" , 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence) , IEEE 2020

[2] Rafael T. Sousa, Gabriela T. F. Curado, "*Evaluation of Classifiers to a Childhood Pneumonia Computer-aided Diagnosis System*" , 27th International Symposium on Computer-Based Medical Systems , IEEE 2014

[3] Okeke Stephen ,Mangal Sain ,Uchenna Joseph Maduh , "*An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare*" , Journal of Healthcare Engineering, Article ID 4180949, 2019

[4] Samir S. Yadav, Shivajirao M. Jadhav, "*Deep convolutional neural network based medical image classification for disease diagnosis*" , Journal of Big Data, Springer, 2019

[5] Vanessa Rezende, Adam Santos , " *Image Processing with Convolutional Neural Networks for Classification of Plant Diseases*" , IEEE 2018.

[6] Manali Shaha, Meenakshi Pawar, "*Transfer learning for image classification*" , Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology , IEEE 2018

[7] Z. Xue, D. You, S. Candemir et al., "*Chest x-ray image view classification,"* in Proceedings of the Computer-Based Medical Systems IEEE 28th International Symposium, SaÃÉo Paulo, Brazil, June 2015.

[8] O.Stephen, M. Sain, U. J. Maduh, and Do-Un Jeong, "*An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare*", Hindawi, Journal of Healthcare Engineering, 2019.

[9] Taufik Rahmat, Azlan Ismail, And Sharifah Aliman, "Chest X-Rays Image Classification in Medical Image Analysis" , Applied Medical Informatics, Review Vol. 40, No. 3-4 /2018, pp: 63-73

[10] Sarath Pathari, Rahul U , "Automatic Detection of COVID-19 and Pneumonia from Chest Xray using Transfer Learning" , 2019.

[11] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim, "Deep Learning Applications in Medical Image Analysis" , IEEE, 2018

Movie Recommendation System through Movie Poster using Deep Learning technique

Harshali Desai, Shiwani Gupta
Thakur College of Engineering and Technology Mumbai, India
harshalidesai28@gmail.com,shiwani.gupta@thakureducation.org

Abstract— ***Movie recommendation system plays an important role for all media service providers like Netflix, Amazon Prime etc. to increase their business by providing user with appropriate list of movies from very broader lists. By helping user to choose movies of their interest, media service providers thus attract huge traffic which indeed helps them to increase revenue. Very often users choose to watch movie by just looking at its poster. So, users can immediately gain an idea about the movie from the movie posters. From movie posters user can easily know the genres of the movie. In this paper we have proposed Movie recommendation system which recommends movies based on the movie genres predicted through movie posters. Firstly, the movie posters are obtained by performing web scraping. These movie posters are then used to train the Convolution Neural Networks. Convolution Neural Networks then predicts the genres of the movie then based on predicted genres hybridization of collaborative and content-based filtering is applied to obtain recommendations.***

Keywords—Movie recommendation system, movie posters, Convolution Neural Networks, hybridization, collaborative filtering, content-based filtering

I. INTRODUCTION

There is tremendous rise in usage of recommender systems in various E-commerce sites, movie Web Services, music Web Services etc. from last few decades. Recommender system have therefore become a crucial part of daily online activities. In very simple term, recommender system aims to find or suggest relevant items to the user based on user's behavior, past history etc. Almost all application uses recommendation system so as to increase their revenue by meeting the user's expectations and giving them best user experience. Specifically, in the field of entertainment, various movie applications have millions of movies for the user to browse and watch. But for users to browse movies by their own and then select movies becomes a time-consuming activity. At this place, movie recommender systems come into the picture which helps the user to find the right movie by minimizing the options. Movie Recommendation Systems aims at suggesting relevant movies to the users.
The basic algorithms for recommendation systems are:

- Collaborative filtering: Recommends based on similar users.
  This method works by finding similar group of users from a large set of users, which is similar to the targeted user by studying their similar taste.

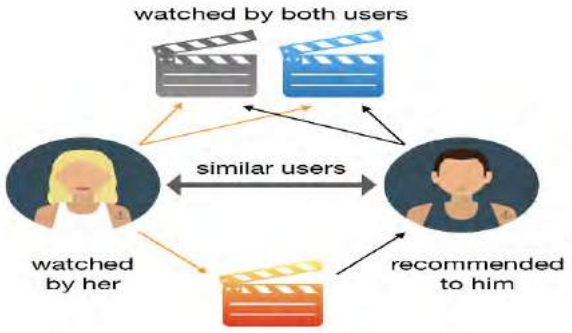


Fig. 1. Collaborative Filtering [1]

- Content-Based Filtering: Recommends based on user's history itself.
  In this method, similarity between the items of same user is found. Also, user's history is also considered for finding similar items for him.

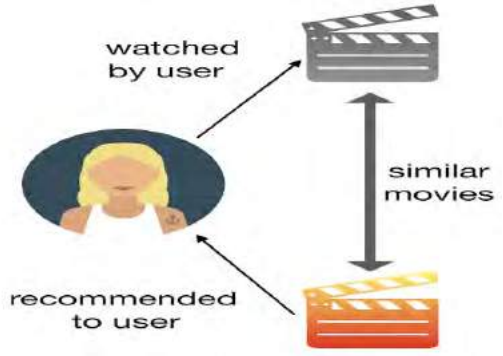


Fig. 2. Content-Based Filtering [1]

- Hybrid Method: Hybrid of above two approaches.

Apart from the above basic algorithms there are many other approaches developed from last few decades for recommending movies which we will study in literature survey section. Based on literature survey, a proposed methodology for movie recommendation system using the movie posters will be explained in third section. Third section, gives an explanation about our proposed methodology to recommend movies based on genres predicted movie posters and discuss how the proposed model will work.

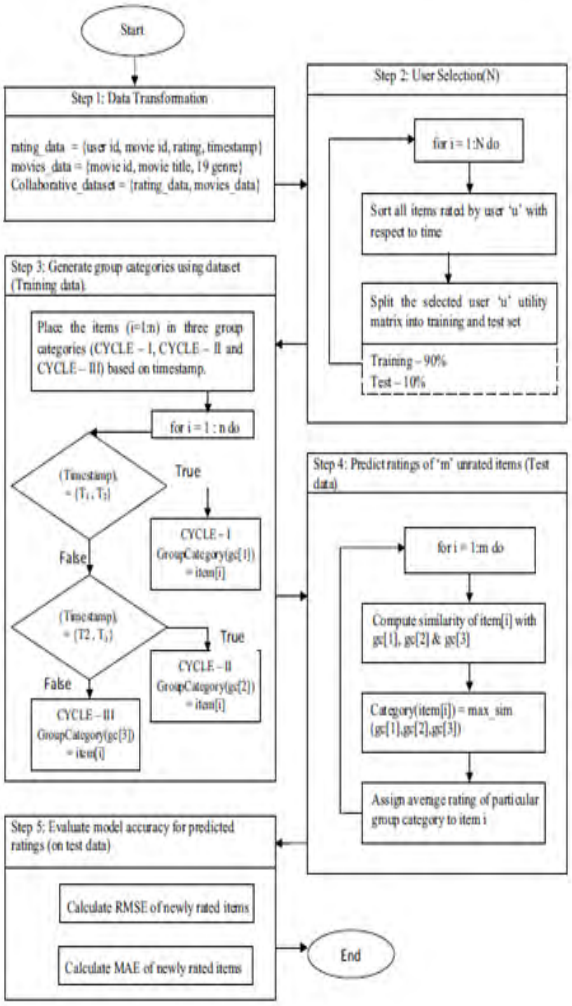
II. LITERATURE SURVEY

The authors of paper [3] have proposed a novel algorithm where Recommendation system deals with analysing the behaviour of user over the passage of time and predict the item ratings. According to authors, a lot of work is done to improve recommendation system. But still most model fails due to sparsity and also suffer from high computational complexity due to usage of heavy

deep learning and machine learning approaches. To deal with this issue, authors have proposed a novel Time Fly algorithm for building simplified recommendation system. The authors have carried this experiment on various versions of MovieLens dataset and compared their model with another algorithms. The proposed model is depicted below.

Fig. 3. Flowchart of Time Fly algorithm

In paper [4] the authors have proposed a movie rating prediction algorithm named 'MCBF-SVD', which is



based on explicit data and implicit data from movie datasets to predict the future ratings of movies from users with a certain degree of activity. In RF algorithm, they firstly, removed the users who rated less than 50 movies and their rating records, then calculates the difference value between the average rating of each user and the average rating of all users, after that deletes the users with the larger difference value. Finally, they kept the remaining users' rating data for carrying out further process. The novel Rating filtering (RF) algorithm proposed by the authors removes users with excessive rating differences to increase the MAE and RMSE. So, the main idea of the RF is to delete the users with large differences from the average rating of all users. Further, authors have proposed MCBF-SVD to alter rating according to the movie categories. Based on this, they used SVD to predict the future rating of movies from users. Their experimental results proved that the MCBF-SVD could effectively reduce errors of rating prediction models. In addition, this method can assist to increase the variety of recommended movies and alleviate the cold-start issue in theory. The authors used 2 famous movie datasets: hetrec2011, movielens-2k-v2 (an extension version of MovieLens -10M) and ml-latest.

The authors of paper [5] have used Natural Language Processing technique to generate more consistent version of Tag Genome, which is a side information that is associated with each movie in the Movie Lens 20M dataset. Also, they have proposed a 3-layer autoencoder so as to create more compact representation of tags which can improve the accuracy and computational complexity. Finally, the authors have combined the proposed representation with matrix factorization technique so as to develop a unified framework that outperforms state-of-art models by at least 2.87% RMSE and 3.36% MAE. The authors have firstly tried to reduce the total number of tags by combining the similar genome tags together. To eliminate the effect of freely user-created tags, the authors proposed to apply a mapping process: original tags which share the common context are grouped into a new tag associated with a composite score. Then a natural language processing technique named word2vec is used to cluster the same meaning tags.

So, authors have used spaCy library for implementing pre-processing step and for calculating the similarity score between two tags. They have kept a fixed threshold value of 0.65, which indicates that if the similarity score of two tags exceeds this value than the 2 tags is consider to have same meaning. The similar tags are then clustered, which gives reduced representation of tags. Finally, a composite score is assigned to the new tag.

The above step slightly improves the accuracy, but authors have further used autoencoders to discover a latent feature embedded in raw data. So, to keep reducing the dimension of genome tags and learn hidden structures they attempt to apply an autoencoder to newly created tags in the previous step. Then the proposed model is integrated with Matrix Factorization techniques, SVD and SVD++.

The authors in this paper [6] have proposed movie recommendation model based on word vector feature. The authors have used Doc2Vec model to extract the semantics, grammar and word order of the sentence, then transform it into a fixed dimension vector, then calculate the similarity of the vector and finally apply it to the collaborative filtering recommendation algorithm.

In the paper [7], authors have proposed a method for effectively recommending preferable movies for each user by using community user's movie rating information and movie metadata information with deep learning technology. A simple and effective item recommendation model is used based on Word2Vec algorithm with metadata. The proposed method uses various metadata of movie, such as movie director, actor, production year, production cost, movie tag etc. The values of these metadata are embedded as vector and are used as input and output of proposed Word2Vec network. Movie embedding is also used as input and output with meta data embedding. This input output data is obtained from user's viewing history and purchase history. The inputs are initialized with pretrained embedding using the Word2vec algorithm. Two

methods are used for obtaining pretrained metadata vector. The input embedding is generated by concatenating those of pretrained metadata embedding and movie embedding.

III. PROPOSED METHODOLOGY

In this paper we proposed a model which first aims to predict the movie genres from movie posters and then recommend movies based on the predicted genres. We will see how proposed methodology will be carried out phase wise.

Phase 1: Data Cleaning and Preparation

To train our Convolutional Neural Network we need movie posters. These posters are gathered using web scraping approach. Initially we use a dataset which is available on kaggle [8]. From this dataset we require only ImdId, genres, title column. The rest of the columns are eliminated as they are of no use. Now we need to clean our dataset so that we can achieve proper accuracy. So, we'll clean data by eliminating all those rows which contains null ImdbId and genres also we will eliminate all those rows which have empty genres. To perform web scraping process, we need to obtain the IMDB website link for various movies. For this we need to create a IMDB link. The IMDB link is created based on the ImdbId from the dataset. To create IMDB link for particular movie we concatenate IMDB website [9] home page link with the ImdbId for particular movie from the dataset. A Python Framework, BeautifulSoup [10] is used for web scraping process. Since the structure of all movie pages on IMDB website are same. So, through web scraping we can easily retrieve poster link of each movie by simply going to its IMDB page and taking the content of the 'src' HTML attribute corresponding to the poster. Once we have got all poster links, we add them to our dataset by creating a new column.

So, this web scraping process gives as the IMDB poster link from where we can now download our movie posters using these poster links. Before continuing to download posters, we drop all those records which do not have genres and ImdbId defined. Also, some images may not be found on IMDB website so we drop those entries too.

Phase 2: Exploratory Data Analysis and Data Pre processing

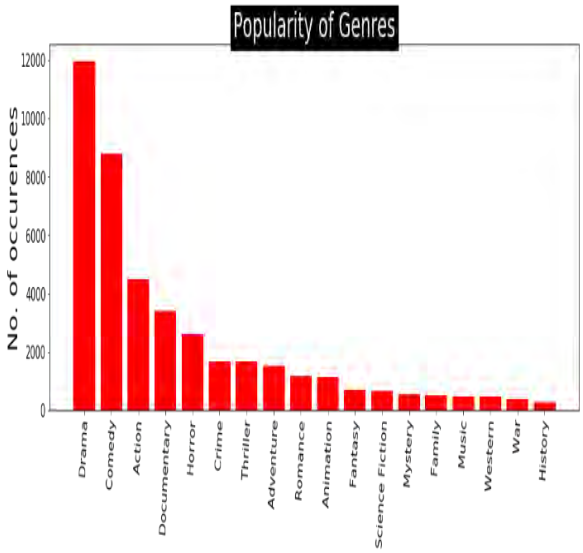


Fig. 4. Genres Visualization

From Fig. 4. we notice that the dataset obtained is heavily imbalance with respect to genres therefore, we use Oversampling technique [11], which adds more instances to less frequent genres in order to balance the dataset. This step is necessary since imbalance in the training set can negatively affect the model performances.

After this we perform image processing by reshaping the poster images so that all of the images should have the same size. As seen in the graph for popularity of genres we can notice that the genres data is heavily imbalanced data. Therefore, we perform Oversampling technique which can we further replaced with other techniques to try for more accurate results.

Phase 3a: Constructing Convolution Neural Network

Now we are building our Customized Convolutional Neural Network by using Keras framework which allow to build Deep Learning model. We will be using Sequential Model, which is the easiest way to build a model, since it allows to build a model layer by layer.
So, we'll stack sequential layers on top of each other.
Once the model is built, we train it with help of 80-20% training and validation dataset respectively. And finally, 10% dataset are used to predict the test instance and evaluate the result.

Let's first understand some terms used to build CNN model.

- Convolution Layer: They are important layer which is used to apply filter to extract features from original image.
- Pooling Layer: It is used to reduce the dimensionality. They are of 3 types- Max pooling, Min pooling and Average pooling. Max pooling will take the maximum pixel value from the part of image matrix. Min pooling will take the minimum pixel value from the part of image matrix. Average pooling takes the average pixel value from the part of image matrix.
- Dense Layer: This is also called as fully connected layer. Fully connected layers are placed before the classification output of a CNN and are used to flatten the results before classification [12]

The customized CNN model will be constructed so that we achieve maximum accuracy. We'll start building our CNN model by stacking sequential layers on top of each other. The first layer includes convolutional layer "Conv2D". Here convolutional layer uses 16, 3X3 filters which will be applied to each part of image.

The second layer includes convolutional layer "Conv2D" and "MaxPooling2D" layer. Here convolutional layer uses 64, 3X3 filter and max pooling layer is used to reduce the dimensionality of the feature maps by converting 2X2 pixel grids of the image to one pixel which is the maximum activation value in that pixel grid. Therefore, total 4 convolution layers of 3X3 pixel filter will be used and 2 Dropout layers. For the two Dropout layers threshold value will be considered

as 0.25. It means that connection whose probability is less than 0.25 will be eliminated.

The same process is been show diagrammatically below in the Fig. 5.

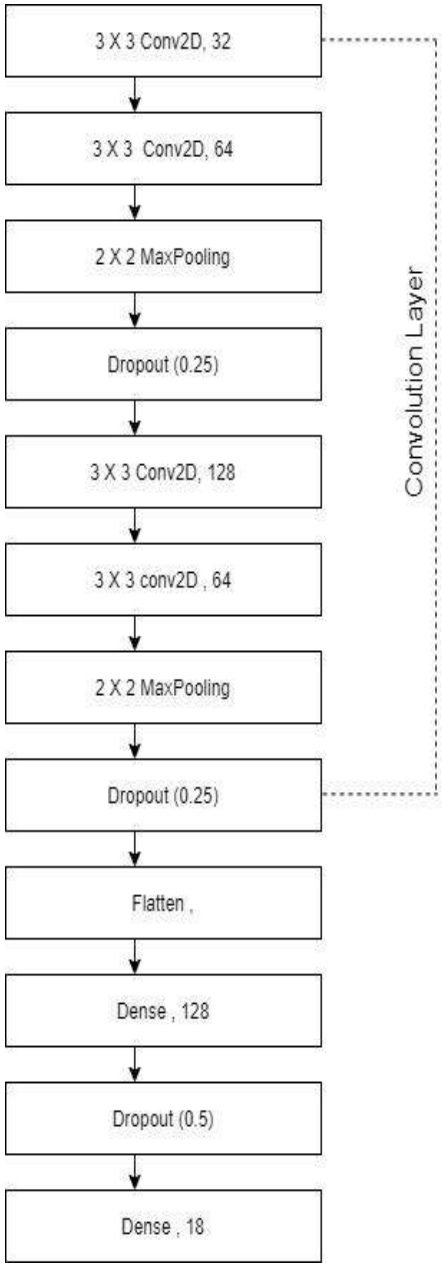


Fig. 5. Customized CNN Model

This above customized CNN model can be further customized in different ways to gain good results. We can also try by hybridizing various CNN variants like AlexNet, ResNet etc. or either taking any particular layers from this highly pretrained models and apply in our customized CNN model.

Phase 3b: Building Recommendation system using hybridization of Collaborative and Content-Based Filtering approach based on predicted genre.

Based on the previously predicted genres, Collaborative and Content-Based Filtering hybrid model will be used to obtain movie recommendation.

Phase 4: Model Evaluation

Once the model is built, we will evaluate it using three evaluation parameters- Accuracy, Recall, F1 Score. Let's see formulas for calculating above evaluation parameters. If the confusion matrix is:

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Fig. 6. Confusion Matrix [13]

- Accuracy: (TP+TN) / (TP+FP+TN+FN). In other words, we can say, ratio of correctly predicted positive and negative observation to total number of observations.
- Recall: TP/(TP+FP). In other words, we can say that, ratio of correctly predicted positive observation to correctly predicted positive and negative observations.
- F1 Score: 2 x (Recall x Precision) / (Recall + Precision). In other words, we can say that, it is the weighted average of precision and recall.

The above proposed methodology flow is depicted in below Fig. 7.

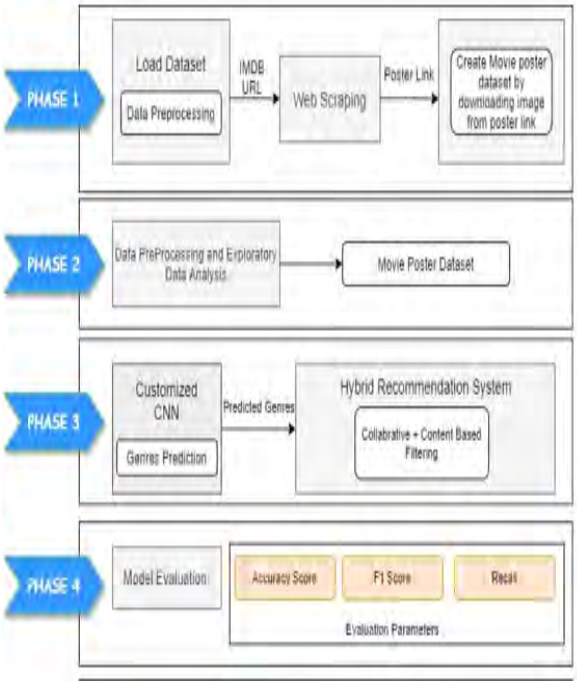


Fig. 7. Proposed Methodology

IV. CONCLUSION

Movie posters gives an idea about movie content and its genres. Based on colours, objects, expressions of actors etc. one can easily determine the genre of movie. Humans are more or less able to predict genre of a movie only by looking at its poster. Therefore, we can say that the poster contains some characteristics which can be utilized in deep learning algorithms to predict its genre.

In this paper, Deep Neural Network (Convolutional Neural Network) is built to classify a given movie poster

image into its genres. Finally, based on these predicted genres, hybrid of Collaborative and Content-Based Filtering model will be used to obtain movie recommendation.

V. FUTURE SCOPE

The movie posters are taken from IMDB websites using the IMDB id and IMDB link. But there are many movies which are very old and not present on the website. Due to this many movies were not found. The proposed model can be tried on other datasets. Also, while doing resampling, other techniques can also be used other than Over sampling. One can also use Matrix factorization techniques like SVD for recommendation model.

References

[1] https://towardsdatascience.com/how-to-build from-scratch-a-content-based-movie recommender-with-natural-language processing-25ad400eb243

[2] https://towardsdatascience.com/how-to-build-from-scratch-a-content-based-movie recommender-with-natural-language-processing-25ad400eb243

[3] Bam Bahadur Sinha, R. Dhanalakshmi, Ramchandra Regmi (2020), "TimeFly algorithm: a novel behavior-inspired movie recommendation paradigm", Pattern Analysis and Applications, https://doi.org/10.1007/s10044-020-00883-8

[4] Xiaoyue Li, Haonan Zhao, Zhuo Wang and Zhezhou Yu (2020), "Research on Movie Rating Prediction Algorithms", 2020 5th IEEE International Conference on Big Data Analytics, 978-1-7281-4111-4/20/$31.00

[5] Tan nghia duong, Tuan anh vuong, Duc minh nguyen, Quang hieu dang (2020)," Utilizing an Autoencoder-Generated Item Representation in Hybrid Recommendation System", 10.1109/ACCESS.2020.2989408

[6] Gaojun Liu, Xingyu Wu1 (2019), "Using Collaborative Filtering Algorithms Combined with Doc2Vec for Movie Recommendation", 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC 2019).

[7] Yeo Chan Yoon, Jun Woo Lee (2018), "Movie Recommendation using Metadata based Word2Vec Algorithm", International Conference on Platform Technology and Service

[8] https://www.kaggle.com/rounakbanik/the-movies-dataset?select=links.csv

[9] https://www.imdb.com/

[10] https://pypi.org/project/beautifulsoup4/

[11] https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/

[12] https://towardsdatascience.com/simple-introduction-to-convolutional-neural-networks-cdf8d3077bac#:~:text=There%20are%20three%20types%20of,task%20on%20the%20input%20data.

[13] https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

[14] https://sandipanweb.wordpress.com/2017/12/16/data-science-with-python-exploratory-analysis-with-movie-ratings-and-fraud-detection-with-credit-card-transactions/

[15] https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/

Prediction of depression using Machine Learning and NLP approach

Amrat Mali,Dr. RR Sedamkar
*Thakur College of Engineering and Technology Kandivali Mumbai, Indi*a
amrat014@gmail.com,rr.sedamkar@thakureducation.org

Abstract— ***Today, for Internet users, micro-blogging has become a popular networking forum. Millions of people exchange views in different thought of their lives. Thus, micro blogging websites are source of opinion mining data or Sentiment Analysis (SA) information. Because of the recent advent of micro blogging, there are a few research papers dedicated to this subject. In our paper, we concentrate on Twitter one of the blogging sites, to explore the opinion of the public. We will demonstrate how to collect real-time Twitter data and use algorithms such as Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BOW) and Multinomial Naive Bayes (MNB) for sentiment analysis or opinion mining purposes. We will be assessing positive , negative feelings using the algorithms selected above for the real-time twitter info. The following experimental evaluations show that the algorithms used are accurate and can be used as an application for diagnosing the depression of individuals. We worked with English in this post, but it can be used as any other languages. English in this document, but it can be used for any other language.***

Keywords— NLP(Natural language Processing) , Machine Learning , Reddit, Social networks, depression.

I. INTRODUCTION

Depression as a common mental health disorder has long been defined as a single disease with a set of diagnostic criteria. It often co-occurs with anxiety or other psychological and physical disorders, and has an impact on feelings and behaviour of the affected individuals [1]. According to the WHO report, there are 322 million people expected to suffer from depression, equal to 4.4 percent global population. Nearly in-risk individuals who living South-East Asia (27 percent) and Western Pacific region (27 percent) like China and India. In many countries depression is still under-diagnosed and left without any adequate treatment which can lead into a serious self-perception and at its worst, to suicide [2]. In addition, the social stigma surrounding depression prevents many affected individuals from seeking an appropriate professional assistance. As a result, they turn to less formal resources such as social media. With the development of Internet usage, people have s-tarted to share their experiences and challenges with mental health disorders through online forums, micro-blogs or tweets.

Their online activities inspired many researchers to introduce new forms potential health care solutions and methods for early depression detection systems. They tried to achieve higher performance improvements using various Natural Language Processing (NLP) techniques and text classification approaches. Some studies use single set features, such as bag of words (BOW) to identify depression in their posts. Some other papers compares performance of individual features with machine learning classifiers [9] – [12].

Recent studies examine the power of single features and their combinations such as N-grams + LIWC [13] or BOW+LDA and TF-IDF+LDA [14] to improve the accuracy results. With almost 326 million active users and 90 million publicly distributed tweets to a wide audience, Twitter the popular social networking sites [17]. Many researchers have used Twitter data successfully as a source of insights in the epidemiology of users tweeting emotions, depression and other mental disorders.

As an online discussion site conducted by multiple groups or "sub-reddits," Reddit social media is widely used. It is also used for discussions on stigmatic subjects because it enables the users to be totally anonymous. The posts of Reddit users who wrote about mental health discourse were studied by Choudhury. Features such as self-concern, weak linguistic style, decreased social participation, and the expression of hopelessness or anxiety predicted this change.

II. RELATED WORK

To provides insight into depression detection, there are different types of research exploring the relationship between mental wellbeing and language use. Sigmund Freud [18], dating back to the earliest years of psychology, wrote about Freudian slips or linguistic errors to expose the authors' inner thoughts and feelings. Various approaches to the relationship between depression and its language have been established through the development of sociology and psycholinguistic theories.

For example, according to the cognitive theory of depression of Aaron. [19], affected people appear to view themselves and their environment in mostly negative terms. Via derogatory words and first-person pronouns, they also express themselves. identify self-preoccupation as their typical feature, which can evolve into an intense stage of self-criticism. Other scholars have been inspired by these hypotheses to come up with empirical evidence for their validity. For instance, in three separate periods of their lives, Stirman and Pennebaker[19] compared the word use of 300 poems written 9 suicidal , 9 non-suicidal authors.

The findings indicate that first-person singular pronouns (I, me or we) were used by suicidal poets. Depressed students used more negative words and fewer positive words of feeling, according to his findings. In order predict the of depressive symptoms, Zinken etal .[20] investigated the psychological importance of syntactic structures. He assumed that in its word use, a written text may barely differ; however, it could differ in its syntactic structure, especially in the construction of relationships between events. Analyzing the roles of cause and insight. Studies on depression and other mental health problems have brought new challenges with the advent of social media, the Internet era. In order to capture user behavioural patterns, online domains such as Facebook , Twitter or Reddit have provided a new forum for groundbreaking analysis with a rich of text data and social metadata.

As an online discussion site conducted by various groups or "subreddits," Reddit social media is widely used. Since it allows the complete privacy of users, it is also used to address stigmatic subjects. Choudhury etal.[18] reviewed the posts of Reddit users who wrote about the debate on mental health and later moved to fix suicidal ideation problems. In the recent past, in a large research community, shared tasks potentially applicable to different circumstances have become significantly common.

RISK, or the Early Risk Prediction Conference and Labs Assessment Forum, is a public competition that enables researchers from various disciplines to participate and collaborate on the production of reusable benchmarks for the assessment of early risk detection technologies used in various fields, such as health and safety,

III. Problem Statement

Patients of mental problems such as Alzheimer's, depression, anxiety and neurodegenerative diseases. The most depressed place in the world is India. Using Natural, after taking input as text data, results will be extracted from the dataset. Language processing algorithm and classification algorithm algorithm to find the data as depression data and prediction or not as a suicidal post.

IV. Proposed System

A. Data Pre-Processing

We will take our first look at it now we have received our data, looking for missing values and selecting which sections of the data set will be useful for our classifier. We're also going to start pre-processing text information with natural language tools. Some exploratory data analysis, visualisations end with this portion.

Before we move to the feature selection and training level, we use the NLP tools to pre-process the dataset. To divide the posts into individual tokens, we use tokenization first. Next, we remove all URLs, punctuations, and stop words that might lead to erroneous results if ignored. Then, we apply stemming to reduce the words to their root form and group related words together.

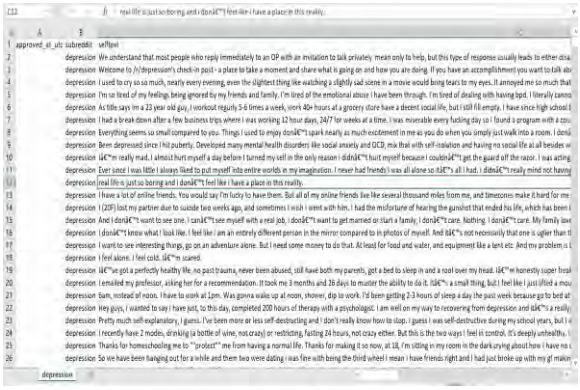


Fig. 1.Dataset

B. Extracting functionality

After data pre-processing, we feed our models with the characteristics which reflect the language habits of users in Reddit forums. To explore the linguistic use of the users in the blogs, we use the LIWC dictionary, LDA topics, and N-gram features. To encode words to be used by different classifiers, these methods of text encoding are used. N-gram modelling is used to analyse the characteristics of the documents. In text mining, NLP, it is widely used to calculate the probability co-occurrence of each input sentence as a unigram and bigram as a function for depression detection[9],[40]. As a numerical statistic for n-gram modelling, we use the Term Frequency-Inverse Document Frequency (TF-IDF), where the value of a word is highlighted with respect to each corporate document.

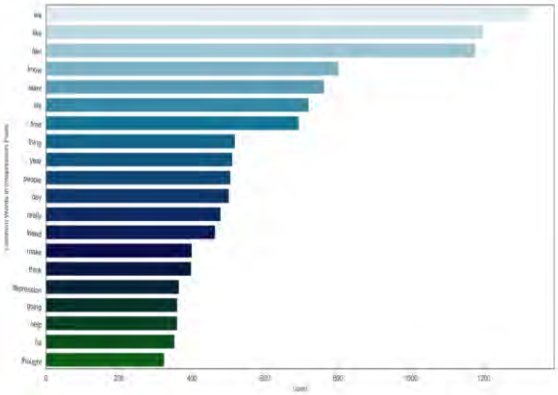


Fig. 2. Bar Plot for Top Words

V. Proposed Methodology Architecture

In computational linguistics, topic modelling is an important method to reduce the input of textual data feature space to a fixed number of topics [20]. Hidden topics such as subjects connected with anxiety and depression can be extracted from the selected documents via the unsupervised text mining approach. It is not generated by a predetermined collection of pre-established terms in contrast to LIWC. However, it produces a category of non-labelled terms automatically.

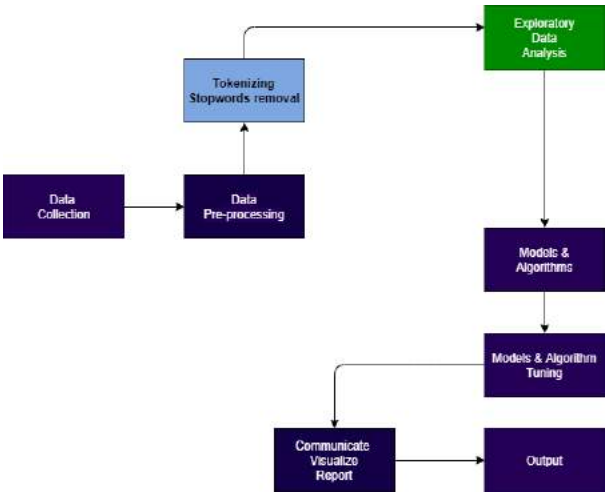


Fig. 3. Block Diagram of Proposed System

We use classifying methods to quantify the probability of depression among the users in order to quantify the presence of depression. Using Logistic Regression and Support Vector Machine, Random Forest, Adaptive Boosting and Multilayer Perceptron classifier, the proposed structure is built..

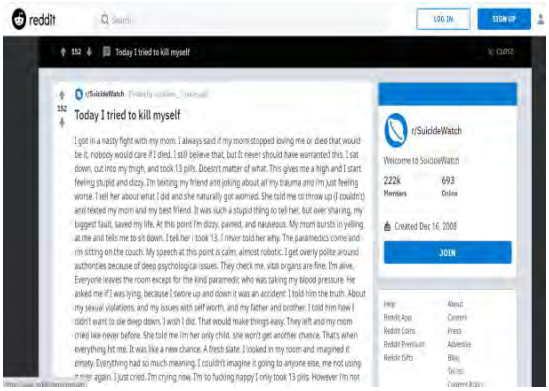


Fig. 4. Live Blog Data

Fig. 5. Full blog Data

Adaptive Boosting (AdaBoost) is an ensemble technique that can make one strong classifier combine several weak classifiers[56]. It is commonly used for problems of binary classification, A special case of the artificial neural network, Multilayer Perceptron (MLP) is mostly used for modelling complex relationships between the input, output layers[58]. It is able to discern data that is not only non-linearly separable due to its many layers and non-linear activation[59]. In our analysis, we used the MLP method ,two hidden layers with 4 and 16 perceptrons to correct all the characteristics in order to ensure accuracy of the comparison.

Since depression also affects psychomotor functions[60], we can find terms that represent the symptoms of low energy, exhaustion or inverse insomnia and hyperactivity (tired, I 'm tired or sleepy). It is also articulated somatically (my brain, discomfort, hurt) via the symptoms of the body. Unigrams and bigrams in regular posts, unlike depression-indicative posts, contain the terms identifying the events that happened quite in the past (time, month ago, year ago, last year).

Fig. 6. Most top Words used

To evaluate the connexion between the textual data and the features themselves, we selected 68 out of 95 characteristics. In view of psycholinguistic characteristics resulting in association provided in the features extraction, we transformed every depressive and non-depressive post into numerical values. The Psychological Mechanisms (0.19) followed by the Linguistic Aspects (0.17) and Personal Concerns (0.16) show the greatest correlation. c)GSM Module Working.

The findings indicate depressed individuals used more self-oriented references with respect to the mental concentration of depressed and non-depressed users and prefer to shift their attention to themselves (I, me, and mine) (0.17). The work of [25], [26], [38] is confirmed by the findings. With strongest focused on the present and future, their posts contain more negative feelings, depression and anxiety. Based on our results, when applied to the design tools, LIWC may play an effective role in data detection models.

We developed a topic model to quantify the hidden topics extracted from the posts, which acts as a depression triggering point. LDA requires that the number of topics produced be specified. Any parameter change can trigger a change in the accuracy of the classification. For this purpose, an acceptable value needs to be identified.

VI. CONCLUSION

Basically, Naïve Bayes and KNN classification will be used to final classify the text used. And finally achieving

the outcome was suicidal note or not as the post or email. After model prediction deployment, it is possible to allow users to input text and receive predictions about their mental state.

References

[1] D. Greene and P. Cunningham. "Practical Solutions to the Problem Diagonal in Kernel Document Clustering", Proc. ICML 2006.

[2] J. Grimmer and B. M. Stewart, “Text as Data: The Promise ,Pitfalls of Automatic Content Analysis Methods for Political Texts,” Political Analysis, January 2013, pp. 1–31, doi:10.1093/pan/mps028.

[3] F.“Machine Learning in Automated Text Categorization,” ACM Computing, vol. 34, No. 1, March 2002, pp. 1-47.

[4] J. Grimmer and B. M. Stewart, “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts,” Political Analysis, January 2013, pp. 1–31, doi:10.1093/pan/mps028.

[5] G. D. Guo, H. Wang, D. Bell, Y. X. Bi and K.Greer, “Using kNN model for automatic text categorization," Soft Computing, 10(5), pp. 423- 430, 2006.

[6] A. S. Patil, B.V. Pawar, “Automated Classification Websites using Naive Algorithm,” Proceedings of the International of Engineers and Computer Scientists, Hong Kong 2012, Vol. I, 14-16.

[7] [9]L. Jiang, C. Li, S. Wang, and L. Zhang, “Deep feature weighting for naïve Bayes and its application to text classification,”Engineering Applications of Artificial Intelligence, vol. 52, June 2016, pp. 26-39 H . Lasisi and A A. Ajisafe, “Development of stripe biometric based Fingerprint Authentications Systems in Automated Teller Machines,” 2012, IEEE, ISBN. 978-1 -4673-2488-5, pp. 1 72- l75.

[8] Hamid Haqani, Mir Saleem, Shoaib Amin Banday, Ab RoufKhan, “Biometric verified Access Control of Critical Data on a Cloud,” International Conference on Communication and Signal Processing, April 3-5, 2014,India.

[9] Q. Yuan, G. Cong, and N. M. Thalmann, “Enhancing Naive Bayes with Various Smoothing Methods for Short Text Classification,” WWW 2012 Companion, April 16–20, 2012, Lyon, France, ACM 978- 1-4503-1230-1/12/04.

[10] V. Lertnattee and T. Theeramunkongt, “ Analysis of Inverse Class Frequency in Centroid-based Text Classification,” International Symposium on Communication and Information Technologies 2004 (ISCIT 2014), pp. 1171-1176, Sapporo, Japan, October 26-29, 2004. 2014 | London,UK.

[11] David M W Powers, “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation,” School of Informatics and Engineering, Flinders University of South Australia, Technical Report SIE-07-001, December 2007.

[12] Abdul Razaque, Fathi H. Amsaad, Chaitanya Kumar Nerella, Musbah Abdulgader, Harsha Saranu, “Multi-Biometric System Using Fuzzy Vault”, 978-1-4673-9985-2/16/$31.00©2016 IEEE

[13] Debanjan Sadhya, Sanjay Kumar Singh, Bodhi Chakraborty,“Review of key-binding-based biometric data protection schemes”, IET Biom.© The Institution of Engineering and Technology 2016

[14] JinXin Xu “An Online Biometric Identification System on Two Dimensional Fisher Linear Discriminant”, 978-1-4673- 9098-9/15/$31.00 ©2015 IEEE.

[15] Sabout Nagaraju, Latha Parthiban “Trusted framework online banking in using multi-factor authentication and privacy protection gateway”,Nagaraju and Parthiban Journal of Cloud Computing: Advances, Systems and Applications (2015)4:22.

Real Time Driver Tracking and Attendance Management System with validation using Face Recognition

Gayatri Supatkar,Pooja Shiv,Vidya Raut,Snehal Warade
Department of Computer Engineering Cummins college of Engineering for women Nagpur, India
supatkargayatri@gmail.com,poojashiv2599@gmail.com,vidya.raut@cumminscollege.edu.in,snehalwarade@gmail.com

Abstract--***This paper presents Real Time Driver Tracking and Attendance Management System. Which is a GPS based tracking system. It has many applications in real world like vehicle tracking, person tracking, product tracking, etc. Driver tracking system is implemented to track bus drivers and manage their attendance for in between stations in the route by using Global Positioning System and basic face recognition algorithm for validating the bus driver. This system as a mobile application provides facility to the driver to send the live location to the station admin and to the admin to track the driver location as per the bus route specified. This paper presents the development of the driver tracking system with GUI application for sharing the actual location of the driver.***

Keywords—Global Positioning System, Android Application, Face Recognition.

I. INTRODUCTION

Now a days, attendance monitoring and tracking driver's location is very essential for almost every transport organization. Typically, attendance can be managed by two ways, manually and automated. Manual process requires a separate man effort to take attendance and extra time to do that, also the management of data will lead to wastage of papers or sheets. On the other hand, automated tracking and attendance management system uses bar-code badges, finger print or biometric scanners. The provided information through these devices will be received by computer system for processing. Using an automated system for attendance monitoring reduces the errors of manual system. But these automated systems require heterogeneous devices got to be located within the organization which is expensive. In this paper, we introduce smart-phone-based attendance and Driver location tracking system.

I. LITERATURE SURVEY

There are many techniques and methods carried out till now for monitoring attendance and for tracking location. we have gone through some of the papers based which have followed the tracking approach. [4]in this paper they used wireless technology for monitoring attendance and for that they used iris recognition method.

Automated attendance and location tracking system is easiest and fastest way to monitoring employee's data. [2]in this paper they described how you can retrieve data in real time i.e. in every second/minute/hour. when your smartphone moves from one location to another location, data will be varying according to it. [5] This paper explains the use of Global Positioning System (GPS) and Global System for Mobile Communication (GSM) for vehicle tracking and monitoring purpose using SIM800 module. In our work, we addressed the problem utilizing smartphones internet connectivity. [3]in this paper they used GPS for tracking employee's attendance and registration. we can easily detect others location. It describes the mobile application which is GPS enabled that tracks staff location and manage attendance using latitude, longitude and IMSI number.

[10] In this paper, authors described about the development of vehicle tracking systems hardware prototype and GUI application for displaying the actual position.

A. PROJECT SURVEY

We visited at the local bus stop on March 19,2020 at 4:00 PM, from where we tried to gather information about the driver's daily routine and the overall manual process carried out by the respective depo manager to manage the data of the bus drivers. We have also discussed about the problem faced by the drivers while travelling from one bus stop to next with regarding to submit attendance. We came to know that the driver needs to get out of the bus at every but stop and visit respective bus stop office and enter the attendance in the register manually, which not only need large amount of papers or registers and also the time.

II.PROPOSED SYSTEM

We have proposed a location based smart attendance and location tracking system implemented as an Android mobile application that communicates with the remote server during which the database is found. Internet connectivity (Wi-Fi/4G) is required for connecting to database residing within the remote server. Our proposed system doesn't require any peripheral aside from smartphone which can reduce computational time and price of placing an additional device. Any driver reached to the bus station with a smartphone and running application will be tracked and one can submit the attendance through application developed.

The system is designed for testing of Driver's attendance which will help state transportation team for validating of the

Fig 1: Work Flow (Admin Interface)

Driver. This application will monitor driver's attendance and location. One of the measures to validate the driver is mobile number of driver and main is current location (GPS) of smart phone. We have implemented application for state transport in which they can track driver current location and note driver's attendance.
An application is divided into two interfaces, i.e. driver and admin. By proposing this system, we have reduced the wastage of paper and also tried to reduce the time and efforts needed to submit and manage attendance.

III. METHODOLOGY

We have gone through lots of research paper, project surveys and finally conclude with solution and also implemented it. Now a days, everyone has smart phone so we decided to make application in which driver has to give their attendance. We have provided the unique id and password to the driver after registration. By using that unique id and password drivers has to login and then they have to share their current location for that they have to on their GPS. After that driver has to select their respective route and after selecting route they have to select respective location in which they have to submit their attendance. Application process the data provided by driver and match it with database.
The whole system has been divided into two major categories:
1) App for mobile
2) Accessing location through
Map

A. SOFTWARE REQUIREMENT

Software requirement consist of database, application program and server.

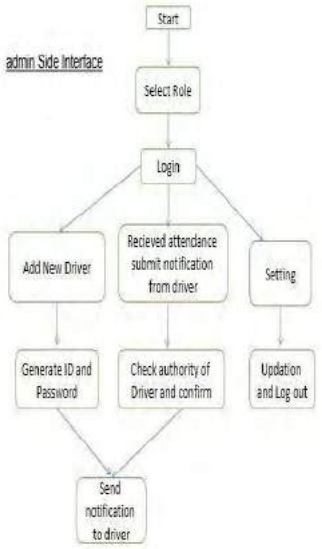
Database: we used MySQL and PostgreSQL for storing data which is easy, fastest and can store the large number of records in real time.

Application program: Application program is developed with programming language using IONIC framework and AngularJS. Application program provide user friendly
interface to Driver and admin

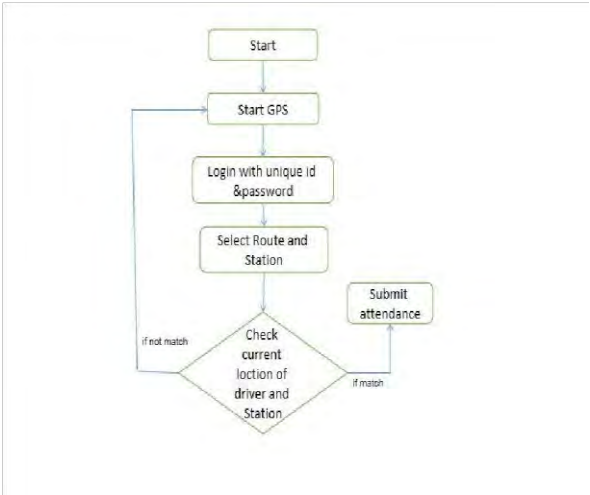
B. HARDWARE REQUIREMENT

The basic requirement of the location attendance tracking system is an android device, which will run the application. The other requirement is a personal computer on the server side, which will store the database.

User authentication is one of the major factors. Every driver is authenticated based on his unique user identification number. This unique identification number is the number which is given by the admin. The identification number and password along with other information is also saved in the



employee device.
Drivers needs to install the APK of the application in their smart phones. Admin/Driver first has to login/Register himself. After that he has to select his respective role(driver/admin).



Then there will be three options Add new driver, Manage notification and settings. In add new driver tab, admin can register new driver. After, registration the unique id and password will be provided to the drivers by admin. The notifications page will display the notifications regarding the submission of attendance and using setting tab, admin can edit his profile. Admin will be notified about the drivers location in the form of google map by displaying the live location of the driver.

Fig 2: Work Flow (Driver Interface)

In driver Interface, driver will log in using the User id and password provided by the admin. Then he has to select his respective route. After selecting route, he has to select his station location. Then he has to click on

submit attendance button. After clicking on this button, the live location of the driver will be fetched using GPS. If it is matched with the station location then only the l current location of the driver will be submitted to the admin with driver's personal details for confirmation identity otherwise same process will happen. After submission, the attendance notification will send to the admin. Admin will check and confirm the attendance.

Abbreviation and Acronym

1) GPS:

Global Positioning System (GPS) is a satellite-based navigation system owned and operated by the United States government. The service is available globally and is free to anyone with a GPS receiver.

The United States Department of Défense (USDOD) originally put the satellites into orbit for military use, but later they were made available for civilian use. GPS is a network of orbiting satellites that send precise details of their position in space back to earth. The GPS receivers use this information and trilateration to calculate a user's exact location.

The GPS receiver measures the distance to each satellite by the amount of time it takes to receive a transmitted signal from each satellite.

2) IMEI:

The IMEI number or in other words International Mobile Equipment Identity is a unique 15-digit code that precisely identifies the device with the SIM card input. The first 14 digits are defined by GSM Association organization. The last digit is generated by an algorithm named Luhn formula and it has a control character.

C. METHOD OF LOCATION TRACKING:

GPS coordinates are required for the program to instantly determine the driver's current location, based on the coordinates received. Using GPS, we will obtain both x and y-coordinates up to six decimal points with the assistance of ground and space satellites. To interpret the coordinates, the program must be integrated with Google Maps APIs so that users can view the visual location of the coordinates receive. A credential is required to use the Google Maps APIs service, which can be obtained by placing a request through Google console.

Fig 3: Sample Google Map for station location

D. DRIVER VALIDATION APPROACH

The application needs to be installed on an android device with an active internet connection, GPS and a camera. This program can only recognize one face at a time. The primary objective of the program is to be able to take attendance using mobile devices with accuracy. For the project to succeed, it must employ a location tracker and face recognition to submit proper attendance system. The face recognition requires the driver to have direct interaction with the device, while the GPS locator specifies the device's location. To summarize, this program requires the following core functions:

- A driver's current location
- Face recognition of driver

Facial Recognition Technique

A camera is needed to use facial recognition. Before deploying the application to users, it must be initialized with the required data set images, which will be processed at the start of the program. Thus, every Driver and admin would use the same application, except it will have been initialized with different images.

Once the program has been loaded with the driver images, it will be able to recognize driver's faces by using an appropriate algorithm to compare the current frame image with the one that has been initialized. Initializing the images in the program before generating the installer provides much greater reliability because drivers cannot easily alter the initialized images.

Location Detection

GPS coordinates are required for the program to instantly determine the driver's current location, based on the coordinates received. Using GPS, we will obtain both x- and y-coordinates up to six decimal points with the assistance of ground and space satellites.

To interpret the coordinates, the program must be integrated with Google Maps APIs so that users can view the visual location of the coordinates receive. A credential is required to use the Google Maps APIs

service, which can be obtained by placing a request through Google console.

Submit Attendance

The system automatically updates attendance in the database for any faces that the program could recognize. Driver mobile devices are remotely connected to the local database. The information updated is student name, x-coordinate, v-coordinate, number and timestamp. Maps.

E. LOCAL BINARY PATTERN HISTOGRAM

Local Binary Pattern Histogram Only a few face recognition algorithms are provided in the OpenCV library, including Local Binary Pattern Histograms (LBPH), Eigen-faces and Fisher-faces. This project uses LBPH technique, which takes a different approach compared to the other methods. In LBPH, characterization of features is done locally. The LBPH algorithm comes from a visual descriptor for pattern classification mainly used in computer vision. In this project, images were set to 127X 127 pixels.

It is necessary to maintain the image size to avoid affecting area rate while recognizing face. Because the LBPH algorithm is very susceptible to scaled images. That is, once the algorithm extracts a feature, the program can only identify the person when given a picture at an equivalent scale (in pixels).

The first requirement of the LBPH algorithm is to convert the image to grayscale mode. Grayscale images are not in black and white or binary.

Grayscale mode may be a series of numbers, each of which represents a special intensity. Having images in grayscale mode represents a big advantage when using the LBPH algorithm, as the image are often treated as a vector to extract valuable information. Next, for each pixel in the grayscale image, we select a neighbouring pixel of size 8 surrounding the centre pixel. The LBP value is calculated supported the centre value by thresholding it to a 3 X 3 array. The intensity level threshold is set to 8. Formal description of the LBP operator is often given as equation

$$LBP_{P,R}(X_c, Y_c) = \sum_{p=0}^{P-1} S(i_p - i_c)\, 2_p$$

Fig 6: Calculating LBP

$$S(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ 0 \text{ if } x \leq 0 \end{cases}$$

(1), where the notation (P, R) denotes a neighbourhood of P sampling points on a circle of radius R:

Formally, given a pixel at (Xc,Yc), the resulting LBP are often expressed in decimal form as in equation (1),where intensity ip and ic are respectively gray-level values of the central pixel and P surrounding pixels within the circle neighbourhood with a radius R, and performance s(x) is defined as:

The operator LBP (P, R) creates 2p different output values, matching to 2p different binary patterns formed by P pixels within the neighbourhood. The basic LBP operator is invariant to monotonic grey-scale transformations maintaining pixel strength order within the local neighbourhoods.

The histogram of LBP labels calculated over a region can be exploited as a texture descriptor.Fig.6 is an example of calculating the LBP value with a neighbouring size 8 pixel.

In the above figure, the worth 4 is that the centre pixel. each of the values represents a colour intensity. The expected output is an 8-binary digit for each pixel LBP calculation. After performing the LBP calculation, the value is stored in a 2D array with the exact same dimension as the input image. With 8 adjacent pixels converted to binary digits, we have a total of 2^8 = 256 possible combinations of local binary patterns. The stored result in an 8-bit array can be processed to obtain a decimal value. This process is visualised in Fig.5.

For the aim of illustration, we start at the highest right and move clockwise (the blue boxes indicate the sequence) to accumulate the binary string. the sequence of collecting the binary string doesn't matter provided we use an equivalent sequence for all other Local Binary Pattern (LBP) calculations.

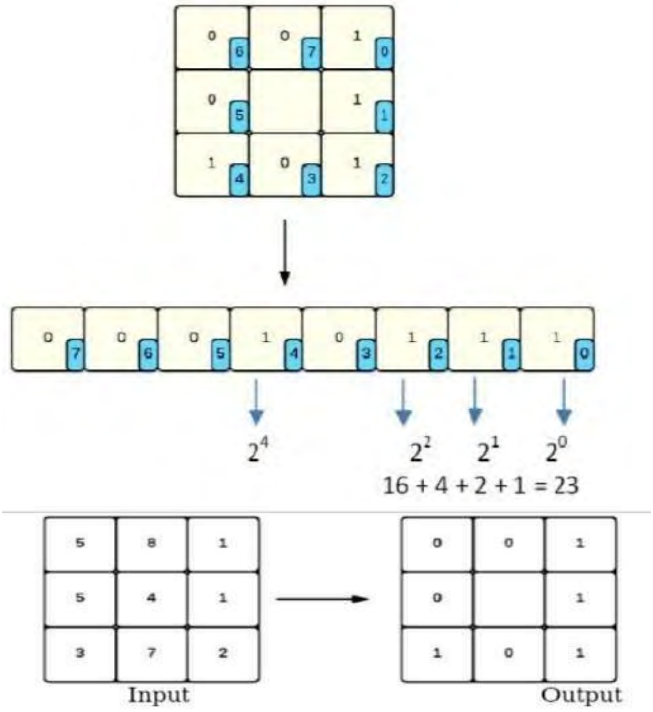


Fig 5: 8-bit Binary Representation

Fig. 4 illustrates on the proper the output value from the original image on the left. The process of thresholding, accumulating binary strings and storing the calculated LBP value is repeated for each pixel in the input image.

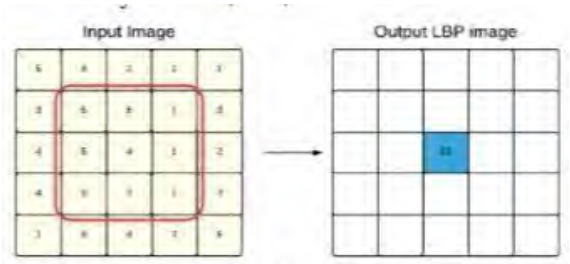


Fig.4.The output value from the original image

After obtaining all LBP values for every pixel, the next step is the histogram. The histogram represents the number of times each LBP pattern occurs that acts as a feature vector. The LBPH algorithm represents the local structure of a picture by calculating the histogram efficiently and summarising the histogram across different blocks to summarise, the steps to create LBP histograms are as follows:

1. Convert image to grayscale
2. Calculate the LBP for each pixel
3. Create a histogram supported each of the LBP values
4. When new faces are provided, generate the LBP histogram exactly as was done for the trained image.
5. Recognition comes when a new histogram matches the histogram pattern of a trained

Authors and Affiliations

[1] Prof. Vidya Raut
Professor at Cummins college of Engineering for women, Nagpur. Department of Computer Engineering

[2] Miss. Gayatri Supatkar
Student of Cummins college of Engineering for women. Nagpur. Department of Computer Engineering

[3] Miss. Pooja Shiv
Student of Cummins college of Engineering for women, Nagpur. Department of Computer Engineering

[4] Miss. Snehal Warade
Student of Cummins college of Engineering for women, Nagpur. Department of Computer Engineering

RESULTS AND DISCUSSION

The application of driver tracking and attendance management system was successfully tested on local machine and android application. After logging into the application, the driver needs to verify his identity through the face recognition and then by clicking on the 'submit attendance' button in Figure 7, shown below:

While submitting the attendance it is mandatory for the driver to choose the route he is travelling and then the station for which he is submitting the attendance. The data of the routes in Maharashtra have successfully feed into the database. The Figure 8 shows the procedure to select the bus route and the station name.

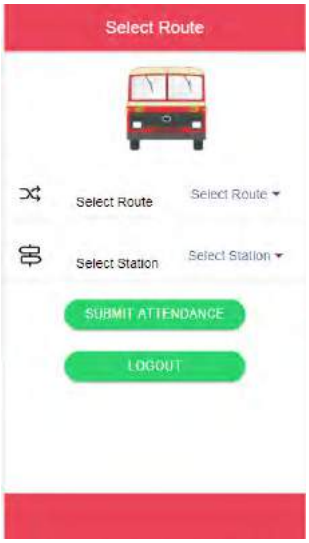


Fig 7: Submit Attendance Screen

We discussed about the application developed with the staff of state transportation Maharashtra government throughout the development process. They appreciated that there is no such applications used to manage driver's attendance and application like this will make this process more efficient and reduce the man efforts.

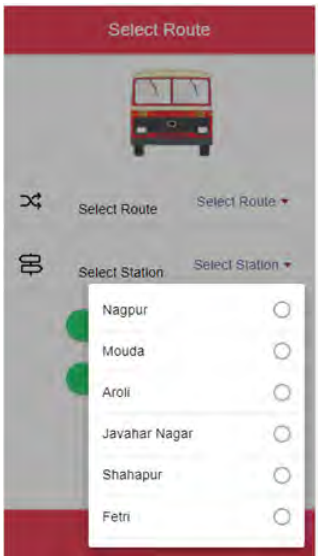


Fig 8: Application screen to select bus route and bus station

CONCLUSION

The Driver attendance tracking system uses the application was successfully designed and tested for real time data. The system has the advantages of small size, low costs, full featured and powerful expansibility. It can be easily installed and used by the drivers to submit the attendance and reduced the burden of managing attendance manuall. The validation of driver's location using the GPS system made it efficient and accurate in terms of location detection and submission. This system is an application-based system and can run on platforms like android. This is an intelligent and sophisticated mobile driver attendance management and tracking system. This system proved to be much more efficient and produced good results in sending driver's current location to Admin or respective depo manager, also in keeping track of travelling history of drivers.

References

[1] A smart location-based time and attendance tracking system using mobile application (IJCSEIT)Feb 2015 shermin sultana, Asma Enayet, Ishrat jahan mouri

[2] Real Time Tracking System (IJARCET) Keshav Agrawal, Shradha Dewarkar, Sakshi Salokhe, SujayRamesh, Vidya pujari

[3] GPS enabled employee registration and attendance tracking system. (ICCICCT) Albert Mayan 2015 J.M. Yusuf Khan, S. P. Avinaash Ram

[4] Kadry, S., & Smaili, M. (2013)." Wireless attendance management system based on iris recognition", Scientific Research and essays, 5(12), 1428-1435.

[5] Vehicle tracking and monitoring system to enhance the safety and security driving using IOT, a. anusha, syed musthak ahmed.

[6] A smart, location-based time and attendance tracking system using android application, shermin sultana1, asma enayet1 and ishrat jahan mouri1, international journal of computer science, engineering and information technology (ijcseit), vol. 5, no.1, february 2015.

[7] A location-based time and attendance system, mohammad salah uddin, member, iacsit, s. m. allayear, n. c. das, and f. a. talukder, international journal of computer theory and engineering, vol. 6, no. 1, february 2014

[8] Attendance system using a mobile device: face recognition, GPS or both? 1geetha baskaran, 2ahmad farhan aznan, international journal of advances in electronics and computer science, issn: 2393-2835 volume-3, issue-8, aug.-2016

[9] Smart college bus tracking system, prashantha n c1, rashmi p2, rashmi p s3, triveni d, international journal of advance engineering and research development volume 5, issue 05, may -2018,

[10] Real time vehicle tracking system based on arm7 gps and gsm technology, pradip v mistary,r h chile, 2015 annual ieee india conference(indicon)

[11] A Realtime vehicle tracking system, m. Humphries, p.radev, m.shirvaikar, proceedings of the thirty-seventh southeastern symposium on system theory,2005.ssst 05.

Real Estate Price Prediction Using Machine Learning Algorithm

Palak Furia,Dr. Anand Khandare
Department Of Computer Science Thakur College Of Engineering And Technology Mumbai, India
palakfuria18@gmail.com,anand.khandare@thakureducation.org

Abstract— ***Property Technology (PropTech) is the next big thing to disrupt the real estate market using technology to simplify operations and operations. Goods here refer to buildings or cities. It can be seen as part of a digital transformation in the real estate industry and focuses on both the technological and psychological changes of the people involved and could lead to new functions such as transparency, unprecedented data, statistical data, machine learning, blockchain and sensors are also part of PropTech. In India, there are many websites that divide the house and the land where property prices differ in price in the apartment and there are cases where the same apartments have different prices and as a result there is a lot of ambiguity.Sometimes buyers may feel that the price is not suitable for a particular listed apartment. . We suggest using machine learning techniques and automated methods to create an algorithm that can predict house prices based on specific input features. location, size, community, square foot, number of bedrooms. Commodity prices are closely linked to our economy. Apart from this, we have no price measures for the large amount of data available. Therefore, this project uses machine learning to predict house prices. One heuristic data commonly used in the analysis of housing price deficits is the Banglore city suburban housing data. Recent analysis has found that prices in that database are highly dependent on size and location. To date basic algorithms such as line deflection can eliminate errors using both internal and local features.The previous function of forecasting housing prices is based on retrospective analysis and machine learning. A local line model and a random forest model, vague assumptions, In addition, a multi-dimensional object model with two training items can also be used to evaluate house prices where something that predicts the "internal" price of a house is used, and the non-objective component can count neighbors 'preferences. The aim is to solve the problems of relapse where the target variable is the value and the independent variable region. We've used hot code coding in each of our institutions .The business application of this algorithm is that classified websites can directly use this algorithm to predict the values of new properties to be listed by taking variable input and predicting the correct and appropriate value.***

Keywords— Machine learning, linear regression, lasso regression, decision tree, data collection, data cleaning, oulier reduction, real estate, price prediction.

I. INTRODUCTION

We are in need of a proper prediction on the real estate and the houses in housing market field. we can see a mechanism that runs throughout the properties buying and selling buying a house will be a life time goal for most of the individual but There are lot of people making huge mistakes in united states of America right now when buying the properties most of the people are buying properties unseen from the people they dont know by seeing the advertisements and all over the grooves coming around the America one of the common mistakes is buying the properties that are too expensive but its not worth it. In the housing market 2017, there is a survey that in the year 2016 the house sold in a area were about

5.42 million but the starter home inventory down up to 10.7% from 2015. collapse in the year 2007 and 2008 there was an economic so there were several economic indicators that give the clue of impending disaster, this situation is currently happening and the economic indicators are suggesting that the housing prices are getting high as people are using the real estate to know the current economic situations, the US government fails to produce the data about the house prices so it is becoming difficult to buy the properties so the people who are in need to buy houses are using the Internet as source to search so there is evidence there is a correlation between housing sales and housing prices. In general, real estate may have the valuation of land may be obliged to furnish. A quantitative measure of the profit is carried out by many different Players in the commercial center. Business worth will be evaluated through that requisition. From claiming valuation systems Also methods that reflect those nature Of property and the condition under which those provided for. The property might well on the way exchange in open market under many conditions and circumstances, people are the unaware amount the current situations and they start losing their money, the change in prices of properties would affect both the common people along with the government, to avoid circumstances there is a need of price prediction. Many methods are used in the price prediction like a hedonic regression in this I am trying to predict the predict the real estate price for the future using the machine learning techniques with the help of the previous works. The methods that I have used are linear regression lasso regression and decision tree.

II. RELATED WORK

In the meantime each frame can be moved to a refresh at a simple launch from functionalities. The training framework comes with continuous e. Each person tends to move from the brochure to the traffic lights. That is

the primary goal of this would be expected for installation costs to be agreed on the customer system. Those signaling strategies can be a long process that those customers need communication with a global provider. The ground operator gives a valid A view on the forecast cost forecasting. This strategy poses a significant risk directly the effect a global operator can give to bad customer information. They hire those straightforward ones the calculation should see the cost. This same analysis was used to foresee the best place where customers bought houses. The information used here comes from those Mumbai sleeper board since 2009. In the end, Tom's ability to use this exact return predicted the average of all square distances. This forecast shows a square the feet of the house will be finally lifted to Toards 2018. (Bhagat et al .; 2016)

Mankind's wealth is measured Lastly the purchase of a home involves a limited measure of seeking an integrated option. As for illustration, the same how to provide a

House can be more complicated than that. There is a need for both buyers and sellers should receive the same amount of profit. A different model will be loaded Finally that is the Cox recovery model for those predictions. This model may be proposed starting with the Survival Study. The information used for this prediction comes from a website called Trulia. He added that it is difficult to find the actual sales and website time this is because the target time is beyond the details. He also outlines uncontrolled learning styles it is more popular than other methods. (Li and Chu; 2017) and states that the recovery method helps to predict prices with the help of using each with signs related to the house and its surroundings.

(Li and Chu; 2017) It is very important that the effort made by Toward for the value of the house. House value can have an impact on looking at different budget items. As we all know, China is one of the most densely populated countries in the world. Here the author tries to guess how to help banks provide their clients with mortgage loans. That prediction compares to Cathy's house services list provided by China. The data are obtained from the Taipei input phase of increasing efficiency after data collection using an algorithm to study the neural network to predict the price and accuracy of the forecast that can be obtained using

RMSE (ROOT MEAN SQUARE ERROR) and MAE (IMEAN ABSOLUTE PERCENTAGE ERROR). (Li

and Chu; 2017) (Willmott; 1981)

(Park and Bae; 2015) In 2005 there seemed to be a lot of development in the US housing market so America was forced to crash and it shows that the US housing market declined to 30 to 60% in the big cities continued for many years, after November 2012 it began to recover because investment was low and therefore it was necessary and therefore the author is trying to research and develop a forecasting model to find out if the closing price is high or low by using machine learning to find information and predict the future. Here using the KDD model discovery information used here seems to be compiled from different data sets and uses the WEBA software to find many algorithms like the decision tree used for relational database, here in the park and bay using RIPPER, C4.5 (J48) , Buzzy Bayesian and Ada-Boost whole algorithm is used under different conditions RIPPER is used to select class and quantities, Blank Bayesian is used to classify data set into different classes by calculating probability distribution and AdaBoost is used to improve classification and here you make methods two of the three alternatives separated by 10 folds and 10 folds to verify its achievable results RIPPER more predictability compared to others.

(Piazzesi and Schneider; 2009) For those who foresee that product value differently the arrangement can be quite complex. Cost forecasts are used primarily in the import business sector. But forecasting from supply demand can be pretty complicated due to the fact that there is a consolidation power along the way. A neural programming model needs to be made to predict stock value. This provides an overlap between those shares And benefits. In this model, the manufacturer has applied the need for stock details. The next project will see those shares return on stock holdings. Direct returns will be made with the first information available. At the same time multiplicity of first-order information is performed. After that Fourier analysis, that data suspension will be completed. This creates an MLP with a segmental neuron. The calculation of the neural system can be seen soon. This statistic gives you a selective accuracy of the forecast made and It gives you great comfort tolerance. The biggest obstacle is that the business knowledge of the poles continues to come in excess and forecasts become difficult.

The author(gu et al.; 2011) says that housing rate entails the various financial Hobby additionally it is each the authorities and the peoples so there is in need of correct Forecasting so the 3 researchers from key laboratory advanced a brand new version the use of the Genetic algorithm and the help vector machine. They have truely cited theRegression theorem of the aid vector gadget and introduced a new function referred to as Kernel with the assist of karush-kuhn-tuckers(ktt) conditions. Here they have combined The genetic algorithm with the svm and named it as g-svm in which the kernel functions May be in chromosomes and every will divide into three segments the writer is aware of the Fitness version so that they have calculated the fitness price of for each chromosome so there Will much less percentage of over becoming version and 3 operations selection, crossover and Mutation operation are completed and the consequences are acquired . There may be a evaluation Among the gray version and gsvm and gsvm executes the consequences quicker and extra Particular as suggested through the founders.

The authors(limsombunchai; 2004) try and provide a more accurate prediction at the House charges to improve performance to the actual estate found in new zealand he shows That most peoples in new zealand have their personal houses the pattern facts is received From one of the relied on real property organization so we can believe that there'll no longer be an Errors within the facts right here he as compared the hedonic rate and the synthetic neural community Theory while conducting the hedonic rate version there may be speculation based at the preceding Works it appears to have a advantageous dating. Within the neural;l community the writer uses the Skilled records if you want to avoid the prediction errors the work strategy of

neural networks Is said absolutely and at last whilst comparing the effects the writer says that artificial Neural community has extra performance when compared to the alternative one.

The authors(selim; 2009) have in comparison the a couple of regression evaluation over theArtificial neural networks by way of the use of the 60% statistics for the residence pricing prediction numerous Comparisons had been made of their predictive overall performance they've as compared with the Different education size and choosing the statistics of their length ie) the sample statistics size various For the performance detection. For calculating the error one-of-a-kind equations are used Mean absolute percent mistakes and absolutely the percent errors, here absolutely the Percent divides the houses into three distinct tiers based at the fe(forecasting Error) possibilities definitely six distinctive comparisons are made for more performance, right here it's Clear that if there may be sufficient or sufficient information length artificial neural community can perform Better or else the outcomes will be one of a kind as said by using(willmott; 1981)

The two authors (wu and brynjolfsson; 2009) from mit have performed approximately the Prediction that how the google searches the housing charge and income across the world Suggesting that within the present world each prediction percent factor is correlated with The following 12 months house income. The author wellknownshows approximately the correlation among them housing Charge and their related searches and the positive courting between them. The information is Taken from the google seek this means that the quest queries through the use of the google trends And with the help of a national affiliation of real-tors the facts is amassed for all of the States gift in the usa of the usa and discovered the best quantity of houses Sold throughout the year 2005 and the recession starts over 2009 by way of the usage of the auto regressive (ar) version, through the usage of it the connection between the hunt queries and housing market Signs they have got expected the baseline for housing price prediction and they're Properly proven in the figures and suggesting that if there call for to residence and there could be Call for in residence maintain appliances.

The writer offers the brief element about how the random wooded area algorithm is used for The regression and category, boosting and bagging are stated to be the methods which Produce a many classifiers the difference among the boosting and bagging is as stated with the aid of (liaw et al.; 2002) is the successive tress, the factor weights are calculated and majority Will take for the prediction.Throughout the year 2001 (nghiep and al; 2001) he proposed the Random forset that's associated with bagging and it offers extra randomness the entire Process of random forset classification and regression are stated here for the regression .

III. PROPOSED WORK

In India, there are multiple real estate classified websites where properties inconsistencies in terms of pricing of an apartment, and there are some cases where similar apartments are priced differently and thus there is a lot of in-transparency. Sometimes the consumers may feel the pricing is not justified for a particular listed apartment but there no way to confirm that either. We propose to use machine learning and artificial intelligence techniques to predict housing prices based on certain input features.

Technology and tools

1) Python
2) Numpy and Pandas for data cleaning
3) Matplotlib for data visualization
4) Sklearn for model building
5) Jupyter notebook, visual studio code and pycharm

as IDE

6) Python flask for http server
7) HTML/CSS/Javascript for UI

The Data Collection

It is a technique by way of which the researcher collects the information from all of the relevant sources to discover solutions to the research problem.The Bangalore home prices dataset was taken from kaggle.com. The dataset consists of 9 columns and 13,321 rows. The next very important step is data cleaning. Data cleaning is the procedure of identifying and disposing of (or correcting) faulty statistics from a dataset, table, or database.Some of the columns had been a drop from the data frame which did not help in deciding the final price.

The Data Cleaning Process

all of the N.A values were eliminated from the dataset. The incoming data is not uniform and it consists of errors so the data cleaning process was applied on the dataset to make the model more sophisticated. After data collection and data cleaning, comes data analysis. Data evaluation is the manner of evaluating statistics the use of analytical and logical reasoning to have a look at every aspect of the information provided.After analyzing data, prediction models are created using the Regression algorithm.

5.4. Flowchart:

Fig.1. flowchart

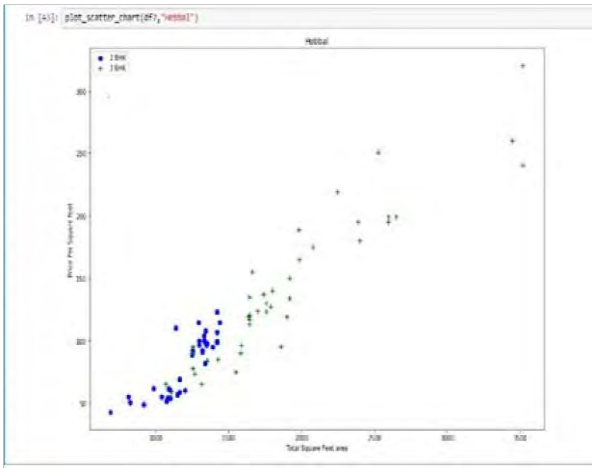


Fig.2. Before removing oultier

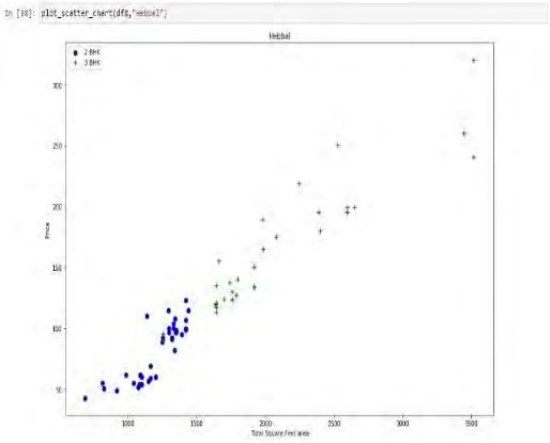


Fig.3. After outlier removal

The scatter chart was plotted to visualize price_per_sqft for 2 BHK and 3 BHK properties. Here the blue points represent the 2 BHK and green points as 3 BHK plots. Based on the above charts the outliers were removed from the Hebbal region using remove_bhk_outliers function.

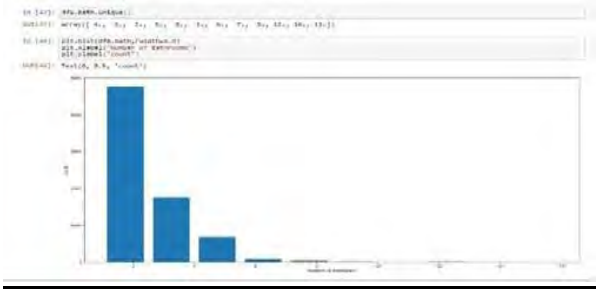


Fig .4. Outlier removal using bathrooms feature.

ALGORITHM:

Linear Regression

Linear regression is a linear approach to modeling the relationship between a scalar response and one or more explanatory variables.

It is a predictive modeling technique that finds a relationship between independent variable(s) and dependent variable(s).

The independent variable's can be categorical(e.g. US, UK, 0/1) or continuous(1729, 3.141 etc), while dependent variable's are continuous.

Least Square Method

The least square method is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve.

During the process of finding the relation between two variables, the trend of outcomes are estimated quantitatively.

This process is termed as regression analysis. The method of curve fitting is an approach to regression analysis.

This method of fitting equations which approximates the curves to given raw data is the least square.

It is quite obvious that the fitting of curves for a particular data set are not always unique.

Thus, it is required to find a curve having a minimal deviation from all the measured data points. This is known as the best-fitting curve and is found by using the least-squares method.

The least-square method states that the curve that best fits a given set of observations, is said to be a curve having a minimum sum of the squared residuals (or deviations or errors) from the given data points. Let us assume that the given points of data are (x1,y1), (x2,y2), (x3,y3), …, (xn,yn) in which all x's are independent variables, while all y's are dependent ones. Also, suppose that f(x) be the fitting curve and d represents error or deviation from each given point.

Now, we can write:

d1 = y1 - f(x1) d2 = y2 - f(x2) d3 = y3 - f(x3) ….. dn = yn – f(xn)

The least-squares explain that the curve that best fits is represented by the property that the sum of squares of all the deviations from given values must be minimum. I.e: Least Square Method formula

Sum = Minimum Quantity

Lasso Regression

Lasso regression is a kind of linear regression that uses shrinkage. Shrinkage is wherein statistics values are reduced in size in the direction of a central point like mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). The acronym "LASSO" stands for the Least Absolute

Shrinkage and Selection Operator.

L1 Regularization

L1 regularization is carried out by lasso regression, which provides a penalty equal to the absolute value of the magnitude of coefficients. This type of regularization can result in sparse models with few coefficients; some coefficients can become zero and eliminate from the model. Larger penalties result in coefficient values closer to zero, which is ideal for producing simpler models. On the opposite, L2 regularization (e.g. Ridge regression) doesn't result in the elimination of

coefficients or sparse models. This makes the Lasso far easier to interpret than the Ridge.

Performing the Regression

Lasso solutions are quadratic programming problems, which are best solved with software. The goal of the

$$\sum \quad yi \sum x \,.\, j \qquad \lambda \sum \quad \beta$$

Which is the same as minimizing the sum of squares with constraint Σ |Bj≤ s. Some of the βs are shrunk to exactly zero, making the regression model easier to be interpret. A tuning parameter, the strength of the L1 penalty is controlled by λ. λ is the amount of shrinkage: When λ = 0, no parameters are eliminated. The estimate is same to the one found with linear regression. As λ increases, an increasing number of coefficients are set to 0 and removed (theoretically, when λ =

∞, all coefficients are eliminated).

As λ increases, bias increases.

As λ decreases, variance increases.

If an intercept is included in the model, it is usually left unchanged.

Decision Tree

A decision tree is a flowchart-like tree structure where an internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. The top node in a decision tree is called as the root node. It partitions the tree in a recursive manner call recursive partitioning. The time complexity of decision trees is a function of the number of records and the number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not rely on probability distribution assumptions. Decision trees handle high dimensional statistics with the right accuracy.

How does the Decision Tree algorithm work?

The fundamental idea in the back of any decision tree set of rules is as follows:

1. Select the best attribute using Attribute Selection Measures(ASM) to split the data.

2. Make the attribute a decision node and then breaks the dataset into smaller subsets.

3. Starts building the tree by repeating this process recursively for each child until one of the situation will fit:

- All the tuples belong to the same attribute value.
- There aren't any extra remaining attributes.
- There are no more instances.

Attribute choice measure is a heuristic for choosing the splitting criterion that partition records into the satisfactory possible way.ASM gives a rank to each function(or characteristic) with the aid of explaining the given dataset. The best score attribute can be selected as a splitting attribute. Attribute selection is a prime undertaking is to identify the attribute for the root node at each level. Most popular selection measures are:

- Information Gain
- Gini Index

Information Gain

Information gain is a measure of the change in entropy. Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy more the information content.

Suppose S is a set of instances, A is an attribute, Sv is the subset of S with A = v, and Values (A) is the set of all possible values of A

Gini Index

Gini Index is a metric to measure how frequently a randomly selected element might be incorrectly recognized. It means an attribute with a lower Gini index should be preferred. The formula for the Gini Index is calculated by subtracting the sum of the squared probabilities of each class from one. It favors larger partitions.

Accuracy of the model:

The dataset was divided into 80% of the training dataset and 20% of the testing dataset. The required libraries were imported and GridSearchCV was used to find the best model. It compares the model on different regressors and different parameters and gives the best score among them. With the help of gridsearchcv we have compared the algorithms ie linear regression ,lasso regression and decision tree . According to the results linear regression gives more accuracy with 84% as compared to other algorithms hence we have used linear regression as our predictive model for prediction of real estate price.

IV. RESULT AND DESCUSSION

TABLE I. ACURRACY TABLE

	Model	Best Score	Best Parameters	Time Complexity	Error score	Accuracy percent
0	Linear regression	0.84776	{‘normalize’: False}	0.0725	0.15224	85%
1	Lasso regression	0.726738	{'alpha': 2,'selection': 'cyclic'}	0.1886	0.22326 2	73%

2	Decision tree	0.71606 4	{‘criterion’: ‘friedman_mse’, ’splitter’: ’best’}	0.2145	0.28393 6	72%

By means of undertaking this test with various device getting to know algorithms its been clear that linear regression are acting better with greater accuracy with 85% percentage and with much less blunders values. Whilst this experiment is as compared with the and to the end result done these algorithms predicts properly. This task has been performed with the the primary intention of this mission is to determine the prediction for costs which we've effectively completed the usage of exclusive system studying algorithms like a linear regression, lasso regression, decision tree, random wooded area, multiple regression, guide vector gadget, gradient boosted trees, neural networks, and bagging, so it’s clear that the linear regression area have greater accuracy in prediction when as compared to the others and additionally my studies presents to locate the attributes contribution in prediction. So i might agree with this studies may be beneficial for each the peoples and governments and the future works are stated under each system and new software program generation can help inside the destiny to expect the fees. Fee prediction this can be advanced by way of including many attributes like environment, marketplaces and lots of different related variables to the houses.

Fig.5. Predicting Price

Fig.5 show some of the values predicted for the entered data i.e. location, area, number of room, number of bathroom. The anticipated records can be stored inside the databases and an app may be created for the humans so they could have a quick idea and they'd make investments the cash in a more secure way. If there's a possibility of realtime records the statistics may be linked to the h2o and the gadget getting to know algorithms can be immediately connected with the interlink and the software surroundings may be created.Huge information and its associated technology.

V. Coclusion

Many algorithms used here to successfully increase the percentage of accuracy, various researchers have done this work and used practical techniques such as line engraving, lasso restoration, command tree, which are considered the best models in pricing forecasting. These are considered as basic models with the help of advanced algorithms data algorithms such as random forest, linear growth trees, multiple perceptron meters and integrated learning models are used and prediction accuracy is obtained at a high level. Results and experiments of this species using machine learning and advanced data mining tools such as Put, Rapid Miner will have more influence on pricing forecasts.

References

[1] Bhagat, N., Mohokar, A. and Mane, S. (2016). House price forecasting using data mining, International Journal of Computer Applications 152(2): 23–26. URL: http://www.ijcaonline.org/archives/volume152/number2/26292-2016911775

[2] Breiman, L. (1996). Bagging predictors, Machine learning 24(2): 123–140. Chang, P.-C. and Liu, C.-H. (2008). A tsk type fuzzy rule based system for stock price prediction, Expert Systems with applications 34(1): 135– 144.

[3] Ganjisaffar, Y., Caruana, R. and Lopes, C. V. (2011). Bagging gradient-boosted trees for high precision, low variance ranking models, Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, pp. 85–94. 4.Gu, J., Zhu, M. and Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine, Expert Systems with Applications 38

[4] 3383–3386. Koskela, T., Lehtokangas, M., Saarinen, J. and Kaski, K. (1996). Time series prediction with multilayer perceptron, fir and elman neural networks, Proceedings of the World Congress on Neural Networks, INNS Press San Diego, USA, pp. 491–496.

[5] Li, L. and Chu, K.-H. (2017). Prediction of real estate price variation based on economic parameters, Applied System Innovation (ICASI), 2017 International Conference on, IEEE, pp. 87–90.

[6] Liaw, A., Wiener, M. et al. (2002). Classification and regression by randomforest, R news 2(3): 18–22.

[7] Limsombunchai, V. (2004). House price prediction: hedonic price model vs. artificial neural network, New Zealand Agricultural and Resource Economics Society Conference, pp. 25–26.

[8] Mirmirani, S. and Cheng Li, H. (2004). A comparison of var and neural networks with genetic algorithm in forecasting price of oil, Applications of Artificial Intelligence in Finance and Economics, Emerald Group Publishing Limited, pp. 203–223.

[9] Nghiep, N. and Al, C. (2001). Predicting housing value: A comparison of multiple regression analysis and artificial neural networks, Journal of real estate research 22(3): 313–336. Park, B. and Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data, Expert Systems with Applications 42(6): 2928–2934.

[10] Patel, J., Shah, S., Thakkar, P. and Kotecha, K. (2015). Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques, Expert Systems with Applications 42(1): 259–268.

[11] Piazzesi, M. and Schneider, M. (2009). Momentum traders in the housing market: survey evidence and a search model, Technical report, National Bureau of Economic Research. Selim, H. (2009). Determinants of house prices in turkey: Hedonic regression versus artificial neural network, Expert Systems with Applications 36(2): 2843–2852.

[12] Trafalis, T. B. and Ince, H. (2000). Support vector machine for regression and applications to financial forecasting, Neural Networks, 2000. IJCNN 2000, Proceedings of the IEEEINNS-ENNS International Joint Conference on, Vol. 6, IEEE, pp. 348–353.

[13] Willmott, C. J. (1981). On the validation of models, Physical geography 2(2): 184–194.

Wu, L. and Brynjolfsson, E. (2009). The future of prediction: How google searches foreshadow housing prices and sales.

Generalized Approach for Accurate Breast Cancer Diagnosis

Aynaan Quraishi,Jaydeep Jethwa,Shiwani Gupta
Department of Computer Engineering Thakur College of Engineering & Technology Mumbai, India
aynaanq@gmail.com,jaydeepjethwa2401@gmail.com,shiwani.gupta@thakureducation.org

Abstract—**Breast Cancer is one of the major causes of death among women. However, it can be cured easily if diagnosed early. A lot of research has been done on diagnosing breast cancer using machine learning by using different machine learning algorithms. The results seem promising. In this paper, we have tried to improve the performance as well as the generalization of the machine learning model. Using two feature engineering techniques (Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA)) for processing the data and three classification algorithms (Logistic Regression, Support Vector Machine, and Naïve Bayes) were trained on the data obtained through each technique separately. Metrics like accuracy, precision, recall, and stratified cross-validation accuracy were used to compare the performances of each algorithm. Further, they were evaluated using Receiver Operative Characteristic (ROC) curve and Box plot. The results show that the Naïve Bayes algorithm trained on feature engineered data obtained through LDA had a 10-fold stratified cross-validation accuracy of 97.90%, accuracy of 99.41% on 70-30 split of data for training and testing respectively. Also had better Precision and Recall values. (Best overall metrics seen in any literature).**

Keywords—Breast Cancer Diagnosis, PCA, LDA, Naïve Bayes, SVM, Logistic Regression, Stratified K-fold cross-validation, Precision and Recall.

I. INTRODUCTION

The number of database sizes is increasing in the medical field. This influx of a huge amount of data has made it possible for various machine learning, data science applications. Analyzing these data sets with various new techniques has given rise to finding internal hidden patterns that were not possible before. The Wisconsin Breast Cancer Dataset is one such publicly available dataset from UCI Machine Learning Data Repository [1].

Breast cancer is one of the most frequent cancers affecting women. It is one of the deadliest cancers in women. It's believed by the World Health Organization that almost 2.1 million are affected by it each year and it's estimated in 2018 that almost 627,000 women died because of the disease [2]. That puts it at a whopping 15% of overall cancer-related deaths. Detection at the early stages ensures that there is a 30% chance that cancer can be treated effectively [3]. Cancer is a disease which can particularly affect organ and stimulates abnormal growth of that organ. There are two types of tumor namely Benign and Malignant. A Benign is a tumor that doesn't evade its surrounding tissue or spread in the body, therefore it's typically harmless [2]. While a Malignant tumor that may evade the surrounding tissue and spreads in the body; is harmful. Fine Needle Aspiration (FNA) of a breast mass is used to compute the features extracted from the digitized image.

Machine Learning is a widely used technique in the field of AI, in which machines are trained to learn without any human intervention to develop predictive models in health care for effective decision making [4]. In Breast Cancer Diagnosis, machines can be trained to classify and predict the type of tumor present in the breast mass as Benign or Malignant based on the features extracted from the FNA technique and successful classification results into the diagnosis of cancer.

In this paper, a comparative analysis of three machine learning classifiers namely Logistic Regression, Support Vector Machines (SVM), and Naïve Bayes is presented. These machine learning models were trained on the Breast Cancer Wisconsin (Diagnostic) data set from the UCI Machine Learning Repository [1]. Before training the data set on the above-mentioned classifiers, the data was pre-processed also a comparison of the performance of these classifiers on two feature engineering techniques namely Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) was done. 70% of the data set was used to train the classification models and 30% of the data set is used to test the trained models. These models were evaluated on metrics like accuracy, precision, recall and stratified 10-fold cross-validation accuracy. In Breast Cancer Diagnosis, precision and recall are very important factors that give the idea about how accurately a model is classifying the type of tumor as accuracy gives the overall performance of models. This ensures that our model is not biased towards a particular outcome resulting in a better generalization of the model.

II. RELATED WORK

This section deals with the research carried out in the prediction and diagnosis of breast cancers using different machine learning techniques and classifiers. The details of some of the work are given below:

In paper [5], popular data mining algorithms like (Naïve Bayes, RBF network, J48) are used to develop prediction models for classification. The use of 10-fold cross-validation was used to verify an unbiased result. Specifically, those algorithms were used that were immune to data imbalance. Interestingly Weka (Waikato Environment Knowledge Analysis) was used to analyze the dataset and evaluate the performance of data mining. The paper was able to achieve a cross-validation accuracy of 97.36%. Also achieved a precision and recall of 97.4%, 97.9% respectively.

In this paper [6] various machine learning algorithms and as well as deep learning algorithms were used such as Decision Trees, Random Forest, Multi Layer Perceptron (MLP), SVM, and Deep Neural Network (DNN). The important aspect of this paper is that it had a lot of important metrics that were justified and compared with each other thoroughly such as log loss, F1score, accuracy score, AUC score, ROC curves. It focused on the cross-validation score to verify True False Positives and False true positives. While DNN had the best accuracy score of 92%.

This paper [7] uses the SVM machine learning algorithm. The paper experiments and analyses different SVM methods like linear and non-linear SVM. Feature selection was an important part of this paper as in medical diagnosis a small feature set means lower tests and less diagnostic costs. It also decreases the computation time and effectiveness. There were 5 selected features based on their F-score. This paper achieves an accuracy of 99.51% on 80-20 % training-test partition.

In paper [8], a brief comparison of classification models namely MLP, Decision Tree (C4.5), SVM, and K-Nearest Neighbor (K-NN) is studied. Here, ANOVA (Analysis of Variance) was used for feature extraction to reduce noise from data. This paper was able to achieve a maximum of 10-fold cross-validation accuracy of 98.12% on MLP.

Ali Al Bataineh [9] presents a comparative analysis of performances between nonlinear machine learning algorithms: MLP, KNN, CART, NB, and SVM. The performances were evaluated on metrics: accuracy, precision, and recall. The best performing model was MLP with an accuracy of 99.12%, precision 99%, and recall of 99%.

Similar to [5], in paper [10], Bayesian Network and J48 machine learning algorithms were used which are immune to data imbalance. Also, WEKA was used for the analysis and evaluation of these models. This model was able to achieve 10-fold cross-validation of 97.80% with a 0.29% false-negative rate and a 1.90% false-positive rate on Bayesian Network.

III. Experiment & Analysis

3.1 Data Set

All the experiments and analysis presented in this paper is performed on publicly available Breast Cancer Wisconsin (Diagnostic) data set from UCI Machine Learning Repository [1]. There are 32 features and 569 different instances are recorded with respect to these features. The features in this dataset are computed from a digitized image of fine needle aspirate (FNA) of a breast mass. These features describe characteristics of the cell nuclei.

Ten main features about tumour in breast mass around which the dataset is prepared are:-

- Radius,
- Texture,
- Perimeter,
- Area,
- Smoothness,
- Concavity,
- Compactness,
- Concave points,
- Symmetry,
- Fractal dimension.

The dataset is divided in two different classes: Benign and Malignant as ‘B’ and ‘M’ respectively. Out of 569 instances 62.7% are of Benign and 37.3% are of Malignant.

3.2 Data pre-processing

Data pre-processing is the most important step as it deals

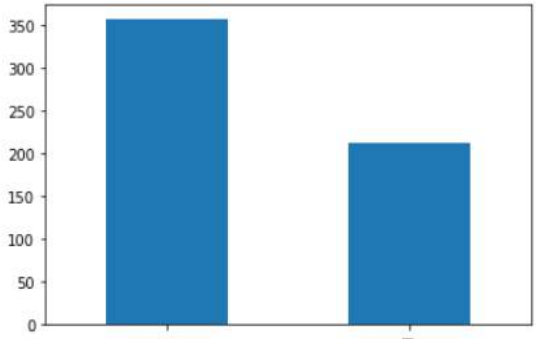


Figure 1: Observations per class

with missing values, normalization, encoding, scaling, etc. Feature engineering techniques like PCA and LDA were used because these techniques modify, combines, and deletes features to build a better training algorithm.

3.2.1 Principal Component Analysis (PCA)

PCA is an unsupervised feature engineering technique to reduce dimensionality and to increase the interpretability of information. It is performed in a way such that there is very little loss of information. This paper [11] verifies the importance of PCA in medical assessment. Also mentioned its applicability to problems like health diagnosis. To perform PCA following steps were followed:

- Standardization of the data.
- Computed the covariance matrix.
- Calculated the eigenvectors and eigenvalues.
- Computed the principal components.
- Reduced the dimensions of the data set.

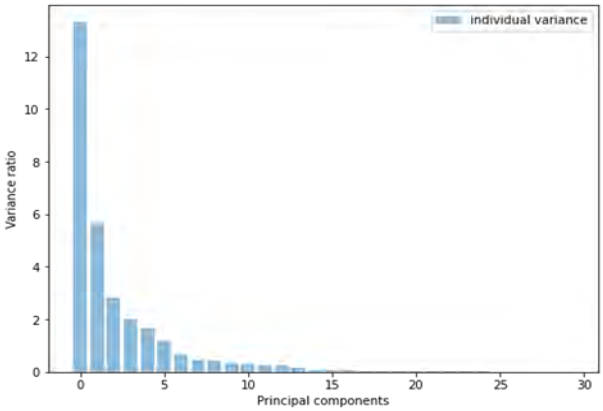


Figure 2: Variance ratio of components

Once the eigenvalues and eigenvectors were obtained, they were set in descending order. The eigenvector with the highest eigenvalues is the main principal components. The principal components were computed in such a manner that the obtained components are highly significant and independent of each other. These components were then selected based on how significant information they contain. This significance of data was judged by the variance ratio in each component.

Figure (2) shows the variance ratio present in each components obtained. Since the top 5 components had the maximum variance and retained 85% of the information, thus they were selected. We also used LDA as we understood that there could be a loss of valuable information by using PCA. This paper [13] goes into depths about the loss of information in PCA therefore as a countermeasure we used LDA.

3.2.2 Linear Discriminant Analysis (LDA)

Unlike PCA, LDA is supervised dimensionality reduction technique. The purpose of this technique is to find the feature subspace that optimizes class separability and hence this requires class labels.

In addition, it preserves most of the discriminative information that was present in original data set [12]. The feature can be reduced to k dimensions ($k < d$, where d is original dimensional space). In our case, we have two classes and hence one component is sufficient to separate feature subspace for better optimization.

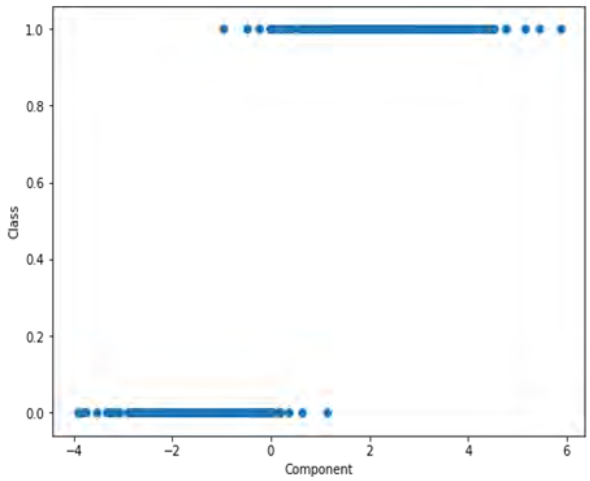


Figure 3: Class separation of data points obtained

The above figure shows scatter points of reduced component and its distribution for the two classes.

3.3 Classification algorithms used

3.3.1 Naïve Bayes

Naïve Bayes classifier makes a naive assumption that the impact of one variable is independent of the values of the other variables [14]. It gives a probabilistic value and is based on the Bayes probability theorem. Naïve Bayes algorithm was used because it was computationally faster and was very easy to use. This paper [15] highlights the advantages of using Naïve Bayes. An additional advantage of using Naïve Bayes is that it can also work on imbalanced data sets.

For our data set, Bernoulli Naïve Bayes was used. Since our data for classification comprises two classes hence Bernoulli Naïve Bayes was the perfect candidate as it only accepts and processes features that are binary in nature [16].

3.3.2 Support Vector Machine (SVM)

Support Vector Machines are defined as discriminative classifiers. For a given classification problem the algorithm outputs an optimal hyperplane by training on given labeled training data which is capable to categorize unseen examples [17].

SVM model represents examples as points in space and is mapped in hyperplane such that the example of different categories (classes) are divided by a clear gap (more the better). For our dataset, SVM is used to classify and find a suitable hyperplane that effectively separates the two classes present. Also, the RBF (Radial Basis Function) kernel was used for defining the decision boundaries in which the boundaries are non-linear for better separation of classes.

3.3.3 Logistic Regression

Logistic regression is one of the oldest machine learning classification algorithm that is used to predict the probability of a categorical variable that has been converted into a binary form. The advantage of using logistic regression was because of its simplicity and its interpretability, ability to sometimes outperform some complex machine learning algorithms. Regularization was used because a large number of k predictors don't increase the model efficiency thus regularization in a way optimizes the algorithm to reduce the unimportant predictors. Specifically, ridge regression was used since it includes all the features of the models and shrinks it. The model being used for medical purposes, thus ridge regression was used to prevent removal of some features.

3.4 Metrics

After training these models, we evaluated the performances of these models on some metrics namely; accuracy, precision, recall, and stratified k fold cross-validation accuracy. Accuracy is the percentage of correctly classified examples. It is calculated using a confusion matrix based on four possible prediction cases which are true positive (TP), true negative (TN), false positive (FP), false negative (FN) [20]. Accuracy tells us about the overall performance of the model.

Precision is the ratio of correctly predicted positive observation i.e. TP to a total number of positively predicted observation i.e. TP + FP. It gives an idea about actually correct classified positive observations out of total observations classified as positive. For example, if precision is 0.8 and our model predicts an observation as malignant, it is 80% of the time correct.

The recall is the ratio of correctly classified positive observation i.e. TP over a sum of total correct classified positive observation and total observations that were positive but are classified as negative i.e. TP + FN. For example, if the recall is 0.5, that means 50% of all malignant observations are identified correctly.

Precision and recall are important factors in Breast Cancer Diagnosis as correct prediction of a malignant class is necessary as it causes cancer. Hence, maximizing both should be our goal. Also, for classification, accuracy is not always a better choice for evaluating model performance [20].

Stratified K fold cross-validation is a technique where the whole data set is divided into K parts (we took K = 10) where the ratio of each class is equal in each part. Out of these 10 parts, (10 – 1) parts are used for training the models, and the remaining one part is used for testing the models and the same process is repeated 10 times [19]. Its

performance is measured as the mean of accuracies obtained after training and testing the model 10 times. Higher the mean accuracy obtained, better generalized our model is. Stratified K fold has an edge while dealing with the datasets that do not have an equal number of observations for classes under consideration [18].

Also, ROC (Receiver Operative Characteristic) Curve was used to evaluate the performances of the trained models. It gives us an overall sense of how much we should trust these models [21]. In classification problems, models give a probability (between 0 and 1). In breast cancer diagnosis, the model gives the probability of a person diagnosed with malignant tumours. ROC Curve is plotted based on True Positive Rate (TPR) and False Positive Rate (FPR) where TPR is the proportion of observations that were actually Malignant out of total observations classified as Malignant. FPR is the proportion of observations that were misclassified as Malignant type. Area Under the Curve (AUC) score of the models is calculated with respect to ROC Curve plotted where the perfect model will have an AUC score of 1.0. A model is said to be performing well if AUC score is above 0.7 and the model with a score of more than 0.85 is great.

IV. Result & Discussion

According to no free lunch theorem, if one of the algorithms performs well in one metric, it could fail in performing on the other. Thus, our goal was to find an algorithm that was performing well consistently on the metrics used by us. Therefore, we set out to compare the metrics. We have the algorithm comparison of Naïve Bayes, Logistic Regression, and Support Vector Machine between the two feature engineering methods used. The performance of models was calculated on the before mentioned metrics and the following results were achieved.

Table 1: Logistic Regression

Metric	On PCA	On LDA
Accuracy	0.9883	0.9883
Precision	0.99	0.99
Recall	0.99	0.99
Stratified K Fold Accuracy	0.9719	0.9772
AUC Score	0.9986	0.9985

Table 2: Support Vector Machine (SVM)

Metric	On PCA	On LDA
Accuracy	0.9766	0.9883
Precision	0.97	0.99
Recall	0.98	0.99
Stratified K Fold Accuracy	0.9754	0.9789
AUC Score	0.9972	0.9945

Table 3: Naive Bayes

Metric	On PCA	On LDA
Accuracy	0.9239	0.9941
Precision	0.92	0.99
Recall	0.92	1.0
Stratified K Fold Accuracy	0.9158	0.9790
AUC Score	0.9644	0.9953

From the above results, we observe that Models are performing very close to each other on data trained on both the feature engineered techniques except Naïve Bayes trained on data obtained through PCA which underperforms as compared to others. For better understanding, we have plotted Box plot with accuracies obtained on each fold of Stratified 10 fold cross-validation technique.

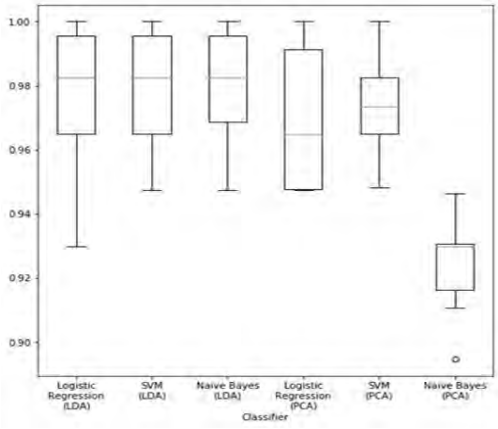


Figure 4: Box Plot for algorithm comparison

Figure 4 shows the Boxplot of the accuracy achieved on 10-fold stratified cross validation. Comparing the box plot results we found that results presented by the classifiers trained on data obtained by LDA had a higher median value than data obtained by PCA. The plot of LDA classifiers shown no outliers but the distribution variance of Logistic regression accuracy is high. While SVM and Naïve Bayes perform similarly but the median value of Naïve Bayes is more than SVM. The boxplot of Naïve Bayes is smaller than SVM indicating lower variance in the distribution of accuracy values in Naïve Bayes. Thus, making it the best algorithm in this case.

From the above analysis, we found that SVM and Naïve Bayes models perform identical on data obtained through LDA. Naïve Bayes has a slight edge because of its precision and recall are on the higher side. Also, for the problem under consideration, if a benign class is predicted incorrectly, then it is not much of a problem as compared to misclassification of malignant class and due to which a patient will be declared healthy and proper treatment will not be given. Hence, a recall of 1.0 and a precision of 0.99 given by Naïve Bayes makes it the best performing model. The same was verified by comparing the ROC curve (Figure 5) and AUC scores.

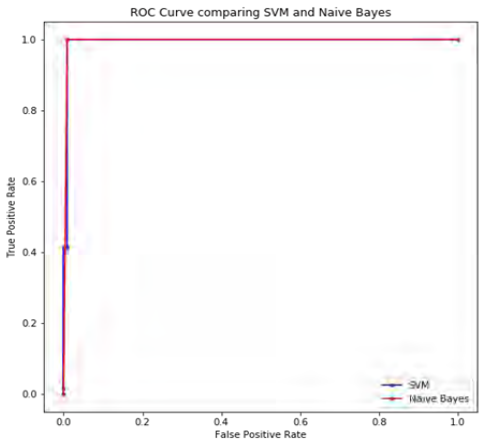


Figure 5: ROC Curve

V. CONCLUSION

Finally, by applying two feature engineering techniques and three classification algorithms, our study was able to find the algorithm which performed the best in all the metrics consistently and was able to generalize well. The algorithm Naïve Bayes, trained on data obtained from LDA, performed best in most of the metrics. It had a 10-fold stratified cross-validation accuracy of 97.90%; the accuracy of 99.41% on 70-30 split of data for training and testing respectively, the precision of 99% and recall of 100%.

Table 4: Accuracies achieved by previous papers

Algorithm	Accuracy achieved	Year
Naïve Bayes [5]	97.36 (Cross-validation)	2018
DNN [6]	92.00	2019
SVM [7]	99.51	2009
MLP [8]	98.12 (Cross-validation)	2018
MLP [8]	99.12	2019
Bayesian Network [10]	97.80 (Cross-validation)	2015

From table 4, we found that only one paper [5] was able to achieve a higher value of cross-validation (98.12) compared to the results shared in this paper. The algorithm used by them was a deep learning algorithm, which requires higher computation time and speed than normal machine learning algorithms. Also, deep learning models require excessive data for training and testing. We managed to achieve higher precision and recall values 99 and 100 respectively compared to 99.2 and 97.85 achieved by them. Our model was able to generalize better and perform outstandingly on most of the metrics.

References

[1] Dua, Dheeru and Graff, Casey, " UCI Machine Learning Repository", 2017. [DataSet] Available: http://archive.ics.uci.edu/ml.[Accessed: July 24,2020].

[2] WHO, "WHO position paper on mammography screening", 2014. [Online] Available: https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ [Accessed:July 15,2020].

[3] Joann G. Elmore, Connie Y. Nakano, Thomas D. Koepsell, Laurel M. Desnick, Carl J.D'Orsi, and David F. Ransohoff, "International Variation in Screening Mammography Interpretations in Community-Based Programs", Journal of the National Cancer Institute, 95(18), pp.1384–1393, 2003 September 17.Doi: 10.1093/jnci/djg048

[4] Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. "Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis", Journal of Personalized Medicine, 10(2), 2020 March 31. DOI: 10.3390/jpm10020021

[5] Vikas Chaurasia,Sauhrab Pal,BBtiwari "Prediction of benign and malignant breast cancer using data mining techniques", Algorithms and Computational Technology,Vol.12(2), pp.119-126, 2018. DOI: https://doi.org/10.1177/1748301818756225

[6] Teixeira, J. L. Z. Montenegro, C. A. da Costa and R. da Rosa Righi, "An Analysis of Machine Learning Classifiers in Breast Cancer Diagnosis", 2019 XLV Latin American Computing Conference (CLEI), Panama, pp. 1-10, 2019. DOI: 10.1109/CLEI47609.2019.235094.

[7] Mehmet Fatih Akay "Support vector machines combined with feature selection for breast cancer diagnosis", in conf. Expert Systems with Application , vol.36, pp.3240-3247, 2009, DOI: https://doi.org/10.1016/j.eswa.2008.01.009

[8] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques", 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), Erode, pp.997-1002,2018. DOI: 10.1109/ICCMC.2018.8487537.

[9] Ali Al Bataineh, "A Comparative Analysis of Nonlinear Machine Learning Algorithms for Breast Cancer detection", conf. International Journal of Machine Learning and Computing, Vol.9, No.3, June 2019.

[10] Borges, Lucas, "Analysis of the Wisconsin Breast Cancer Dataset and Machine Learning for Breast Cancer Detection", Proceedings of Xi Worskshop de Visao Computacional", 2015. [Online] Available: https://www.researchgate.net/publication/311950799_Analysis_of_the_Wisconsin_Breast_Cancer_Dataset_and_Machine_Learning_for_Breast_Cancer_Detection

[11] A. Malhi and R. X. Gao, "PCA-based feature selection scheme for machine defect classification", in IEEE Transactions on Instrumentation and Measurement, vol. 53, no. 6, pp. 1517-1525, Dec. 2004 , DOI: 10.1109/TIM.2004.834070.

[12] E.K. Tanga, P.N. Suganthana, X. Yaob, A.K. Qina, "Linear dimensionality reduction using relevance weighted LDA", The Journal of Pattern Recognition Society, Vol. 38, pp.485–493, 2004. DOI: https://doi.org/10.1016/j.patcog.2004.09.005

[13] Geiger, B., & Kubin, G, "Relative Information Loss in the PCA", in Proc. IEEE Information Theory Workshop, pp.567-571, 2012. DOI: 10.1109/ITW.2012.6404738

[14] V. KaliyaMeiyar and D. Shanmugasundaram, "The Comparative Study for Diagnosing Heart Disease Using", International Journal of Advance Research in Computer Science and Management Studies, Vol.3, Issue 8, pp. 9-19, August 2015.

[15] Pouria Kaviani and Mrs. Sunita Dhotre, " Short Survey on Naive Bayes Algorithm", International Journal of Advance Engineering and Research Development, Volume 4, Issue 11, November 2017. [Online] Available: https://www.researchgate.net/publication/323946641_Short_Survey_on_Naive_Bayes_Algorithm

[16] Pedregosa F., Varoquaux Ga"el, Gramfort A., Michel V., Thirion B., Grisel O., others. (2011), "Scikit-learn" Machine learning in Python. Journal of Machine Learning Research, pp.2825–2830, 12 October 2011.

[17] Awad M., Khanna R. (2015) Support Vector Machines for Classification. In: Efficient Learning Machines. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4302-5990-9_3

[18] Victoria López, Alberto Fernández and Francisco Herrera, "On the importance of the validation technique for classification with imbalanced datasets: Addressing covariate shift when data is skewed", Information Sciences, vol.257 , pp.1-13,2014. DOI: https://doi.org/10.1016/j.ins.2013.09.038

[19] Rohan Kohavi, "A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection", in conf. International Joint Conference on Artificial Intelligence IJCA, Vol.2, pp.1137-1143, August 1995.

[20] Brendan Juba, Hai S. Le, "Precision-Recall versus Accuracy and the role of large data sets", in conf. The Thirty-Third AAAI Conference on Artificial Intelligence(AAAI-19), Vol.33, 2019. DOI: https://doi.org/10.1609/aaai.v33i01.33014039

[21] Carter, Jane & Pan, Jianmin & Rai, Shesh & Galandiuk, Susan. (2016), "ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves", Surgery, Volume 159, Issue 6, pp.1638-1645, June 2016, DOI: 10.1016/j.surg.2015.12.029

Cricket Match Winner Prediction Using Machine Learning

Sarvesh Kharche, Saranjeet Saluja, Rohit Gupta, Yash Shah
Department of Information Technology,Vidyalankar Institute of Technology,Mumbai, India
sarvesh.kharche@vit.edu.in, saranjeet.saluja@vit.edu.in, rohit.gupta@vit.edu.in, sarvesh.kharche@vit.edu.in

Abstract— **Cricket is one of the most famous sports activities in the whole world, and also one of the most popular sports in India. Cricketing occasions inclusive of the Indian Premier League (IPL) and One Day Internationals (ODIs) are thoroughly enjoyed by means of enthusiasts all throughout the country. Fans of the game love predicting the ongoing match results, and this is something that has ended up being a hobby for numerous people who observe the game. This is a sport with an abundant amount of statistics and using this information, we will make an evaluation on whether a team can win an ongoing IPL match or similarly, an ODI match. This prediction is implemented by the usage of system learning algorithms such as Multilayer Perception Classifier, Decision Tree Classifier, K-Nearest Neighbor and Random Forest. The required dataset is obtained via collecting the usage of an internet site and consolidated. As a result, the output is received which lists whether the home team has secured a win in the match or not.**

Keywords— Machine Learning, Data Mining, Cricket Match Winner, Score Prediction, Indian Premier League (IPL).

I. INTRODUCTION

Regardless of the type of format the sport is played in, cricket is a beloved and extremely popular sport in our country, having a massive fan-base. As fans, the people make their own predictions while watching a particular match based on the information given, they have and then, they make a call on who will win the match.

Cricket is essentially a bat and ball game that is played between 2 teams having 11 participants each. Each team comes to bat and has a sole inning in which it seeks to attain as many runs as possible, while the opposite team fields. The innings ends when the full quota of deliveries, which depends on the game format which is being played, or the ten batsmen have been dismissed, whichever comes first. The prime objective is to score more runs & therefore runs are the decisive factor.

There are 3 widely approved formats of cricket on the international level - T20, One Day Internationals and Test match. The scheduled length of the game is the prime distinction among these three formats, which without delay modifies the number of deliveries each team get to play of their respective innings. Test cricket format is the longest one and is acknowledged as the highest general of the sport. Match length is five days in which each squad gets to play 2 innings each. A regular day of a test match consists of three sessions, of two hours for every session.

One Day International i.e., ODI layout is of finite overs, where each team faces 300 deliveries (50 over's). Generally, ODI match falls in any of the two categories: Day or Day- Night match.

T20 is the shortest internationally identified format of this game, where each innings encompasses of 20 overs. This is extra of an "explosive" and greater "athletic" than the other two formats.

This research aims at predicting the result of an ongoing cricket match based totally on the information and data that is available from previous matches. The data that is available for each match includes the teams involved in the match, the venue, the winner of the match, the margin with which they won and the toss decision. We will be performing prediction for all of the matches that have taken place in the IPL. This is executed with the help of using machine learning algorithm for making the prediction of the result of the matches. This research is primarily based on predicting the winner of a cricket match based totally on the records available from previous matches.

II. RELATED WORK

Haghighat et al. [1] described the various prediction techniques including Nave Bayes which can be used to predict the best playing eleven for the team. Although this research work is aimed at specifically basketball, it can be extended to any sport, including cricket. This research work used basics of Python and the Scikit-learn API to implement the project. The different types of machine learning algorithms used in this research work included Support Vector Machine and Gaussian Naive Bayes.

D. Thenmozhi et al. [2] explains the comparison between four machine learning algorithms viz. Gaussian Naïve Bayes, Support Vector Machine, K-Nearest Neighbor and Random Forest applied on an IPL dataset.

Shubhra Singh et al. [3] The paper addresses the problem of predicting the outcome of an IPL cricket match. Factors such as luck and player strength were used as key functions in predicting the winner of a match. The novelty of the proposed method lies in addressing the trouble as a dynamic one and the usage of an appropriate non-relational database, HBase for scalability of application. Out of all the machine learning algorithms used, KNN has been located to be the maximum accurate.

Prakash et al. [4] proposed three variations of predictive models using Support Vector Machines to predict the winners of IPL matches.

Rory P. Bunker and Fadi Thabtah et al. [5] explains generating models and using machine learning algorithms such as Artificial Neural Networks to determine which team will win the match. This research work was generalized to all sports, and so can be utilized for different sports including cricket.

Nazim Razali et al. [6] explains predicting the winner of a Football match in the English Premier League, which is a famous football tournament. This implemented by using Gaussian Naive Bayes, a machine learning algorithm. We have used similar to this in predicting the winner of the cricket match. This analysis makes use of a variety of models to identify and anticipate the winner of the match.

Jalaz Kumar et al. [7] shows data for ODIs was obtained from ESPNcricinfo [8] and scraped using a script by sending one request per second. Furthermore, the matches which finished in tie/draw or were disrupted due to rain were removed from the dataset as a part of data cleaning. Along with the above, the matches between special teams such as World XI or Asia XI were eliminated. The data for the matches reproduced by switching between two teams, i.e., a match between 1. Australia and 2. New Zealand was reproduced as 1. New Zealand and 2. Australia. The dataset thus, extracted and cleaned was converted into a categorical one from a continuous one by using counterfeit variables.

Factors included for analysis include:

- Previous performances of teams,
- The match being played on the home or the away turf
- Innings and
- Home turf upper hand.

The main aim of using a Decision Tree is to create a training model that can be used to predict the class or value of the target variable by using certain tree based rules which is identified from prior data. In Decision Trees, for predicting a class label for a record we start comparing from the root of the tree. We compare the values of the root attribute with the record's attribute. Based on comparison, we follow the branch corresponding to that value and jump to the next node.

In KNN (K-Nearest Neighbors) algorithm the data points are classified based on the points that are most similar to it. It uses the prior data (training data) to estimate what an unclassified point should be classified as.

Random forest is a supervised machine learning algorithm that is used for both classification as well as regression. But however, it is mainly used for problems which involve classification. As we recognize that a forest is made up of trees and more trees means more robust forest.

Similarly, random forest algorithm creates decision trees on data samples and then receives the prediction from every tree and subsequently selects the best solution by using voting. It is an ensemble method that is better than a single decision tree as it reduces the over-fitting by using average of the result.

Thenmozhi et al. [2] In this research the author has predicted the outcome of an IPL cricket match using four different machine learning algorithms viz. Gaussian Naive Bayes, Support Vector Machine, K-Nearest Neighbor and Random Forest.

III. Proposed Methodology

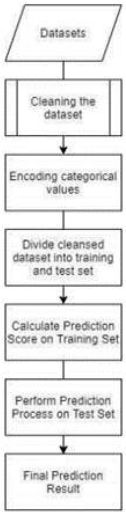


Figure 1. Flowchart of proposed Methodology

A. Dataset Selection

The data-set is obtained from Kaggle[www.kaggle.com/]. The data set consists of various attributes such as season(in years), the city and venue in which the match is being played, the teams involved in the match, the toss winner and decision(field or bat), the player of the match and the umpires for each match. The data set is cleaned by removing certain duplicate values, renaming certain values, and handling missing and null data.

B. Encoding Categorical Values

Certain categorical values such as the teams are replaced with numeric values, where an integer is assigned to each team. For example, all instances of the team 'Mumbai Indians' are allotted the integer value 1. Similarly, other categorical values such as toss decisions which consist of only two values (field/bat) are replaced by 0s and 1s, respectively. The remaining categorical values are dropped, as they have a lot of unique values which is not feasible.

C. Decision Trees

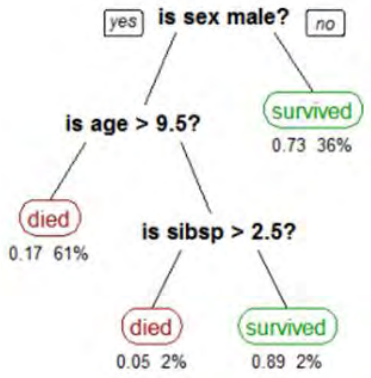


Figure 2. Example of Decision tree Classifier

A decision tree is drawn upside-down with its root on the top. In the figure above, the bold text in black represents a condition/inner node, based totally on

which the tree splits into branches/ edges. The end of the branch that does not split anymore is the decision/leaf, in this case, whether or not the passenger died or survived, represented in red and green textual content, respectively.

Although an actual dataset will have a lot more features and this may simply be a branch in a much larger tree, however, you cannot ignore the simplicity of this algorithm. The feature importance is clear, and relationships can be easily viewed. This technique is more commonly known as a learning decision tree from data and the above tree is called Classification tree because the target is to classify if a passenger survived or died. Regression trees are represented in an identical manner, just they predict continuous values like price of a house. In general, Decision Tree algorithms are known as CART or Classification and Regression Trees. So, what actually happens in the background (growing a tree) involves identifying on which features to choose and what conditions to use for splitting, also knowing when to stop. As a tree typically grows arbitrarily, you will need to trim it down for it to be efficient.

D. Random Forest Classifier

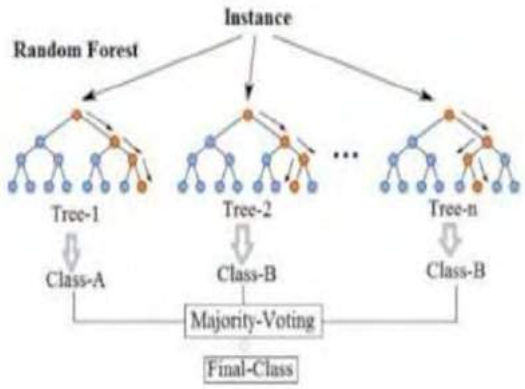


Figure 3. Random Forest Classifier

RF Classifier (Random Forest, RF-algorithm) is used here. RF classifier is an ensemble method which creates a myriad of decision trees during time of training. It finally takes the mode or average of the output classes by these trees.

Usually, as suggested[2],it is assumed that the objects of the data-set U which is used for the RF classifier development, are deleted into classes with the labels from the set $Y = \{1,2,...,\eta_1,...L\}$ (η_1 is the label of the l th class). A herewith, each object z_i ($i = \overline{1,s}$ where s is number of objects in data-set)can be described by the vector $z_i = (z_i^1, z_i^2, ..., z_i^n)$ of the numerical values in the n -dimensional space of features. The data-set U can be considered as the set $\{(z_1, y_1),...,(z_s, y_s)\}$, in which each object z_i has the class label y_i $(y_i \in Y = \{1,2,...,\eta_1,...L\})$.

Random Forest works in two-stages, first stage is to create the random forest by combining N decision trees, and second stage involves making predictions for each tree created in the first stage.

Steps involved in the working of Random Forest algorithm:

1. Select random number of data points from the training set.
2. Build decision tree with respect to each of the selected data points.
3. Choose a number N for the number of decision trees you want to build.
4. Repeat step 1 and 2.
5. For each new data point, find the predictions of the decision tree, and assign the new data points to the category that wins the majority votes.

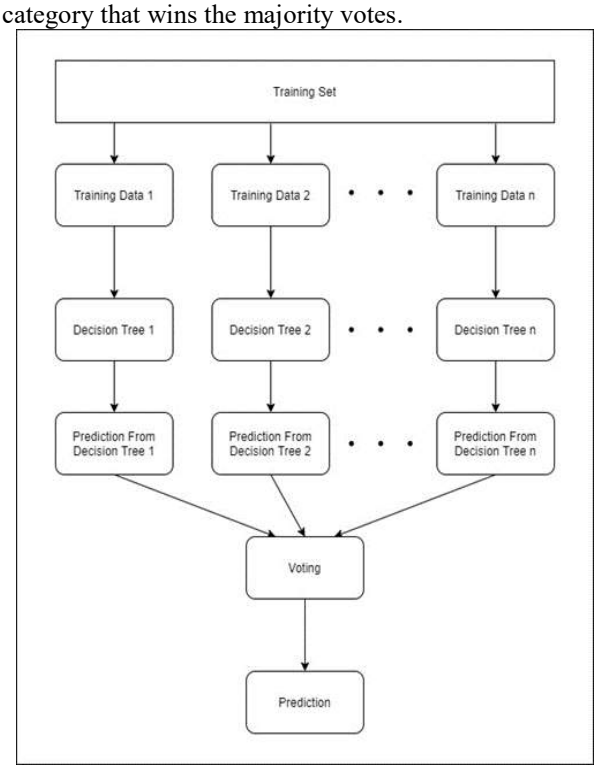


Figure 4. Random Forest Algorithm

Advantages of Random Forest Classifier:

- Highly accurate
- Estimation of attributes relevant for classification
- Generates an internal unbiased estimate of the generalization error as the forest building progresses.

Disadvantages:

- Prone to overfitting of data in a noisy classification.
- For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels

IV. RESULT

Figure 5. displays the final results after prediction. For each algorithm along the x-axis and accuracy along y-axis, the highest accuracy for each algorithm was depicted. Thus, the graph shows the highest accuracy for each algorithm.

Table 1 describes the highest accuracy obtained by each algorithm for a particular generated model. From the

results, it is observed that the most preferred algorithm which provides the highest accuracy is the Random Forest algorithm. This approach gives an overall accuracy of 87% on predicting the winners of the matches.

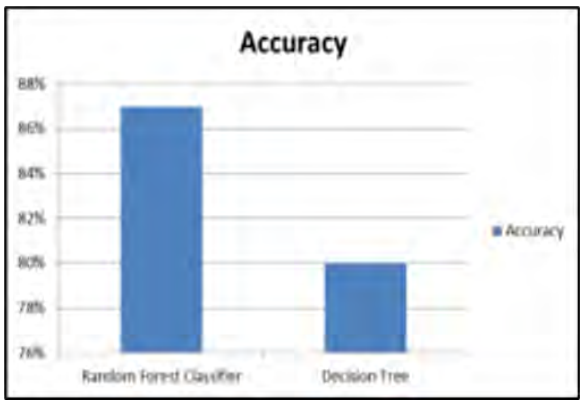


Figure 5. Accuracy achieved by different algorithms

TABLE I. OVERALL COMPARISON OF ACCURACIES

Algorithm	***Accuracy***
Decision Tree	80%
Random Forest	87%

V. CONCLUSION

Thus, the winner of the matches is predicted using various attributes using different algorithms such as Random Forest Classifier and Decision Tree. While the Decision Tree gives us about 80% accuracy, the Random Forest Classifier gives us better accuracy comparatively at 87%. Further, we aim to create a simulation system that will simulate the cricket match between two teams, selected by the user and accurately predict total runs scored by each team and thus predict the winner of the match along with the scorecard.

ACKNOWLEDGMENT

We would like to show our deep appreciation to our project guide, Prof. Yash Shah, who helped us finalize our project, who gave us the opportunity to do the research and provided us invaluable guidance throughout this research.

References

[1] Maral Haghighat, “A review of Data Mining Techniques for Result Prediction in Sports”, Advances in Computer Science : an International Journal (ACSIJ), vol. 2, no. 5, ISSN : 2322-5157, pp. 54–61, 2013.

[2] D. Thenmozhi, P. Mirunalini, S. M. Jaisakthi, Srivatsan Vasudevan, Veeramani Kannan V, Sagubar Sadiq S,” MoneyBall - Data Mining on Cricket Dataset”, Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019),2019.

[3] Shubhra Singh and Kaur, P., “IPL Visualization and Prediction Using HBase”, Procedia computer science, vol. 122, pp. 910–915, 2017.

[4] Prakash, C. D., Patvardhan, C., and Lakshmi, C. V., “Data Analytics based Deep Mayo Predictor for IPL-9”, International Journal of Computer Applications, vol. 152, no. 6, pp. 6-10, 2016.

[5] Rory P. Bunker, Fadi Thabtah, “A machine learning framework for sport result prediction”, Applied Computing and Informatics, vol. 15, no.1, pp. 27–33, 2017.

[6] Nazim Razali, Aida Mustapha, Faiz Ahmad Yatim and Ruhaya Ab Aziz, “Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL)”, International Conference on Material Science and Engineering, vol. 226, no. 1, pp. 012099:1–6, 2013.

[7] [7] Jalaz Kumar, Rajeev Kumar, Pushpender Kumar, “Outcome Prediction of ODI Cricket Matches using Decision Trees and MLP Networks”,2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC).

Smooth Medicare Services Using Machine Learning Techniques

Upasana Patil, Tejaswini Yeole, Sachin Patil, Pratik Chavan, Tukaram Gawali
Department of CE , Shri Vile Parle Kelavani Mandal's Institute of Technology, Dhule, India
upasanapatil26@gmail.com, tejaswiniyeole28@gmail.com, sachinpatil4188@gmail.com, pratikchavan601@gmail.com, t.gawali@gmail.com

Abstract -Health **is a useful gift of nature. It is one among the essential needs of all the citizenry and it's influenced by many factors, such as, food, housing, basic sanitation, healthy lifestyles, protection against environmental hazards and communicable diseases. Health care is the most vital aspect of any society and is prime need for every citizen of every country. Within the recent past vast amount of knowledge is generated in Bangladesh. We used machine learning techniques to predict the quantity of various outpatient of community clinics. A singular approach towards deciding process and better quality is developed by using the machine learning. Machine Learning approach is employed in determining the patient satisfaction in health care sector. To predict the self-care problems of youngsters with physical and motor disability beforehand, an expert system is proposed using machine learning. (KNN) K Nearest Neighbor is proposed to predict self-care problems.**

Keywords—Temperature, Arduino UNO, Patient Satisfaction, Healthcare, Regression, Self-care, Physical and Motor Disability, Health Data Analysis, Community Clinics, Bangladesh

I. INTRODUCTION

Health care is the maintenance or improvement of health via the prevention, treatment, recovery, or cure of disease, illness, injury and other physical and mental impairments in people. A healthy population plays an important role in contributing the economical burden to the government and reduce pressure overloaded hospitals, clinics. We all know the fashionable healthcare system is vital to remain people effective and healthy. In Bangladesh there are 64 districts and 492 sub-districts Dysentery diarrhea, Bacterial and Fungal diseases. According to Sandwip Upazilla Health complex survey 61.4% patients get admitted. During this manner we used ML techniques to predict number of out of door patients who will visit the clinic. Clinic management are getting to be benefited and manage their human resource in better way. Just in case of health care sector patient satisfaction plays a key role in evaluating the quality service they provides. So an expert system is proposed to predict the self-care problems of youngsters with physical and motor disability, Computer-based expert systems are developed using different machine learning classification techniques. Among them, Artificial Neural Network (ANN) is one of the foremost used once because of performance. Among the other used classification techniques, Support Vector Machine (SVM), K Nearest Neighbor (KNN) But a haul with these classifiers is that each of the mentioned classifiers isn't imagined to work well altogether situations.

II. LITERATURE SURVEY

In this paper an endeavor possesses to make a literature survey on totally different aspects of health care watching system.The review of literature is also a vital a vicinity of the analysis in any field.throughout this section, we have a tendency to gift recent connected works on swish Medicare services victimisation Machine Learning techniques.Recent and former literatures have designed totally different prototypes for patient watching system.However, there area unit many limitations for these studies.The sudden development of health technologies fostered the prospect of measurement Associate in Nursing outsize quantity of clinical knowledge with the last word aim to strengthen patients' management [1].Nowadays, physicians and medical researchers will perpetually monitor clinical knowledge of every patient, permitting correct pursuit of the disease's evolution.Such knowledge area unit usually collected and keep in electronic health records (EHR), that holds promise to strengthen potency and quality of attention, creating knowledge a lot of accessible,facilitating health info exchange and ability between attention suppliers [2] Machine learning techniques (MLTs) provide a replacement chance in terms of the management of this info.A growing body of literature shows MLT applications in medical specialty, particularly for developing prediction models victimisation each supervised and unsupervised strategies [3].supply Regression [LR], Support Vector Machine [SVM] and Neural Network [NN] were applied to guage the practicableness of such techniques in predicting hospitalization of patients with HF.we have a tendency to set to match these algorithms, given their increasing quality in clinical settings for prediction of binary outcomes and their ability to find advanced relationshipsbetween the result and predictors and interactions between covariates [4].Logistic Regression (LR) is maybe the manoeuvre most often accustomed predict the incidence of an incident in clinical analysis [5].the recognition of LR is especially associated with its ability to supply important and easy-to-interpret quantities like odds ratios (ORs), which might offer clinical info on the impact ofpredictors on the incidence of the event of interest.However, LR is understood to possess some limitations given its constant quantity assumptions and so the matter to find non-linearities and interactions between covariates.LR was typically used as a benchmark in studies aimed to match totally different MLTs for the prediction of the incidence of a binary outcome [6].Support Vector Machine (SVM) is Associate in Nursing algorithmic program that was developed for binary classification settings with 2 categories [7].SVM works by constructing hyperplanes

of the covariates' area that separates the observations in line with the class they belong to.The separation is travel by augmenting the features' area victimisation kernel functions to permit for non-linear relationships between the result and so covariates.the employment of such kernel functions permits the analysts to find and model advanced relationships, which might be quite common in clinical analysis.SVM showed sensible classification ability in many settings, Associate in Nursingd it's been established to be an honest challenger of different Machine Learning Techniques [8].NNs area unit a generalization of statistical regression functions.Neural Network (NNs) area unit characterised by units, known as neurons, that area unit connected.In its simplest type, the neurons take the information from the input units, i.e., the worth of the predictors inside the dataset, computed a weighted total of the receivedinputs and supply Associate in Nursing output, which, in classification tasks, is that the class expected by the NN for every observation.NNs area unit enforced victimisation several parameters such they're planning to flexibly approximate any swish functions.NNs area unit wide utilised in pattern recognition field that they have recently become terribly trendy in medical analysis, being shown to exceed several different MTLs. Malady|heart condition|cardiopathy|cardiovascular disease} normally occurring disease and is that the most clarification for overtime today.K-Nearest Neighbor (KNN) is that the wide used lazy classification algorithmic program.KNN is that the foremost well liked , effective and economical algorithmic program used for pattern recognition.Medical knowledge sets contain Associate in Nursing outsize variety of options.The Performance of the classifier have gotten to be reduced if the information sets contain wheezy options.Feature set choice is projected to unravel this downside.after we believe patient care in health care sector, knowledge plays a big role in analyzing personal expertise factors that relates with patient satisfaction.Patient satisfaction is measured to be emphatically related to to personal expertise conjointly as hopes.any studies show that the relations of the expert with the patient conjointly as his/her complete expertise correspondingly conducive issue that's thought-about to be vital than the necessary consequenceOrganization and client Satisfaction attention sector is taken into account as a service adjusted and so the patients as client the organization ways is tested to yield the satisfaction.Patient health care {is significant|is critical|is necessary} however conjointly as self-care is in addition important thus, self-care is taking care of own self throughout a healthy and correct approach.Managing common conditions, taking care of body elements and thus the opposite factor that's associated with taking care of own self area unit typically a self-care activity.

III. RELATEDWORK

Machine Learning has proved to be a viable resource for tending suppliers, consecutive step is scaling the creation of intelligence and its integration into the progress at the purpose of decision-making. Cerner itself is construction analytical tools that draw on the volumes of secure, anonymized patient information it already has access to: diagnosis and treatment outcomes, monetary outcomes from claims and writing, request tools, prophetical hospital staffing models, and additional data processing may be a multidisciplinary field wide utilized in the clinical field like prediction of heart condition. Researchers developed varied techniques to predict the center exploitation data processing. Feature choice is employed to predict the malady. Their methodology obtained associate degree accuracy of ninety two.5% for thirteen options and 100 percent accuracy with fifteen options. there's a seven.5% improvement when discarding a pair of options from fifteen to thirteen. Machine Learning helps to contour body processes in hospitals, map and treat infectious diseases and change medical treatments. ..."It may also be accustomed demonstrate and educate patients on potential malady pathways and outcomes given totally different treatment choices. Wireless health observation system or patient observation system involves observation of patients vital organ remotely by suggests that of devices that transfers patient information to remote locations wirelessly. ... Continuous observation of patient health is very important throughout treatment. during this high-speed race of life, it's terribly troublesome to perpetually monitor the patient's body parameter like sugar level, heart rate, temperature so on once the close to ones aren't perpetually accessible whereas the patient is affected by a malady. therefore to attenuate the patient's burden of observation from the hospital or from the doctor's head, this work have given the methodology for checking patient's remotely by the utilization of GSM network and technology referred to as terribly giant Scale Integration (VLSI). This patient observation system records physical characteristics either by regular intervals of your time or by endlessly. during this we tend to work monitor the human health in period of time and alarm the amendment in organic structure for patient health specially for those that area unit affected by diseases. The Embedded system that have a Embedded ARM microcontroller is connected to a sensors and GSM module, the system checks the patients health by the medical signals if any abnormalities area unit found, to transfer the signals to the heart the system uses the patient's phone, were the doctor attracts the conclusion for the signals that have received and sends the medical recommendation to the patient in order that his/her life is saved. The work may be a open supply platform and therefore the elements simulation is completed exploitation Proteus eight computer code. during this work we've got targeted on the patients within the remote space handed by one doctor were it becomes troublesome for one doctor to observe sizable amount of patients. therefore an online based mostly application is developed to observe endlessly and appearance when the parameters like BP, Heart Rare etc

IV. PROPOSED WORK

This work gift the look and construction of a operating epitome for good health care system through machine learning techniques which may offer Quality Health Care to everybody. To urge patient satisfaction from the given information sets we tend to applied the renowned machine learning techniques. It's accustomed analyze vast information sets and their relationship between info} in turn it provides the meaning information. throughout this study we tend to use Machine learning ideas that is dived into 2 classes as supervised and unattended learnings. In supervised learning includes classification, regression prediction and data point analysis that helps informing models that's accustomed classify new and unknown data. In unattended learning we tend to use clusters in decisive the various patterns and similarity between the information. For our studies we tend to used Regression ideas for analyzing the information the results square measure valid by statistical regression and binomial correlation victimization ancient statistical procedure.

A. Classification :

When most dependent variables square measure numeric, supply regression and SVM ought to be the primary select classification. These models square measure straightforward to implement, their parameters straightforward to tune, and therefore the performances also are pretty sensible. thus these models square measure applicable for beginners.

B. Regression :

Regression analysis consists of a gaggle of machine learning ways that enable North American country to predict endless outcome variable (y) support the worth of 1 or multiple predictor variables (x). Briefly, the goal of regression model is to create a mathematical equation that defines y as a perform of the x variables.

Simple regression analysis uses one x variable for every dependent "y" variable. For example: (x1, Y1). multivariate analysis uses multiple "x" variables for every freelance variable: (x1)1, (x2)1, (x3)1, Y1).

C. bunch :

Clustering algorithmic rule is one in every of the foremost standard information analysis technique in machine learning to exactly appraise the immense range of tending information from the body sensing element networks, net of things devices, hospitals, clinical, medical information repositories, and electronic health records etc. The bunch algorithms invariably play an important role to predict the diseases by partitioning the similar patient's information supported their relevant attributes.

Figure1. Supervised vs Unsupervised Learning

V. RESULT ANALYSIS

The main objective of this study is to look at the parameters of patient satisfaction. A framework model has been constructed the system collect the parameter values and compared with their standard point values. patients ID, AuthKey , values of Systole, Diastole pressures, pulse and temperature along side the accurate location with Longitude-Latitude to the doctor. The test on ten different patient has been taken and every one the parameter values of the individual patient is shown within the Table 1. The opinion attitude towards Patient satisfaction consists of 5 determinants. These constructs are determined using Likert scale of 1 to 5 where 1 denotes strongly disagree and 5 denotes strongly agree. The descriptive statistics of predictive statistics are depicted is denoted by chart in Figure 2. The opposite patients P1,P2,P3,P5,P7,P9 and P10 are normal are shown in figure 3.

Normal pulse is 60 to 100 per minutes

Bradycardia – pulse rate is smaller amount than 60 per minutes

Tachycardia - pulse is grater than 100 per minute

Normal vital sign

Systolic = 100 to 140 torr

Diastolic = 60 to 90 torr

	Heart Rate	systolic	Diastolic	Body Temperature
P1	70	110	72	30
P2	78	129	74	26
P3	80	125	74	26
P4	75	137	74	26
P5	66	121	88	32
P6	58	107	76	30
P7	79	124	113	30
P8	60	114	76	42
P9	76	118	90	26
P10	65	104	70	32

Table 1. Parametric values of all the 10 patients

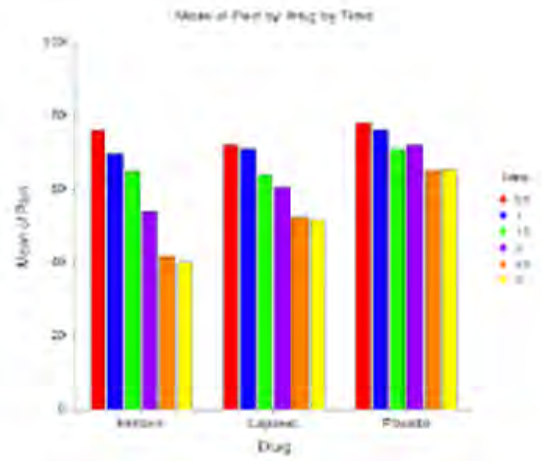


Figure 2. Descriptive Statistics of Patient Satisfaction

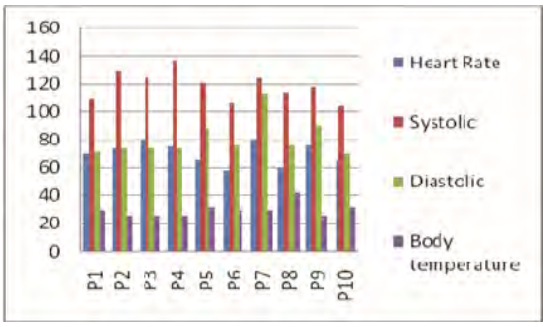


Figure 3. Graph that represents all the individual patients parametric value

VI. CONCLUSION

In our analysis, we have a tendency to foreseen the day of the week once most feminine and kid patients inherit the clinic for treatment. thus the clinic authority might organize support staffs consequently. throughout this we have a tendency to used the knowledge accessible from varied customers to provide a legitimate read purpose to the organization in setting the services related to patient satisfaction .For patient health care we have a tendency to used some machine learning techniques in our paper like KNN, SVM and NN techniques. KNN is value to what extent k-Nearest-Neighbour classifier enhance potency and accuracy amongst patients seeking emergency treatment in Kenya, it had been evident that application of k-NN algorithmic program will greatly facilitate scale back errors in designation, scale back time spent on identification while up potency and effectiveness in treatment. SVM modeling could also be a promising classification approach for predicting medication adherence in HF patients. This prophetic model helps stratify the patients so as that evidence-based call square measure usually created and patients managed fittingly. NN square measure utilised in forefront analysis establishments to hunt out solutions for complicated health issues. These solutions square measure able to offer future quality care and save patients .

VII. ACKNOWLEDGEMENT

The authors are thankful to prof. Tukaram Gawali, for their encouragement, support and guidance for successful completion of the paper.

References

[1]. Jaiee Sitaram Adivarekar, Amisha Dilip Chordia, Harshada Hari Baviskar, Pooja Vijay Aher, Shraddha Gupta, "Patient Monitoring System Using GSM Technology" in International Journal Of Mathematics And Computer Research [Volume 1 issue 2 March 2013] Page No.73-78 ISSN :2320-7167

[2].Mrs.Sonal Chakole1,Ruchita R.Jibhkate, Anju V.Choudhari,Shrutika R.Rawali, Pragati R.Tule , "A healthcare monitoring system using wifi module" in International Research Journal of Engineering and Technology (IRJET)[Volume: 04 Issue: 03 | Mar -2017] page No. 14131417

[3] Sandwip Upazilla Health Complex Reports, 2017

[4] WHO, Global Health Observatory Data Repository: Life expectancy –Data by country, 2015.

[5]. Lis CG, Rodeghier M, Gupta D, Distribution and determinants of patient satisfaction in oncology: Areview of the literature, 2009, 3(3):287-304

[6]. Mpinga EK, Chastonay P, Satisfaction of patients a right to health indicator? Health Policy. 2011, 100(2-3):144-50

[7] .S. McLeod and T. T. Threats, "The icf-cy and children with communication disabilities", in International Journal of Speech-Language Pathology, Vol. 10, No. 1-2, pp. 92-109, 2008.

[8]. W. H. Organization, "International Classification of Functioning, Disability, and Health: Children & Youth Version: ICF-CY", World Health Organization, Geneva, 2007.

[9]. Nithi Pillai,Madhura Kakade,Shreyas Bhimale,Mahesh Bagde , "Real-
Time Health Monitoring System for Remote Places" in International
Journal of Computer Science and Information Technologies, [Vol. 6 (5) ,2015] Page no. 4390-4391

[10] P. Fergus, A. Hussain, D. Hignett, D. Al-Jumeily and K. AbdelAziz, "A
Machine Learning System for Automated Whole-Brain Seizure Detection" in Applied Computing and Informatics [Volume 12, Issue 12016] Page no. 70-89.

[11].Wagner D, Bear M, Patient satisfaction with nursingcare: a concept analysis within a nursing framework, JAdv Nurs. 2009, 65:692–701.

[12].Marton KI, Sox HC, Alexander J and Duisenberg CE,
Attitudes of Patients toward Diagnostic Tests: The Case
of the Upper Gastrointestinal Series Roentgenogram,
Medical Decision Making, 1982, 2:439-448

[13]. G. H. Skrepnek, "Regression methods in the empiric analysis of health care data." Journal of Managed Care Pharmacy 11.3 (2005): 240-251.

[14] V. Morton, T. David J, "Effect of regression to the mean on decision making in health care." BMJ: British Medical Journal 326.7398 (2003): 1083.

[15]. O'Toole RV, Castillo RC, Pollak AN, MacKenzie EJ, Bosse MJ, LEAP Study Group. Determinants of patient satisfaction after severe lower-extremity injuries. J Bone Joint Surg Am. 2008, 90(6):1206-11.

Cloud Computing In Ehealth

Ms.Amruta Patil, Mr.Praful Pawar, Ms..Neha Baviskar, Prof. Ashish Awate,Ms.Janhavi Kulkarni
Department of Computer Engineering,Shri Vile Parle Kelvani Mandal's Institute of Technology,Dhule, Maharashtra, India.
patilamruta9319@gmail.com, prafulpawar7887@gmail.com, nehabaviskar111@gmail.com, ashish.awate87@gmail.com, Janhavik172@gmail.com

Abstract— **Cloud technology is hired to make community along patients, doctors, and care establishments by offering applications, services and moreover through maintaining the record within cloud. Presenting the cloud administrations in the wellbeing area not just encourages the trading of electronic clinical records among the emergency clinics and facilities, yet in addition license the cloud to go about as a clinical record stockpiling focus. This survey paper targets to discuss, examine security challenges and available solutions in cloud computing. Various approaches were used to keep the safety of the health information in the cloud environment. The SPS model, DACAR model are used to enhance security of data.**

***Keywords— eHealth, cloud computing, health records, sensor networks, security, privacy, data buckets*.**

I. INTRODUCTION

The cloud computing is web based climate permits us to utilize software, information and services over the web from any location on any web enabled gadget. Cloud computing gives client another approach to share information assets and that have place with different organizations or destinations**.** It is in any case seen that various spaces being locked in with sharing of clinical information have made the application extremely hard to oversee subsequently the requirement for cloud-based climate which permits communitarian sharing of data over different authoritative areas [1]. Cloud computing gathers the information or data, resources and also provides services to millions of users simultaneously. Data security is the serious issue in cloud computing. Patients these days are higher proponent for his or her own healthcare they are educated to their diseases and increasingly demand access to the most recent technologies. Simultaneously, patients search for the simplest care at the simplest price and are willing to research their decisions.

Subsequently, requests for admittance to non-public patient records are expanding and associations got the opportunity to proceed. By utilizing cloud in medical services quiet information are accessible whenever and anyplace for specialists to dissect and analyze. It has been set up in various scholastic papers that distributed computing offers various advantages going from adaptability, cost viability, spryness improvement of community-oriented sharing of assets [2].

E-health is rising zone within the crossing point of clinical informatics, public fitness and business, regarding fitness offering and information delivered or enhanced through the Internet and related technologies. From a more extensive perspective, the term portrays a specialized turn of events, yet in addition a perspective, a perspective, an attitude, and a dedication for networked, worldwide speculation, to improve medical services locally, territorially, and worldwide by utilizing Information and Communication Technology [3].

The significance of our survey is to collect as much knowledge as viable on how to maintain the security requirements of the cloud-based e-health systems so that this system might be able to storing and transferring the patient health data through a public cloud in a stable and secure manner.

II. CLOUD COMPUTING

Cloud computing is a system for conveying data innovation administrations inside which assets are recovered from the web through electronic devices and applications, as resistance a quick association to a worker. Rather than keeping documents on a restrictive plate drive or local memory gadget, cloud based capacity makes it feasible to abstain from squandering them to a distant data.

However long device approaches the organization, its admittance to the data and furthermore the product bundle projects to run it. It's alluded to as distributed computing because of the information being gotten to found in "the cloud" and needn't bother with a client to be during a particular spot to acknowledge admittance to that.

Cloud Computing Characteristics:

According to the definition, cloud computing has five main characteristics: resource pooling, rapid elasticity, on-demand self-service, broad network access, and measured service [4].

1. **Elasticity:**
The cloud is adaptable and configurable. Customers feel that assets are boundless.
2. **On-demand self-service**:
If necessary, any client can naturally design the cloud without the obstruction of administration professionals.
3. **Shared resources**:
Customers can share assets like organizations, workers, stockpiling, programming, memory, and preparing at the same time. Suppliers can powerfully designate assets as indicated by the vacillations popular, and the customer is totally unconscious of the physical areas of these administrations.
4. **Broad network access**:
The cloud permits a wide admittance to the organization utilizing the Internet from any gadget.
5. **Measured service**:
Diverse cloud administrations can be estimated utilizing various measurements. Itemized use reports are created to save the privileges of clients and suppliers.

Let's see how cloud is used in eHealth:

1) Now a days we are seeing more Electronic Health Records are moved to cloud.
2) Cloud based emails are mostly used in eHealth.

3) Patients and doctors can transfer data between each other from anywhere with the help of cloud for better treatment.
4) Cloud computing permits healthcare platforms to store all the data while avoiding extra costs of maintaining physical servers.
5) Cloud computing gives flexibility to eHealth Departments to increase or decrease their data storage depending on the patients' flow.
6) Cloud computing not only allows its users to access the information remotely, as it includes automation of backups and disaster recovery options which is mostly important for medical data.
7) The mixture of healthcare and cloud has the capability to improve a number of healthcare related functions such as post-hospitalization care plans, telemedicine and virtual medication.

III. LITERATURE SURVEY

In this literature survey we talking about data collection and its security in eHealth. Existing cycles for tolerant clinical information assortment require a lot of work to gather, enter and break down the data. These processes are usually slow and inaccurate, introducing a latency that prevents real-time data accessibility. This situation limits the clinical diagnostics and checking abilities.

I. *A cloud computing solution for patients data collection in health care institutions*

a. Theme: Use of sensors towards existing processes [5].
b. Proposed Method:
Authors actualize an imaginative and minimal effort answer for improve the nature of clinical help conveyance and they address how to incorporate sensors associated with inheritance clinical gadgets which distributed computing administrations to gather, cycle and conveyance patient's crucial information.
c. Experimentation:
Sensors characterized by creators are stacked with programming to gathers, encode, and send information through remote correspondence channels to be put away.
Exchange service is used to access the data which is received from sensors.
Creators used economically accessible remote switches that permit the substitution of the working programming by Linux arrangement.
d. Results / Advantages:
Designed version presents always on real time facts gathering, it removes guide gathering paintings and opportunity of typing errors and it eases the deployment process, as wi-fi networking.
e. Limitations:
There are no administrations upgrades of security and the executives with association of thirty-party framework specialist co-op.

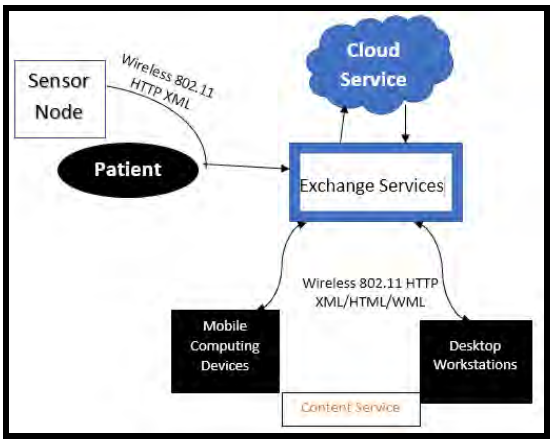


Fig.1. Proposed Solution [5]

II. *DACAR Platform for eHealth Services Cloud*

Theme: Data Capture and Auto Identification Reference (DACAR) project [6].
Proposed Method:
Author has proposed DACAR to develop and implement cloud platform for capture, storage and consumption of data in the eHealth domain.

Experimentation:
Author defined a prototype of the DACAR platform has been implemented using Microsoft .NET 4.0 framework.
In DACAR platform firstly the data capture while data capture DACAR uses radio frequency identification (RFID) it used to identify, authenticate, track and trace medical objects.
Results / Advantages:
It results, the DACAR stage forces little correspondence inertness on application-level messages, and consequently it is adequately proficient to help the turn of events and reconciliation of time basic eHealth applications.
Limitations:
The DACAR model can't enable secure sharing of health care information on a larger scale, and ultimately, to support expert-guided proactive patient-centric health care.
The conceptual view of the DACAR platform is shown in the figure 2.

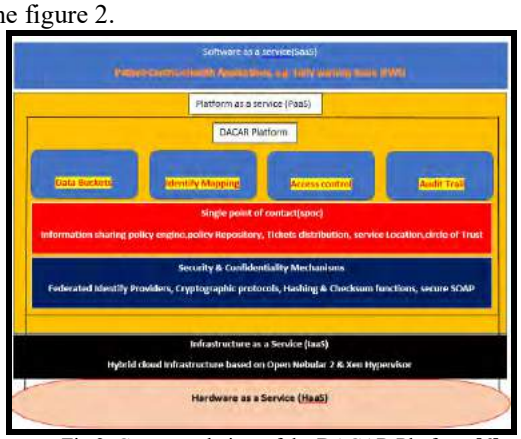


Fig.2. Conceptual view of the DACAR Platform [6]

III. SPS: Secure Personal Health Information Sharing with Patient-centric Access Control in Cloud Computing

a. Theme:
A patient-centric personal health information (PHI) sharing and access control scheme, SPS.[7]

b. Proposed Method:
Author has proposed SPS, a secure and efficient PHI sharing scheme in cloud computing. They also proposed a role based accessed policy named "ESPAC" based on attribute based encryption in [8].

c. Experimentation:
Authors did experiment based on Health-Service Provider (HSP) causes crucial security framework and chooses the cloud service provider where patient's PHI will be stored for future use.
A patient is responsible for defining his/her own attribute-based access policy. Cloud Service Provider (CSP) gives information redistributing administrations and comprises of information workers and administration director. Doctors who request to access stored patient's PHI are identified as Data Access requester (DAR).
Malicious DAR always has an intention to access others' health information to attain some personal benefits.

d. Results / Advantages:
It has been exhibit that proposed conspire is profoundly productive to oppose different potential assaults and malevolent practices.

e. Limitation:
In some cases, the Cloud Service Provider can only store the hashed value of encrypted PHI to save its storage and it can't resist possible inside attack.

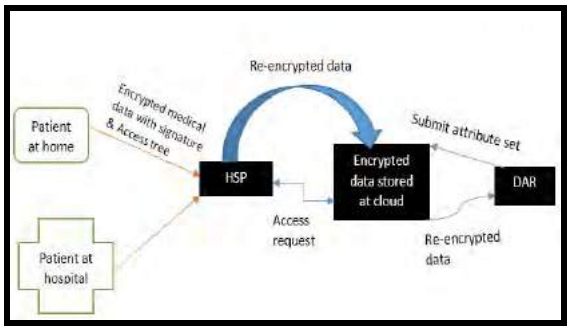


Fig.3. System architecture of proposed SPS [7]

IV. Protecting and Analyzing Health Care Data on Cloud

a. Theme:
Protecting and analyzing Health Care Data on Cloud.
Proposed Method:
Author has proposed a security technology to protect data then they conduct data analysis and lastly they demonstrate feasibility of using cloud in data storage and security [9].

b. Experimentation:
Authors implement security ideas on big data. They use smartphone as a main data producer. They conduct experiment using data mining processes such as linear regression and K-mean clustering with the help of dataset which describes the efficiency of their solution.

References

Data model for their solution is given below in the figure 4.

c. Results / Advantages:
This paper provide feasible and useful solution to secure medical data using cloud.

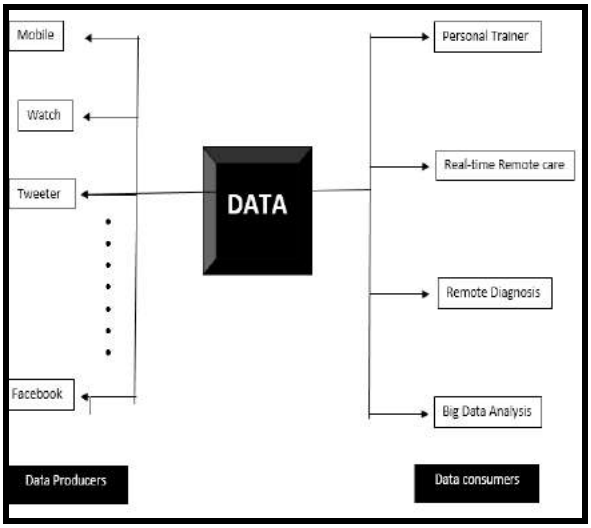


Fig.4. Data Model[9]

IV. DISSCUSSION

In this paper, the different concepts for data security were studied. Paper consist of security techniques using various approaches. Paper started with cloud computing and need of cloud in healthcare and its impact in eHealth.
To do survey we take 4 papers having different approach towards predicting best way for security. From the above papers we conclude that the Cloud Service Provider can only store the hashed value of encrypted Personal Health Information of a patient to save its storage and it cannot defy insider attacks. There is need to change storage of Cloud Service Provider to enhance the security of patients Personal Health Information.

V. CONCLUSION

This literature survey is based on data collection and security challenges in eHealth cloud. This survey illustrates possible all techniques related to security of healthcare information. Cloud computing is dynamically converting our lives in lots of approaches at a honestly brief pace.
The cloud computing arrangements in tending will assist the doctors with remaining in contact with their patients and analyze their ailment adequately at a low cost. It is constantly recalled that cloud computing keeps on being a creating innovation, which suggests that inside the future years the administrations it offers will be bigger than our desires or just on the far side our creative mind.

[1] Zhang R, Liu L. Security models and requirements for healthcare application clouds. In: 3rd IEEE International Conference on Cloud Computing (CLOUD), Miami, FL, USA, USA, pp. (2010).
[2] Abbas, A, Khan, M, Ali, M, Khan, S, Yang, L. A cloud based framework for identification of influential health experts from

Twitter. In: Proceedings of the 15th International Conference on Scalable Computing and Communications (2015), Beijing, China, pp.

[3] G. Eysenbach What is e-health? J Med Internet (2001).

[4] P. Mell and T. Grance, The NIST Definition of Cloud Computing ,NIST, Gaithersburg, MD, USA, 2011.

[5] Carlos Oberdan Rolim, Fernando Luiz Koch, Carlos Becker Westphall, Jorge Werner, Armando Fracalossi, Giovanni Schmitt Salvador Network and Management Laboratory – LRG Federal University of Santa Catarina Florianopolis - SC – Brazil {ober, westphal, jorge, armando, giovanni}@lrg.ufsc.br; fkoch@acm.org

[6] L. Fan, W. Buchanan, C. Th¨ummler, O. Lo, A. Khedim, O. Uthmani, A. Lawson Faculty of Engineering, Computing & Creative Industries Edinburgh Napier University, Edinburgh, UK L.Fan@napier.ac.uk

D. Bell NIHR CLAHRC for Northwest London Imperial College London, UK D.Bell@imperial.ac.uk .

[7] Mrinmoy Barua, Rongxing Lu, and Xuemin (Sherman) Shen Department of Electrical and Computer Engineering University of Waterloo, Waterloo, Canada, pp.

[8] M. Barua, X. Liang, R. Lu, and X. Shen, "ESPAC: Enabling security and patient-centric access control for ehealth in cloud computing," International Journal of Security and Networks, vol. 6, no. 2/3, pp.67–76, Nov 2011.

[9] Danan Thilakanathan∗ †, Yu Zhao∗ , Shiping Chen∗ †, Surya Nepal†, Rafael A.Calvo∗ and Abelardo Pardo School of Electrical and Information Engineering, the University of Sydney, Australia† Digital Productivity & Services, Commonwealth Science Industry Research Organization (CSIRO), Australia Corresponding, pp. Email: Danan.Thilakanathan@sydney.edu.au.

Convergence Of Machine Learning And Blockchain For Securing Future Of Internet Of Things

Darshana Borse, Nikita Hire, Dnyanal Gavale,Ashish Awate
Computer Engineering,SVKM's Institute of Technology Dhule, India
darshanaborse687@gmail.com,nikitashire@gmail.com,dnyanalgavale@gmail.com,ashish.awate87@gmail.com

Abstract: **Recently, IoT Technology has become a part of our day to day lives it is known to be smart things for smart homes over internet. In addition to being everywhere, in almost every face of life, IoT devices are being actively used. As the number of devices connected in an IoT network is usually very large, so it is not easy to secure from getting cracked. Blockchain and Machine learning has emerged as the possible solution for creating more secure IoT systems in future. In this paper, we explained about an overview of the Blockchain technology and Machine Learning with its implementation; In last we have proposed the model to secure IoT by integrating Blockchain network and machine learning algorithm.**

Keywords: Internet of Things (IoT), Machine Learning, Blockchain Technology.

I. INTRODUCTION

The Internet of Things (IOT) is a network to interconnect computing devices which have the ability to transfer data through network, without any human interaction. There is tremendous growth in technologies in past decades, which has resulted the increased use of IoT devices. To better understand in terms of scale, Till 2020, the use of IoT devices had reached to 50 billion. These numbers are expected to increase to 125 billion by 2030.

Recently, Internet of Things (IoT) has received much attention in many fields; its applications are widespread in various domains. When IoT technologies started to be developed by connecting small devices equipped with sensors, there was no serious consideration on the security issues. However, as IoT technology advances and many devices are connected to exchange private and sensitive information, security problems became a major concern. Each layer of the architecture of IoT has its own security challenges and researchproblems. Due to many reasons specific to IoT, providing the security services for IoT is a very challenging task.

As the explosive need of IoT devices results in security challenges. So this paper aims, to provide solutions for these security and privacy concerns by combining machine learning algorithms and Blockchain techniques. As review of ML algorithms and BC techniques employed to protect IOT applications from security and privacy attacks. Based on the review, we highlight that a combination of ML algorithms and BC techniques can offer more effective solutions to security and privacy challenges in the IOT environment. [2].

Blockchain and machine learning (ML) are considered as promising technologies to support secure and sharing of information and model as well as the intelligent network operation and management. In this paper, we specialize in Blockchain and ML, which have a major potential to promote the event of communications and networking systems [3]

II. OVERVIEW OF BLOCKCHAIN

Blockchain is paradigm that consists of a distributed ledger which contains all transactions ever executed within its network, enforced with cryptography and administered collectively by peer-to-peer nodes. Blockchain technology could be a Google Doc. once we create a document and share it with a group of individuals; the document is distributed rather than copied or transferred. Thus, Blockchain allow us to possess a distributed peer-to-peer network where non-trusting members can interact with one another without a trusted intermediary, in a cryptographically verifiable manner. Blockchain consists of multiple blocks, nodes and miners.

1. Blocks: Every chain consists of multiple blocks and every block has three elementsas data, nounce, and hash.
2. Miners: It creates new blocks on the chain through a process called mining.
3. Nodes: it's kind of device that maintains copies of the Blockchain and keeps the network functioning.

A. Implementation of Blockchain

In three domains Blockchain can be deployed:

1. *Public*: In this domain each and each node can send or read transaction and may participate within the consensus process without the requiring any permission.
2. *Consortium area*: In this domain, only defined nodes can participate within the consensus process. The permission to read or send could also be made public or could also be provided only too few authorized nodes.
3. *Private*: In this area, only the organization to whom the network of Blockchain belongs can write transaction to it. Reading of transaction could also be public or restricted to few nodes depending upon the

requirement. This sort of system is usually deployed in industries. [5]

A. Ways to secure IoT using Blockchain

Blockchain Applied in IoT involves the ubiquitous interconnection of various devices with networking and computing abilities, provides a promising way for data collection, analysis, and sharing. Blockchain are often used to secure IoT in such ways:

1. *Secure communication:* IoT devices need to communicate for the aim exchanging data required to process a transaction and to store it during a ledger. IoT device sends an encrypted message using the public key of the destination device, which is then stored within the block-chain network then asks its node to urge public key of the receiver from the ledger. Then they encrypts the message using public key of the receiver, during this way, only the receiver will be ready to decrypt the sent message using their private key.
2. *Authentication of users:* The sender digitally signs the message before sending them to other devices. The receiving device then gets the public key from the ledger and uses it to verify the digital signature of the received message.
3. *Discovering legitimate*: IoT at large scale: With potentially lots of IoT devices are to be connected on an equivalent network, there's an urgent need to get the flexibility to get devices at scale and to discern legitimate and illegitimate nodes. It receives information from other nodes and sends its information to other peers on network.
4. *Configuring IoT*: Blockchain technology helps a lot in establishing a trusted and secure configuration for IoT devices. Approaches that seem relevant here are:
   a) Properties of IoT like Configuration details and the last version firmware validated are often hosted on the ledger. During bootstrap, the Blockchain node is asked to urge its configuration from the ledger. The configuration is required to be encrypted in the ledger to prevent the invention of IoT network topology or its properties by analysis of the information stored in the public ledger.
   b) The hash value of latest configuration file for every device can be hosted in the ledger.

III. OVERVIEW OF MACHINE LEARNING

According to E.Tom Mitchell, definition of machine learning is "a computer program said to be learn from experience E with respect to some task T and some performance measure P, if the performance on T, as measured by P, improves experience E". The machine learning is a gripping tool for providing solutions to the problems and enhances the performance of the developed system based on data sets [3]. The training sets are fed to a learning algorithm to produce a trained "machine" that carries out the desired task. Learning is made possible by the choice of a set of possible "machines", also known as the hypothesis class, from which the learning algorithm makes a selection during training [4].

A. Classification of Machine Learning

Machine Learning consists of three types of learning methodologies

1. *Supervised Learning*: In this learning, the training set consists of for every data instance has the input x and the corresponding output y. Supervised learning consists of two types Classification and Regression. Algorithms of supervised learning for classification are nearest neighbour, neural network, Support Vector Machine, Bayesian's theory and for Regression are linear regression, decision tree, and neural networks.
2. *Unsupervised Learning*: In this learning there is only input x there is no label or output y present to the data. Here the similar instance are grouped or clusters together. k-means, self-organizing maps, and anomaly detectors, fault detection, network operational configuration, energy efficiency management, Gaussian mixture, Neural Networks these are some algorithms of unsupervised learning.
3. *Reinforcement Learning*: Reinforcement learning is to take suitable action to get maximum rewards. Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there's no answer but the reinforcement agent decides what to try to perform the given task. In the absence of coaching dataset, it's sure to learn from its experience.

B. Security Threats in IoT

IoT may be introducing a lot of benefits to our modern life but it also has one major drawback i.e. security threat. IoT security risk could even more significant on user side where they often not aware of potential threats. IoT security threat can be used to steal critical data from people as well as organization. Attacker can exploit security vulnerability in IoT infrastructure to execute sophisticated cyber- attack. Following are some most used security threats.

Denial of Service – A Denial of Service (DoS) attack deliberately tries to cause a capacity overload in the target study by sending multiple requests. Attacker who implement DoS attacker don't have intention to steal data however it can be used to slow down or disable service. Even for seconds of halt of security service can be huge threat. One of the most dangerous types of a DoS

attack is when DDoS attackers use thousands of Internet protocol addresses to request IoT services, making it difficult for the server to distinguish the legitimate IoT devices from attackers.

Man in the middle attack – In this attack, attacker breach into the communication channel between two systems in attempt to intercept the message from them. Attacker gain control over their communication and send illegitimate message to participate system. Such attack used to hack IoT devices in home or organization.

Jamming attack – Attackers send fake signals to interrupt the on-going radio transmissions of IoT devices and further deplete the bandwidth, energy, central processing units (CPUs), and memory resources of IoT devices or sensors during their failed communication attempts

C. *Need of Machine learning in IoT*

Machine Learning is categorized in three ways that is supervised learning, unsupervised and reinforcement learning. Supervised learning works better when we know the environment variable i.e. output to correspond with every input. We use unsupervised learning; mainly this learning is used to categorize the characteristics. On other hand reinforcement learning is different from these two learning, in this learning reinforcement agent decides from their own positive or negative experience. Machine Learning consist many algorithms which can be effectively used to secure IoT devices. To distinguish normal IoT packets following algorithms can be used:

1. K-nearest neighbors algorithm (KN)
2. Support vector machine with linear kernel
3. Random Forest using Gini impurity scores(RF)
4. Neural Network (NN)

IV. LITERATURE SURVEY

Table I summarizes the different machine learning, Blockchain, ensemble model and integrations of Blockchain and Machine Learning research papers. It specifies the methodology, gaps and approaches used to secure IoT. There are still many challenges in securing IoT because convergence of machine learning and Blockchain should have high level of computational complexity to be suitable for the resource limited IoT devices

.

Paper No.	Author	Theme	Methodology	Limitation
1	Ankur Jaiswal .elt (2016)	Machine learning help to prevent DDOS attack.	If the utilized resource is more than the given threshold value, the incoming packets is identified as anomalous. A traffic monitoring node which continuously monitor the packets from different network layers.	In this paper the system is not fully automated, so that very little human intervention is required for monitoring, detecting and preventing the attack.
2	Aissam Outchakoucht .elt (2017)	Dynamic access control based on ML and Blockchain in the IOT	This paper aims on the concept of the Blockchain and ML algorithms to ensure a totally distributed infrastructure and to provide a dynamic, optimized and self-adjusted security policy using Reinforcement Learning.	Blockchain technology has some intrinsic drawbacks especially when talking about privacy, required time for block validation
3	Rohan Doshi .elt(2018)	DDoS detection in IoT network traffic with a variety of machine learning algorithms.	This paper author works on the Stateless Features of the packets and according to that features they classify the packets as a normal packet or malicious packet with five machine learning algorithms	The linear SVM classifier performed the worst, suggesting that the data is not linearly separable.
4	NAZAR WAHEED .elt(2020)	Detailed review of ML algorithms and Blockchain techniques	In this paper, author have reviewed the latest threats to IOT and proposed solutions for each layer of IoT architecture using Block Chain techniques and Machine Learning algorithms.	Author didn't design and developed a privacy-preserving IoT framework, which should also offer privacy-preserving data sharing as well as privacy-preserving data analysis.
5	Rahul Agrawal .elt(2018)	Continuous security in IoT using Blockchain	In this paper Continuous security is established in the system with seamless user authentication using crypto-tokens.	Crypto-token generation can be improvised by using an ensemble learning approach which uses the best models specific to the input data or a weighted combination thereof.

Table I. Survey Table

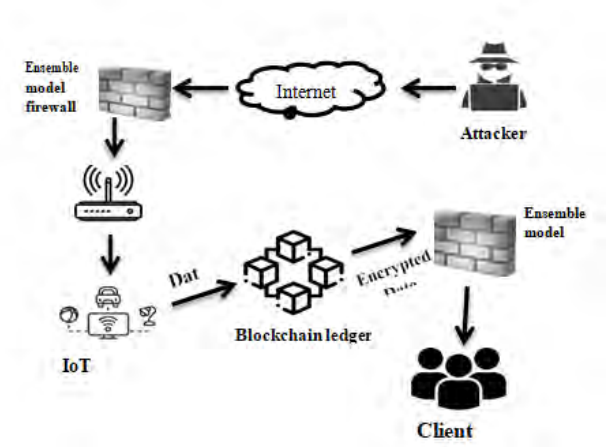
- *Machine Learning Related research*

Paper	Attack Learning	Approach	Accuracy
Machine Learning DDoS Detection [Rohan]	DDoS Attack	KN-	.999
		LSVM	.991
		DT	.999
		RF	.999
		NN	.999

Table II. Comparing ML approaches

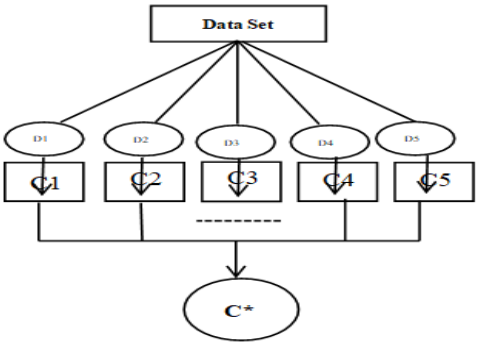
Table II summarizes the different machine learning approaches used to secure IoT. There are still many challenges in securing IoT because machine learning should have high level of computational complexity to be suitable for the resource limited IoT devices.

V. PROPOSED METHODOLOGY



The system overview of secure Ensemble model is provided above i.e. fig1. Each IoT data provider pre-processes IoT data instances, encrypts them locally using their own private keys, and records them in a Blockchain-based shared ledger by generating transactions. Existing Blockchain mechanisms can be employed to manage the encryption capabilities of data providers. Then this encrypted transaction of data is passed through the ensemble firewall model. The data analyst who wants to train an Ensemble model can get access to the encrypted data recorded in the global ledger, and assemble a secure training algorithm with several building blocks, such as secure comparison, secure polynomial addition. During the training process, interactions between the data analyst and each data provider are necessary for exchanging intermediate results.

The ensemble model works on the mechanism of voting. For example first we give 1 instance to all the algorithms and they predict whether that instance belongs to class 0 or class 1. If the majority algorithms classify that instance in class 0 then it belongs to that class otherwise class 1. As Bagging is the best famous representative of parallel ensemble learning methods.



- Blockchain Model

In order to store the encrypted IoT data in the Blockchain, we define a special transaction structure. The transaction format has two fields: input and Output which is mainly important. The input field contains the data provider's address, the encrypted data, and the IoT device type. While, the output field contains the data analyst address, the encrypted data, and the IoT device type. The ledgers are used to store the encryption keys and data is made more confidential. First sender digitally signs the message then creates its hash using its private key. Then sender ask ledger for the public key of receiver and encrypt the whole message. Now the sender sends data through network. After passing through firewall, on receivers' side it decrypts by digital signature using public key of sender that is stored in ledger. The transaction is valid only when the calculated hash and protected hash is same. If user is valid then receiver decrypts message using its private key.

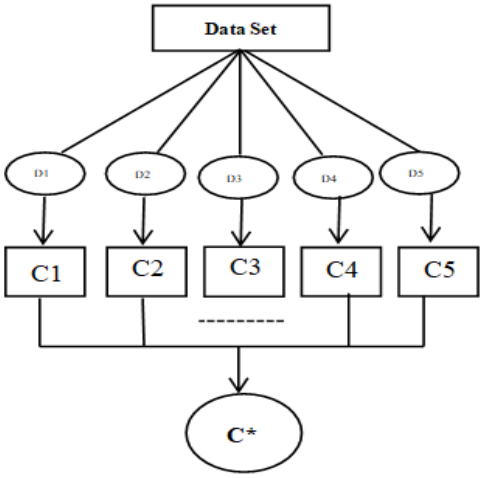
- Packet Detection

Traffic Capture: The process records the source IP address, destination IP address, packet size, and timestamp of all IP packets sent from smart IoT devices as well as from client to IoT devices .

Feature Extraction: For each packet stateful and stateless feature are generated. In which stateful feature has bandwidth and destination IP address while stateless consist packet size and protocols

- Machine Learning Model

After these, An ensemble learning model is introduced in which we used bagging model with all five algorithms like Random Forest, Nearest Neighbour (NN), Support Vector Machine (SVM), Decision Tree and Naïve Bayes, respectively.. Then, we will train bagging model. That is, we have big data set in this data set first we will randomly select the instances with replacement and create new 'n' data set. Then classify these n data set using 'n' classification algorithms i.e. classifiers. After that we will combine all the output of the existing classifiers to create strong classification model.Fig2 .Bagging Technique



In fig2, C1 is SVM classifier, C2 is Random Forest classifier, C3 is K-NN classifier, C4 is DT classifier, and C5 is Naïve Bayes classifier. And D1 to D5 are data sets.

CONCLUSION

Due to many reasons specific to IoT, providing the security services for IoT is a very challenging task. As many IoT platforms are different from each other so that it is very difficult to design universal and homogeneous security systems that can be applied to different IoT platforms. So, we introduce model which include Blockchain ledger and Ensemble model in Machine learning algorithm to get strong model. This paper attempts to briefly explore the technologies related to the convergence of Blockchain and ML.

References

[1] Hyungsik Shin ; Ho Kyoung Lee ; Ho-Young Cha ; Seo Weon Heo ; Hyungtak Kim 2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)

[2] Nazar Waheed, Xiangjian He, Muhammad Ikram, Muhammad Usman, Saad Sajid Hashmi, and Muhammad Usman.2020. "Security and Privacy in IoT Using Machine Learning and Blockchain: Threats and Countermeasures". ACM Comput. Surv. 53, 3, Article 1 (April 2020).

[3] Yiming Liu ; F. Richard Yu ; Xi Li ; Hong Ji ; Victor C. M. Leung "Blockchain and Machine Learning for Communications and Networking Systems" Year: 2020 | Volume: 22, Issue: 2 | Journal Article | Publisher: IEEE

[4] Osvaldo Simeone, Fellow, IEEE (Invited Paper) "A Very Brief Introduction to Machine Learning with Applications to Communication Systems" IEEE Transactions on cognitive communications And networking , vol. 4, NO. 4, December 2018Madhusudan Singh; Abhiraj Singh; Shiho Kim "Blockchain: A game changer for securing IOT data" 2018 IEEE 4th World Forum on Internet of Things (WF-IoT) Year: 2018 | Conference Paper | Publisher: IEEE

[5] Rohan Doshi; Noah Apthorpe; Nick Feamster "Machine Learning DDoS Detection for Consumer Internet of Things Devices" 2018 IEEE Symposium on Security and Privacy Workshops Year: 2018 | Publisher: IEEE

[6] Meng Shen; Xiangyun Tang, Liehuang Zhu; Xiaojiang Du ; Mohsen Guizani "Privacy-Preserving Support Vector Machine Training Over Blockchain-Based Encrypted IoT Data in Smart Cities" 2019 IEEE Internet Of Things Journal

[7] Bin Jia; Xiaohong Huang; Rujun Liu; Yan Ma; "A DDoS Attack Detection Method Based on Hybrid Heterogeneous Multiclassifier Ensemble Learning" 2016 Research Article

[8]Rahul Agrawal, Pratik Verma, Rahul Sonanis, Umang Goel, Dr. Aloknath De,Sai Anirudh Kondaveeti, Suman Shekhar "CONTINUOUS SECURITY IN IOT USING BLOCKCHAIN" Bangalore. Year: 2018 | Publisher:IEE

Credit Card Fraud Detection Using Machine Learning Algorithm

Hitesh Nikam,Kirtish Wankhedkar,Nishant Patil,Uddhav Sharma,B.R.Nandwalkar
Dept of Computer Engg.Shri Vile Parle Kelavani Mandal's,Institue of technology,Dhule
hiteshnikam021@gmail.com,wankhedkarkirtish19@gmail.com,patilnishant222@gmail.com,uddhavsharma786@gmail.com,nandwalkar.bhushan@gmail.com

Abstract--**According to the world's current scenario, we can quickly notice that all the countries are highly expanding on digital platforms. All of the sectors are maximum dependent on the internet. The superiors of the country are also working for the development of the country in technological aspects. The amount of cashless transactions taking place all over the internet is very high. Due to this, the risks of fraud and theft are also increasing. Financial scams are taking place in an excessive number. The data of credit cards people use during the transaction is getting stolen and is used to commit fraud. In this paper, there are some techniques discussed through which the credit card frauds can be detected. The objective of this study is to choose the best machine learning algorithm for credit card fraud detection.**

Keywords—Local Outlier Factor, Isolation Forest, Random Forest, Support vector machine, Logistic Regression, Decision Tree, K-Nearest Neighbor

I. INTRODUCTION

In Today's growing technological world, the growth of the e-commerce sector has experienced tremendous growth because most people find it an efficient way to purchase things. The e-commerce offers a variety of payment options like net banking, credit card, and various online transaction apps; as the e-commerce users are increasing, the fraud in online transactions is also growing, and the major one is the credit card fraud. To commit fraud in these varieties of purchases, a fraudster gathers the card details. Most of the time, the real cardholder isn't aware that somebody else has seen or taken his card information and using it for their own benefit.

II. LITERATURE SURVEY

[1] Pawan Kumar and Fahad Iqbal proposed different system techniques of fraud detection, user-friendly and secure. The system examines the achievability of credit card fraud detection based on outlier mining, applies outlier detection mining based on distance sum into credit card fraud detection and proposes this detection procedure and its empirical process. They approach three algorithm Isolation forest, Local outlier factor and Support vector machine. Their observation was Isolation Forest had detected 73 errors while Local Outlier Factor has detected 97 errors along with SVM detecting 8516 errors. Isolation Forest features a 99.74% correct than LOF of 99.65% and SVM of 70.09. So, the Isolation Forest performed far better than the Support vector machine and slightly better than Local outlier factor.

[2] Maja puh, Ljiljana Brkić proposed a paper for fraud detection in credit card transactions by focusing on three machine learning algorithms Random Forest, Support Vector Machine, Logistic Regression. They applied two learning approaches, static and incremental learning, on chosen algorithms. In a static method, training and testing were done once using all datasets. For incremental learning, divided data into two chunks, and training and testing were done on each data chunk separately, which generated two models. For Random Forest, the number of trees is set T=100. For SVM, use Gaussian radial basis function as the kernel was cost parameter c set to 10, and gamma was 0.01. For logistic regression, we set C to 100 and used L2 regularization. And observation is given in table 1.1

Table 1.1 Algorithm accuracy

Static learning	Random Forest	0.9148
	SVM	0.8877
	Logistic Regression	0.9114
Incremental learning	Random Forest	0.9013
	SVM	0.8678
	Logistic Regression	0.9107

From the observation, they conclude that SVM shows the most inadequate performance, and the difference between the performance of Logistic Regression and Random Forest is slight.

[3] Imane SADGALI, Nawal SAEL, Fouzia BENABBOU focus on that results and try to compare the same dataset. Their research investigated a comparative study of data mining techniques on the same generated dataset. Their task is done on a generated dataset, containing approximately 60.000 transactions across 12 attributes. These attributes include transaction and client information. There is a significant imbalanced data in the dataset where 99.72% of transactions are of the non-fraudulent class. The applied machine learning algorithm was Decision Tree, Support Vector Machine, Random Forest, K-Nearest Neighbor.

Table 1.2 Algorithm accuracy

Supervised Learning technique	Accuracy
Decision Tree	78.9%
Support Vector Machine	99.7%
Random Forest	82.5%
K-Nearest Neighbor	97.1%

And by this observation table1.2, they conclude that performance of four supervised machine-learning techniques, decision tree, k-nearest neighbor, random forests and support vector machines, for credit card fraud detection. Support vector machines have proven to be the best of the others.

[4] The author uses the latest machine learning algorithms to detect anomalous activities, called outliers. First, they obtained the dataset from Kaggle, a data analysis website which provides datasets. Inside the dataset, there were 31 columns out of which 28 are named as v1-v28 to protect sensitive data. The other feature represents Amount, Time and Class. The amount is the amount of money transacted. Time shows the time gap between the first transaction and the following one. Class 0 represents a valid transaction, and 1 illustrates a fraudulent one. This data is fit into a model, were Local Outlier Factor and Isolation Forest Algorithm are applied to it.

[5] Phuong Hanh Tran, Kim Phuc Tran, and other author proposed two machine learning approaches for credit card fraud detection using one-class support vector machine (OCSVM) with the optimal kernel parameter selection and T2 control chart. They investigate the performance of the algorithms on real data set of online e-commerce transactions from European credit cardholders, and they conclude that the OCSVM method outperforms to T2 control chart in all fields of comparison, has an Accuracy percentage of 96.6%, FPR of 8.5%, and F-score of 100%. Were the T2 control chart method with the fast calculation time is still very useful in detecting fraud; it has an Accuracy percentage of 93.6%, FPR of 16%, and F-score of 100%.

As we go through all paper, we have seen that Isolation forest is the best method for detecting fraud. In this paper, we are explaining the specific data attribute and working of the Isolation forest model for detecting fraud cases.

III. CHALLENGES IN CREDIT CARD FRAUD DETECTION

A) Data deficiency:

Due to the issue of personal data, the biggest problem in dealing with Credit card fraud detection is that real data is hardly ever available for training and exploration.

B) Imbalanced data:

As mentioned earlier, on the global level, the fraudulent transaction has amounted to less than 0.05% of the total transactions. They are resulting in highly imbalanced classes. If this issues had not been taken into deliberation, any machine learning algorithm that classifies correct only genuine transactions would perform outstanding, with an accuracy level above 99%, disregarding the fact that all the minority class transactions are classified falsely.

C) Behavioral variation:

Fraudulent behavior tends to alter over time to avoid detection. Therefore, the CCFD predictive model should not be static, i.e., constructed once and never updated

D) Cost-sensitive problem:

CCFD is a cost-sensitive problem, meaning that the cost produced by the genuine misclassifying transaction (false positive) is different than the cost of misclassifying fraudulent one (false negative).

IV. WORKING OF MODEL USING ISOLATION FOREST ALGORITHM

A) Isolation Forest Algorithm:

Isolation forest work on the essence of isolation anomalies also isolation forest algorithm is an unsupervised learning algorithm. The Algorithm recursively generates separations on the sample by arbitrary selecting an attribute and then randomly sort out a split value for the attribute, between the minimum and maximum values allowed for that attribute. So, whenever the random forest trees simultaneously produce shorter path lengths for particular samples, they are highly reasonable to be anomalies. An advantage of this Algorithm is that it works with a vast data set and several dimensions benefits of Anomaly Detection Using Isolation Forests.

B) Data Preprocessing

For handling class imbalanced, we have chosen the SMOTE method. The sampling strategy parameter, representing the desired ratio of the number of samples in the majority class over the number of models in minority class, may vary. The value used in experiments equals to 0.4 (the number of minority samples is increased to 40% of majority class size). It was determined empirically by choosing the best performing value among various attempts[2].

Features transformed using PCA (Principal Component Analysis) for dimensionality reduction. Because of the unknown meaning of the original or the constructed features, there was not enough information to create additional features[2].

. The chosen features, from a financial institution database:

• Amount: the amount of transaction hold in bank currency

• Transaction type: national, international or e-commerce.

• Date and time: date and time of the transaction have been in YYMMDDHHMMSS format.

• Transaction channel: the channel of the incoming transaction (ATM, electronic payment, merchant application or)

• Billing address and Shipping address: address of billing with shipping for the customer and address of merchant or point of service for the purchase, address of ATM for withdrawal.

• Merchant group: transport, food, cloth, ticket, etc…..[3].

C)Design System:

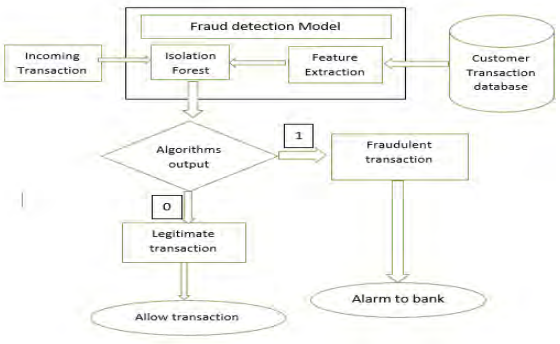


Fig 1 Fraud detection model

V. CONCLUSION

We present in this paper real-time data-driven fraud detection approaches using Isolation Forest. This Algorithm proved accuracy and efficiency in detecting out fraudulent transactions and minimizing the number of false alerts. The use of this Algorithm in a credit card fraud detection system results in detecting or predicting the fraud probably in a concise period within real-time. This will eventually prevent the banks and c from great losses and customer satisfaction with the bank services.

VI. ACKNOWLEDGMENT

The authors are thankful to Prof.Bhushan.R.Nandwalkar, for their encouragement, support and guidance for successful completion of the paper.

References

1. Pawan Kumar and Fahad Iqbal, "Credit card Fraud Identification Using Machine Learning Approaches",IEEE 2019, of Saveetha unviersity.
2. Maja Puh,Ljiljana Brkic, "Dectecting Credit Card Fraud Using Selected Machine Learning Algorithms",IEEE 2019 ,univsersity of zagreb faculty of electrical enginnering and computing MIPRO,May 20-24.
3. Imane Sadgali ,Nawal Sael,Fouzia Benabbou, "Fraud detection in credit card transaction using machine learning techiniques",IEEE 2020,University Hassan II Casablanaca.
4. S P Maniraj,Aditya Saini ,Swarna Deep Sarkar,Shadab Ahmed, "Credit Card Fraud Detection using Machine Learning and Data Science",IJERT,Vol. 08,issue2019.
5. Phuong Hanh Tran, Kim Phuc Tran, Truong Thu Huong, Cédric Heuchenne, Phuong HienTran, Thi Minh Huong Le "Real Time Data-Driven approaches for Credit Card Fraud Detection",ICEBA 2018.

Review On: Applications Of Augmented Reality

Mayank Gindodiya , Tejas Bhavsar, Uzair Shaikh, Ashish Awate
Computer Engineering,S.V.K.M's IOT, Dhule, India
gindodiyam@gmail.com,tbhavsar121gmail.com, uzair11120007@gmail.com, ashish.awate87@gmail.com

Abstract— **Augmented Reality (AR) technologies for supporting various applications have been an academic research topic for around many years now. In the last 10 years, AR technology is getting used for solving industrial problems. In this paper, applications of AR have been explored.**
This paper aims to show, through the results of a systematic literature review, the current state of AR technology in industries and different applications and the advantages of using AR technologies. In this paper a total of 5 papers are discussed which solely describes applications of the AR technology. The paper also gives a brief history of the AR and an illustrated discussion of the same. Paper also proposes an idea of overcoming some limitations of the AR to make it more usable in today's world

Keywords— Augmented Reality(AR), Mobile Device, 3D object manipulation, device-based interaction technique, shopping assistance by AR,
pokemon.

I. INTRODUCTION

Augmented Reality: augmented reality (AR) refers to all cases in which the display of an otherwise real environment is augmented by means of virtual (computer graphic) objects.
The commonly accepted definition of AR as a technology:

1. Combines real and virtual imagery.
2. Is interactive in real-time and Registers the virtual imagery with the real world in 3 dimensions (3D).

Augmented Reality allows the user to interact with the system in a more efficient manner. This enhances the user experience (UX) and the growth of the system. Augmented reality is now being implemented on mobile computing devices that include digital cameras. In such implementations, the view that is currently being captured by a camera can be displayed as a scene on a screen of the mobile device, and data about items that are shown in the scene may have textual annotations added to them. Non-visible objects may also be represented by annotations. The applications of Augmented Reality include many fields of multi-diverse measure. Examples of Augmented Reality include Archaeology, Architecture, Urban design and planning, Industrial Manufacturing, Commerce, Literature, Visual Art, Immersive Video Gaming. Within recent trends, Augmented Reality has revolutionized social media platforms with recent major groundbreaking innovations like LIDAR (Light Detection and Ranging). Although, the hardware component requirements are often a variety of complicated, fragile, and delicate electronic components like MEMS (Microelectromechanical Systems), Sensors like accelerometer, GPS (Global Positioning System), Solid State Compass, Gyroscope, Camera system, HUD (Head-up Display), HMI (Human Machine Interface), VRD (Virtual Retinal Display). Despite its implementation complexity Augmented Reality shows itself as a promising contender for mainstream media consumption channels.
[6] The history of the AR is described in points along with its year.

1. In the late 1950s, Morton Heiling developed a simulator called "Sensorama".
2. In 1962, Ivan Sutherland created "Sketchpad", the first computer graphic user interface. Also in the same year first HMD was patented by Morton Heilig, but never produced(Biocca & Levey, 1995).
3. In 1966, Ivan Sutherlan developed "Ultimate Display" with a cathode-ray tube screen (Ivan Sutherland Biography, 2009).
4. In 1975, Myron Krueger created an artificial reality lab which was called 'videoplace'(now called Hamilton, 2011).
5. In 1980, Steve Mann developed wearing computers (Hamilton, 2011).
6. In 1989, Jaron Lainer coined the term Virtual Reality or VR.(Hamilton, 2011)
7. In 1990, Tom Caudell coined the term 'Augmented Reality' or AR (Hollerer and Feiner, 2004).
8. In 1992, L. B. Rosenberg developed 'Virtual Fixtures', one of the first functioning AR systems (Rosenberg, 1993).
9. In 1998, Ramesh Raskar, Greg Welch, and Henry Funchs introduced 'Spatial Augmented Reality' to UNCG (Raskar, welch and Funchs, 1998).
10. In 1999, Hirokazu Kato developed the first ARToolKit in Japan(Kato and Billinghrust, 1999).
11. In 2000, Bruce Thomas created the first outdoor mobile AR game called 'ARQuake'.(Thomas et. al. 2001).
12. In 2008, Wikitude released the AR Travel Guide[6].
13. In 2009, Esquire magazine collaborated with Robert Downey Jr. and used AR in their magazine to give AR experience in reading magazines.
14. In 2013. the company Volkswagen used Augmented Reality to show its car manuals.
15. In 2014, Google made available the google glass to the market, which is an example of wearable augmented reality.
16. In 2016, Niantic Inc. released Pokemon GO that uses Augmented Reality to view and catch Pokemon characters.
17. In 2018, IKEA launched an app that can preview the customers' home decor items before the actual purchase.

II. LITERATURE REVIEW

Augmented Reality-Based Learning Environment to reinforce Teaching-Learning Experience In Geometry Education.[1]: As suggested by the title the paper focuses on better teaching the geometry subject using augmented reality. Visualization supported by a computer can play a really important role in every field because it is often used very smartly to beat the limitations of the normal teaching methods. This technology is taken into account in the realm of science and arithmetic classroom and supports theoretical underpinnings in understanding the advantages also as limitations of augmented reality-based learning environment (ARLE) experiences. one of the topics which are difficult for The scholars to know is geometry in mathematics education. to deal with the matter, the paper describes a framework of mobile-based ARLE systems[1]. During this paper, 3D Geometry was chosen from mathematics for implementation. The target of such an application is to assist the scholars in achieving LTM retention although they need different learning and retaining abilities.

The paper concludes that AR can help understanding concepts of geometry and its three-dimensional space in a much efficient manner.

3D Object Manipulation Techniques In Handheld Mobile Augmented Reality Interface.[2]: The paper describes how the virtual objects in augmented reality are manipulated and interacted with in different ways on a handheld system. The AR has gained much popularity in mobiles and tablets due to consumer-oriented communication products nowadays, especially touchscreen smartphones[2]. The paper suggests some of the methods interact with it, they are-

1. Touch-based interaction in which the objects are manipulated using the touch screen of the smartphone.
2. Mid-air gesture-based interaction in which the camera is used to track the gestures of hand and finger to manipulate the object associated with the AR.
3. Device-based interaction uses a special device to communicate with the object. An example would be using motion sensors in the mobile to play games or using a motion control device to play cricket or tennis on a Playstation.

The paper concludes with the techniques and its advantages and disadvantages, describing if it's practical or not to use AR and such techniques in today's world. Also Currently, many of the existing handheld mobile AR applications are not considered very practical due to insufficient functionality and they do not fully answer to the needs of the users.

Prospects Of Augmented Reality In Physical Stores Using Shopping Assistance App[3]: This paper describes the utilization of AR in various aspects of retailing. The previous research works limit itself to enhancing the user experience. It includes a GPS tracking system that defines itself in locating a specific person or a product inside a building. This research suggests combining in-store navigation using AR and therefore the ability of the app to cater to the requirements of the purchasers and acts as a customized shopping assistant. The paper introduces a model(mobile application) of AR which will assist the patrons by navigating them to the list of products that the user wants to locate for e.g., during a mall, with the utilization of in-store GPS that uses wifi, RFID, INS and AR image capturing technologies. within the sort of augmented imagery, shoppers also can scan a specific product or a whole rack using their mobile camera and may get various information just like the specifications of the products, multiple promotional offers available.[3] The shopper's experience is enhanced because they will make use of both physical store experience combined with e-assistance on a real-time basis.

This paper concludes that the shopping assistant app utilizes augmented reality technologies to provide personalized advertising and instore shopping assistance like in store navigation, customized product-specific information for the shoppers.

Pokemon Fight Augmented Reality Game[4]:

The Pokemon Fight card Game is predicated on the thought of cards related to a specific Pokemon character which may be wont to fight. The sport uses augmented reality technology to bring Pokemon into the important world. Each Pokemon has one attack move. The Pokemon character can attack the opponent's Pokemon with its attack move. The sport is often played by two players at a time. Using augmented reality makes the sport more fun and enjoyable to the players[4]. This enriches the gaming experience of the user. This technique captures and scans the cardboard and compares the scanned card with database images and loads the precise pokemon model as per the matched image. This leads to a game console to display the result.

A tour guiding system of historical relics based on augmented reality[5]: a game-based guidance system for Yuanmingyuan and a time travel game called MAGIC-EYES has been proposed with Augmented Reality technology. Six interactive modes are designed in the proposed system to guide tourists to visit the specified place[5]. MAGIC EYES also makes use of such sensors in mobile phones as cameras, gyroscope, and global position system (GPS) to identify the images of recognizable objects, the viewing direction of tourists, and the geographical location information[5]. As the test was conducted by the staff, they created two groups and they provided a guidance system to group 2 only. The results of the test suggest that Group 2 achieves a stronger authentic feeling of historical legend about Yuanmingyuan than Group 1.

III. DISCUSSION ON SURVEY

The paper illustrates the use of AR, methods of interaction with AR objects, and use of AR technology in various aspects of business, education, tourism, etc. apart from this there exist many applications of AR which is not described in the paper in detail. Some of the techniques listed in the paper are still not feasible to

implement due to its limitations. But as the world has progressed, it has made some of its applications easy and convenient to implement. Today AR is used in medical training, business logistics, design and modeling, repair and maintenance, retail, tourism industry, classroom education, entertainment properties, etc. As there exist many pros of the AR on the other hand it also has some cons.

Privacy and Security Concern: increased reality needs tons of assortment, generation, and analysis of huge sets of information. Thus, it will increase the danger of problems regarding privacy and security.

Issues concerning Intrusiveness: AR systems record the atmosphere in real-time. This may raise legal problems, a similar approach capturing pictures of random people, and their personal properties.

Can Promote Risky Behavior: AR will hide cues within the globe. A number of these cues naturally facilitate people to avoid dangers. however, the technology will create a private less alert concerning his or her surroundings.

It may be Costly: Businesses will use it to enhance their services or processes. However, implementing the technology needs technical experience, similarly to money prices. As of the instant, it's solely obtainable to a giant or financially capable organization..

IV. PROPOSED SYSTEM

After looking at the security concerns of the AR we have proposed an idea to overcome it. The idea emphasizes the protection of user data and the prevention of malpractices of data. This can be done by implementing a proper organization for AR data storage. Making a regular analysis of the data and its data points. The collection of data should be done with users' permission and will. Security mechanisms should be installed across every platform that uses AR mechanisms. If possible these platforms must be connected in Blockchain so that data cannot be changed or stolen.

Proper authentication should be done for users interacting with AR devices. This will ensure the security of data and no intrusion.

V. CONCLUSION

This paper describes various applications of Augmented Reality along with its limitations. AR has a large scope in future but due to current limitations it is not feasible to use it. AR can ease various tasks as described in the paper like assistance in education(geometry), accuracy of interaction between application objects and users, personalized AR assistance for shopping, enhancing the gaming experience and tourism. Future research leads towards improving the performance.

References

[1]Shubham Gargrish, Archana Mantri, Deepti Prit Kaur Chitkara ,'Augmented Reality- Based Learning Environment to Enhance Teaching-Learning Experience in Geometry'- from University Institute of Engineering and Technology, Chitkara University, Punjab,

[2] Eg Su Goh, Mohd Shahrizal Sunar (Member, IEEE), And Ajune Wanis Ismail, '3D Object Manipulation Techniques in Handheld Mobile Augmented Reality Interface' from Faculty of Engineering, School of Computing, University Technology Malaysia, Johor Bahru 81310.(2019)

[3]Ashok kumar.P a, Murugavel.R , 'Prospects of Augmented Reality in Physical Stores's using Shopping Assistance App' from VIT-BS ,Vellore - 632014, India b Associate professor VIT Technology Management ,Vellore - 632014 ,India.(2020)

[4] Akshay Karkera, Sushil, Dhadse, Vinayak Gawade, Kavita Jain, 'Pokemon Fight Augmented Reality Game' from Xavier Institute Of Engineering Mumbai University Mumbai, India.(2018)

[5]. Xiaodong Wei, Dongdong Weng, Yue Liu, Yongtian Wang, 'A tour guiding system of historical relics based on augmented reality' from Beijing Engineering Research Center of Mixed Reality and Advanced Display, Beijing Institute of Technology, 100081. (2016)

[6] Steve Chi-Yin Yuen,Gallayanee Yaoyuneyong, Erik Johnson, 'Augmented Reality: An Overview and Five Directions for AR in Education' from The University of Southern Mississippi(2011).

Identity Resolution In Social Network Using Recommender System

Mayuresh Pandey, Ravita Mishra
Information Technology, K.J Somaiya Institute of Engg. and Information Technology,KJSIEIT, Mumbai,
Computer Engineering, Thakur college of Engg. & Technology. Mumbai, India
mayurp.pandey@gmail.com ,m.ravita@gmail.com

Abstract— **Online Social Networks are online platforms where users can post and share their social and professional life. Following other individuals with similar personal and professional career interests, day to todat activities, backgrounds or real-life connections also a main job of online social networks. . To enjoy the different services that various online social networking sites offer, a user creates an identity on each of them. The identity he creates constitutes three categories namely profile, content and connection network. There is an absence of a global identifier that can map a user's presence uniquely in the online domain. Due to this, his online identities remain extrange, isolated and difficult to search. Identity search methods have been proposed by literature using simply profile at-tributes of the user but the content and network attributes have yet to be explored. This system proposes an identity search algorithm that uses content attributes of the user in addition to profile attributes thus improving the traditional identity search algorithm. The proposed identity search algorithms are used to find a user's identity on Facebook, given her identity on LinkedIn. Including the two identity search algorithms, each exploiting distinct attributes of identity, improves the filtering accuracy among a candidate sets of similar user-profiles and helps to identify the required user more precisely. With the help of collaborative filtering algorithm similarity between user's identity searches will be improved. The proposed system uses the memory-based collaborative filtering algorithm. The paper is organized into five main parts. The first section includes the introduction of online social networks and recommender system, the second section includes the detailed state of the art and third part contain the proposed system and algorithm. Fourth section includes Methodology and Algorithm, Fifth section includes experimental results and observation.**

Keywords— Online Social Networks (OSN), Facebook,LinkedIn, Collaborative filtering, Jaccard Similarity.

I. INTRODUCTION

1.1Online Social Networks: The world now a day is undergoing advancements in terms of paper-based records to the electronic ones. Due to this, at the time of data entry, there is no proper verification or validation done which results in false or duplicate records in the electronic systems. Entity resolution refers to the means of linking and deduplication of the records in the areas of statistics and database management. Identity resolution is a unique type of entity resolution that specializes in identity management. Entity resolution is also known as record linkage and deduplication in the areas of statistics and database management. Record linkage [1], constructed in the statistics community and is used to identify those records in one or multiple datasets that refer to the same real-world entity. The very same task is often called and studied as record deduplication in the database and artificial intelligence communities. Entity resolution techniques can be expanded to different contexts, first, identity resolution, especially in the intelligence and law enforcement communities, often suffers greatly from the missing data problem. Missing values, if present in many fields of a record, can present a big challenge for identity resolution techniques. Second, identity resolution needs to handle not only duplicates caused by entry errors or data ambiguities but also intentional identity fraud and deception, which tend to be hidden and concealed. Third, identity resolution techniques may need to be adjustable to different evaluation criteria. For instance, false positives may be less tolerable than false negatives for identity authentication that grants access to a critical facility. In contrast, a high false positive rate may not be a big concern when a detective search for records related to a crime suspect with limited information. Therefore, accurate identity resolution requires a careful design that considers the special characteristics of identity records [2].

1.2 Recommender System Introduction: Recommendation system is a technique, which provides users with information, which he/she may be interested in or accessed in past. Traditional recommender techniques such as content and collaborative filtering used in various applications such as education, social media, marketing, entertainment, e-governance and many more. Content-based and collaborative filtering has many advantages and disadvantage and they are useful in specific application. Sparsity and cold start problem are major challenges in content and collaborative filtering. Challenges of content and collaborative filtering can be solved by using hybrid filtering. Hybrid filtering combines the features of two recommender system like content and collaborative; content-based filtering improves the classification accuracy and collaborative model easily gives the best-predicted result of a latent factor model [17].

1.3 Recommender System Uses: Malicious users create multiple accounts on different networking sites to enhance reachability to targets. In this, our solution can help searching for malicious user's multiple online

identities. Malicious users exploit online social media for activities such as Phishing, Spam, Identity theft etc.

II. LITERATURE SURVEY

A detailed survey on recommendation system is provided by Balraj et.al. [3] presented an immense idea about different taxonomy of recommender system, gaps exist in current system, application domain such as television, research agency, restaurant booking, job search and many newer research areas. Anitha et. al. [6] presents different types of Recommendation system for online resources, such as blogs, forums, social networking websites, bookmarking websites, and video and chat portals. Makdalini et. al [4] presents the different challenges and solution of online large social networks. Mamadou et. al. [8] presents another type of recommender system (Job Recommender System) and offer job based on Facebook extracted field. With the help of support vector machine and content based filtering analyses different field. Paridhi et. al , 2013 [1], they proposed an identity resolution system and spoke briefly about Identity, Identity Resolution, Identity search methods and Identity Matching methods. Paper [5, 7], 2015 presents the idea about collaborative filtering and their application. Cheng-Ta et. al [2], 2014 presents network search method as an advancement to find the similarity between the two users. The method used to find the similarity was by using Jaro Distance Formula on the Profile attributes of the user. Jiexun LiAlan G. Wan, [12], 2015 proposed a methodology which finds deviation in Identity attribute matching, self-mention matching and Network search of the user. They used a dataset which was a Real-World Dataset. It openly speaks of the crawler used to extract the data from Facebook [15]. Facebook data can be accessed with the help of services provided by the Facebook's developer team. The crawler used accesses the users friend list to acquire the network of the user which for further used for predicting the relationship of the user on the online social media. Dr. Shalini et. al [16] focuses on the different types of NoSQL databases. Ravita [11, 17] proposed the new frameworks of identity resolution and application of collaborative filtering.

Problem definition: There is a need for a system that can link a user's profile on multiple online social networks by using the means of a global identifier. An identifier is required to accurately identify a user among the billions of isolated profiles within an online social networking site. Such a system can collate a user's multiple online social network profiles as a single identity. The society faces problems such as malicious users who post content which are not good for the society. The need to identify such users also arises. HR of a company can sometime feel the need to check the candidate's background who is giving the interview to ensure if the candidate is suitable for that company or not.

III. PROPOSED SYSTEM DESIGN

People rely on social media for different purposes. Not all the online social media provide the same service. For instance, users may exploit LinkedIn for professional connections while Facebook for personal connections such as chatting with friends and family and may use Twitter for sharing public information and his/her views. The proposed system will find the users having their accounts across two online social networks. The system will be provided with a user's LinkedIn profile from which it will acquire attributes of that user and then will proceed to find the user's profile on Facebook. The system acts as a hub for a person's online social network accounts i.e. we merge the information of a single user having accounts at multiple online social networking sites.

The system will take the input of user's profile attributes such as name, gender, location, education, occupation etc. On basis of the attributes given as input to the system, the system will be performing attribute matching with Facebook attributes and will try to find a match for the user. After the profile attribute matching is done, we get Shortlisted candidates. The system will then extract the content of the shortlisted candidates that the user creates or is being shared which includes likes, follows, comments, posts etc. and then start with the content matching. The content matching is done to increase the accuracy of finding the appropriate user's account across Facebook. Once the profile attribute matching and content matching is performed, the result gets displayed on the web page.

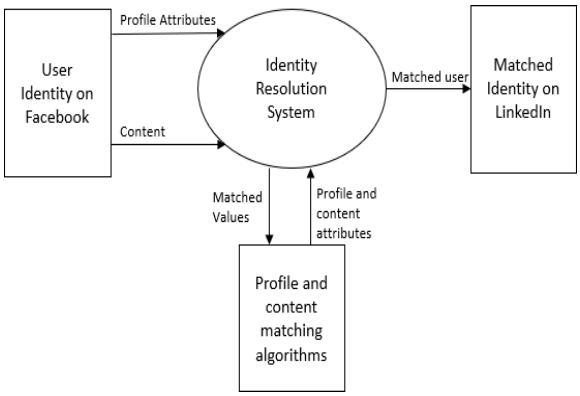


Fig: 1. Identity matching System [12]

When a Facebook or LinkedIn user receives an email from our recommender system saying that some of his friends would be perfect for a given job, he can approve or reject the matches. Its feedback, when provided, is included into databases. Through LinkedIn we can find the suitable candidates whose score is good and has friend recommendation through Facebook. Through Facebook we can capture some important fields which will be similar to LinkedIn fields. The table 3.1 depicts the important field for Facebook and LinkedIn for data matching are [16].

Table 3.1 Attributes of Facebook and LinkedIn

Job	Facebook	LinkedIn
Title	Work history	Headline
Description	Education history	Position
Responsibility	Quotes	Title
	About me	
	Interest	

The identity resolution system makes use of some algorithms to search individuals and find the accuracy and determine the associated records. For example, there can be more than one individual with the name John Snow. Through this system other attributes of the individual such as contact number, email id, working at etc. can help to differentiate among the individuals and detect the matches. These attributes are not accurate for many reasons such as entry errors or intentional deception. Identity resolution techniques need to be adjustable to different evaluation criteria. Therefore, a careful de- sign is required to maintain the accuracy of the identity resolution while considering the special characteristics of the data present in the records. The main block of system is discussed below:

System Development Life Cycle: The Developer's page of both Facebook [2] and LinkedIn [9] contain a wide range of API's from where the data is gathered. For this, we need to create a developer's page account to get an access token that never expires. The access token is used to get authentication to extract the data from Facebook and LinkedIn databases. Through this API extraction, we are able to get the unique ID of the users who have their accounts in the respective OSN's. The extraction of user's profile information from these unique ID's becomes tedious and hard to refine. Hence, we have decided to extract the data by manual intervention and at the initial stage limit our dataset within our college.

Data Cleaning The data obtained will be cleaned by removal of any missing or irrelevant data. Missing data is the data which is not public or the data which is not available for a given attribute. This happens when the user keeps his account private or if the user does not update his account. Irrelevant data is the data which mismatches with the given attribute. For example, the name of the user can be his name along with his contact number. The irrelevancy of the data occurs when the user does not register properly on the given online social network.

Data Storage The data collected will be stored in a flat file. The flat file we used is a comma delimited file that is .csv file. The data which is stored can be used for linking the users. The data stored in the flat file is in unstructured format.

IV. METHODOLOGY

To optimize the performance of identity resolution process we propose a system that utilizes a pairwise comparison string matching algorithm. Here P1 and P2 are two profile A1, A2 are the profile attributes. Collective clustering does require defining a similarity function on clusters of references. Each cluster formed is considered the same real-world individual. Unlike transitive closure that uses single-linkage clustering, collective clustering uses an average linkage approach instead. It defines the similarity between two clusters ci and cj as the average similarity between each reference in ci and each reference in cj:

$$sim(ci, cj) = 1|ci.R| \times |cj.R| \sum r \quad \text{----------Eq}^{n}(1)$$

$$\in cj.R, r' \in cj.Rsim(cj.r, cj.r')$$

where |c, R| represents the number of references in cluster c.

ALGORITHM 1

1. Initialize each reference as a cluster
2. Compute the similarity between each cluster pair
3. Find the cluster pair with the maximal $sim^*(c_i,c_j)$
4. If $sim^*(c_i,c_j)$ is greater than threshold
   a. Merge c_i and c_j
   b. Go to step 2

proposed an identity resolution system using clustering methods where the networks of the individual a social group were being clustered together by the collective clustering algorithm. To optimize the performance of identity resolution process we propose a system that utilizes a pairwise comparison string matching algorithm. Here P1 and P2 are two profile A1, A2 are the profile attributes.

ALGORITHM 2

Input: Profile Attributes of one OSN.
Output: Matched Identity on the second online social network

1. Start
2. We input the name of profile P1.
3. Find the Similarity (P1. A1, P2. A1).
4. IF similarity is greater than a threshold value then predicts it as a match 1 or 0.
5. Repeat process for all the attributes.
6. Similar profiles are tested if values (total 1's) > threshold then it shows that two profile belongs to a single person.
7. Stop

Modified Algorithm: the above algorithm give improved result and accuracy also increases as compare to hierarchical clustering algorithm. New algorithm uses advanced matching technique that will match profile and attributes correctly. NLP parser also gives better result for similarity calculation between two OSNs (two user profile their specified attributes).

V. EXPERIMENTAL SETUP AND ALGORITHM IMPLEMENTATION

The proposed system is using data of 1,000 users from data collected from the respective OSNs. Hardware and software Requirements of the proposed system is, 4GB RAM. Intel 1.66 GHz Processor Pentium 4, Quad-core CPU. For designing purpose, we used NoSQL (MongoDB), PHP, HTML and CSS and JavaScript.

5.1 User Interface Design

User Interface: A search bar that provides the user to enter the name or few letters of the name in order to find the required user on the online social network. The user interface is the first page that the user will see when the user opens the system. The search bar accepts only characters as an input. In this system two different types of matching (similarity) evaluated first profile matching and content matching.

Fig. 2. Interface of Social Linker

Profile Matching: It shows the profile attributes of the user such as Email, living in, Birthday and Studied at across OSN1 that matches to the attributes such as Email, and Working at, living in and studied at across OSN2. It displays the profile attribute matched percentage. Profile matching uses simple algorithm to compare basic attributes of two OSNs.

Content of User: The main part is this system is to match the contents of two different OSNs of same person. With the help of Jaccard similarity and Hadamard similarity, it calculates the matching contents of two different OSNs. Fig. 3 shows the content of the user on different online social networks.

Fig. 3 Profile match of user

The content contains the posts of the users on their respective accounts on different social media sites. The posts of their respective social media is shown in this page. The Fig. 4. shows the content of the user on Facebook.

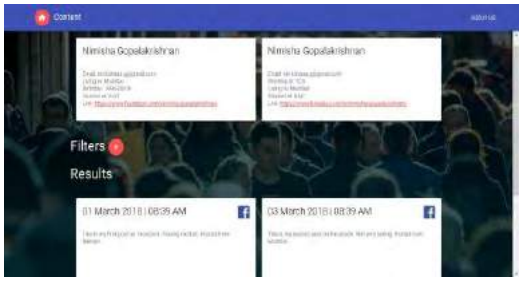


Fig. 4 Content of Facebook User

The Fig. 5 shows the content of the same user on LinkedIn with same credential and (name, email-id and their date of birth, company profile). With the help of similarity matching it shows the same users information on other social media.

Fig. 5. Content of LinkedIn User

Content Matching: It displays the percentage accuracy of the content matched. The percentage displayed is done using Jaccard distance formula. It matches the content of OSN1 with the content of OSN2 of the same user accounts on different online social networks. Fig. 6 shows the content of user on Facebook. The below figure exactly match the contents of two different OSNs of same user and result shows the percentage of matching of their contents.

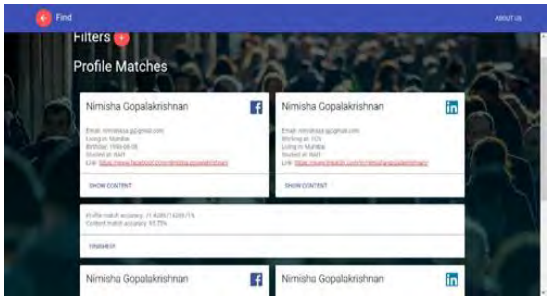


Fig. 6. Content Matching of User

5.2 Database of the different online social networks

The figures below show the gathered data from the different online social networks. The database used for this purpose is MongoDB for simplicity and size. The JSON file of the first OSN are populated in our

main database for evaluation purpose. The Fig 7. Shows the database of the first OSN which is Facebook.

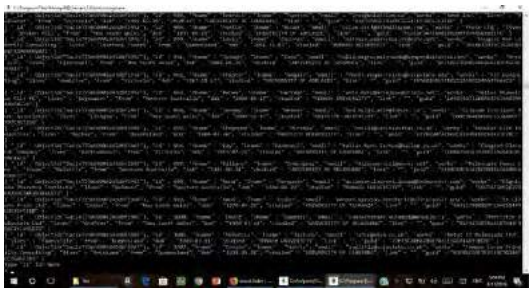


Fig. 7. Database Profiles of OSN1 (Facebook)

For LinkedIn database same steps applied and JSON file of LinkedIn database are populated in main database. The contents of user profile on LinkedIn is shown below:

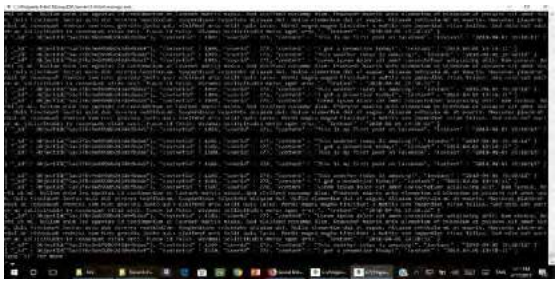


Fig. 8. Database Content of OSN1 (Facebook)

The database profile and contents of LinkedIn users are shown in below. Fig. 8. depicts the Database profile of LinkedIn User and Fig. 9. depicts the database contents of LinkedIn users.

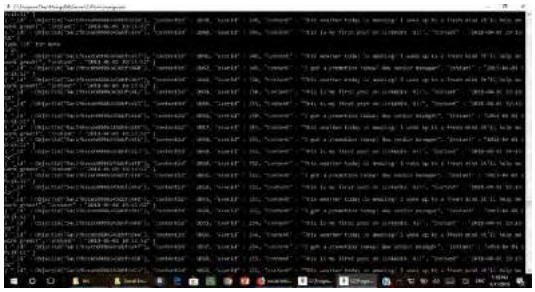


Fig. 9. Database Profiles of OSN2 (LinkedIn)

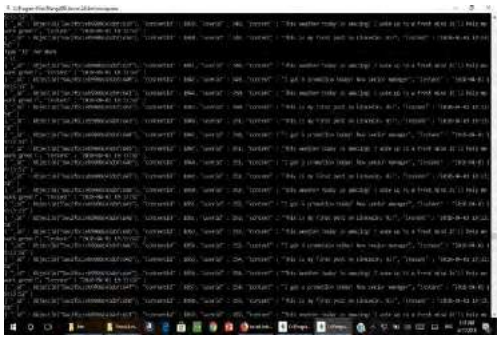


Fig. 10. Database Content of OSN2 (LinkedIn)

The contents based recommender system gives the best contents matching and percentage of similarity is also more than manual system. The system needs some improvements. 1) we are tsetse on small static dataset of Facebook and LinkedIn due to privacy issues, so there is improvement to use API and test on for dynamic dataset. 2) System uses Jaccard Similarity measure, it gives satisfactory results, but improve similarity measure gives more accuracy.3) accuracy of system is 89 % so it should be increased by using advanced machine learning algorithm.

5.3 Test Cases:

Table 5.3.1: Test on search button

Test on search button				
Purpose	Input	Expected Output	Actual Output	Remarks
To test ne in search bar	Nimisha	Nimisha	Nimisa	Pass
Test on search button				
Purpose	Input	Expected Output	Actual Output	Remarks
Click search button	Diya	Diya	Diya	Diya
Alphanumeric test				
Purpose	Input	Expected Output	Actual Output	Remarks
To check if it accepts alphanumeric value	Ajju1997	Invalid	Invalid	Pass
Percentage Test				
Purpose	Input	Expected Output	Actual Output	Remarks
To check if it displays the accurate % of profile matching	Diya	Near 65%	66.21 %	Pass
Duplicate Value Test				
Purpose	Input	Expected Output	Actual Output	Remarks
To check if it displays the accurate % of profile	Diya	Near 65%	66.21 %	Pass

matching				
Display user's attribute				
Purpose	Input	Expected Output	Actual Output	Remarks
To check if it displays the accurate % of profile matching	Diya	Near 65%	66.21 %	Pass
Profile matching test				
Purpose	Input	Expected Output	Actual Output	Remarks
To check if it displays the accurate % of profile matching	Diya	Near 65%	66.21 %	Pass
Content of user check				
Purpose	Input	Expected Output	Actual Output	Remarks
To check if it displays the accurate % of profile matching	Diya	Near 65%	66.21 %	Pass

VI. Conclusion

Proposed system addresses the problem of identity resolution in online social network. Earlier research paper presented a solution by simply matching profile attributes of a user in two online social networks and they perform clustering algorithm for individual. We have proposed an identity resolution technique that perform pairwise matching on identity attributes and it match the content on both OSNs and self-mention search which access only the public information to find the candidate identities. Here we compare the two identity search algorithm, each exploiting distinct dimensional attributes of an identity, improves the filtering accuracy among candidate set of similar user profiles and helps to identify the required user more precisely. With the help of this system, we can accurately identify the correct identities of users from the two online social networks. Even through this work has focused on LinkedIn and Facebook, we believe that extension of identity search methods proposed in this work can be applied to similar social networks such as Twitter, Instagram and Google+ with minor changes. The future work that we are trying to implement is that increase the accuracy of system and applying the NLP parser to match the attributes. Machine algorithm also improve the attribute match. The current system is implemented with the help of hierarchical clustering algorithm but we can improve this system with the modified clustering algorithm.

References

[1] Paridhi jain et. al, " '@I seek 'fb.me': Identifying Users across Multiple Online Social Networks", Pages 1259–1268, doi.org/10.1145/2487788.2488160.

[2] C. T. Chung, C. J. Lin, C. H. Lin and P. J. Cheng, "Person Identification between Different Online Social Networks," 2014 IEEE/WIC/ACM,Warsaw, 2014, pp. 94-101, doi: 10.1109/WI-IAT.2014.21.

[3] Balraj Kumar, Neeraj Sharma, "Approaches, Issues and Challenges in Recommender Systems: A Systematic Review", Indian Journal of Science andTechnology9(47) · Dec 2015 DOI: 10.17485/IJST/2015/v8i1/94892.

[4] Magdalini Eirinaki, Jerry Gao, IraklisVarlamis, Konstantinos Tserpes, "Recommender Systems for Large-Scale Social Networks: A review of challenges and solutions", http://dx.doi.org/10.1016/j.future.2017.09.015 0167-739X/© 2017, Elsevier B.V.

[5] Yangi, Bing Wui, Kan Zhengi, Xianbian Wang2, Lei Lei3, "A Survey of Collaborative Filtering-Based Recommender Systems for Mobile Internet Applications", May 26, 2016, version July 7, 2016. DOI: 10.1109/ACCESS.2016.2573314.

[6] Anitha Anandhan, Liyana Shuib, Maizatul Akmar Ismail, and Ghulam Mujtaba, "Social Media Recommender Systems: Review and Open Research" February 27, 2018, IEEE Vol. 6, 2169- 3536, DOI: 10.1109/ACCESS.2018.2810062.

[7] Adomavicius, G., Tuzhilin, A. (2005), "Towards the next generation of recommender systems: a survey of the state-of-the art and possible extensions". IEEE Trans. Knowl. Data En. 17, 734–749.

[8] Mamadou Diaby, Emmanuel Viennet and Tristan Launay, "Toward the Next Generation of Recruitment Tools: An Online Social Network-based Job Recommender System" 2013 IEEE/ACM.

[9] Mahdi Jalili, Sajad Ahmadian, Maliheh Izadi, Parham Moradi, Mostafa Saleh, "Evaluating Collaborative Filtering Recommender Algorithms: A Survey" DOI: 10.1109/ACCESS.2018.2883742, Vol-6, IEEE Access, Nov 2018.

[10] Charu C aggrawal,"Recommender system Textbook", ISBN 978-3-319-29657- 9 ISBN 978-3-319-29659-3, DOI 10.1007/978-3-319-29659-3, Springer, 2016.

[11] Ravita Mishra. "Entity resolution in online social networks (@Facebook and LinkedIn)", Proceedings of IEMIS 2018, Volume 2 January 201 DOI: 10.1007/978-981- 13-1498-8_20.

[12] Li, J., Wang, A.G. A framework of identity resolution: evaluating identity attributes and matching algorithms. Secur Inform 4, 6 (2015). https://doi.org/10.1186/s13388-015-0021-0

[13] Paula R. C. Silva, Wladmir C. Brandaˆo, "ARPPA: Mining Professional Profiles from LinkedIn Using Association Rules", 2015, in press.

[14] Dr. Mamta Madan and Meenu Chopra, "Using Mining Predict Relationships on the Social Media Network: Facebook (FB)", 2015 in International Journal.

[15] Simplified Web Scraping, https://nocodewebscraping.com/how-to-extract-data-from- facebook-page-competitor-analysis.

[16] Lokesh Kumar, Dr. Shalini Rajawat, Krati Joshi," A Comparative analysis of MongoDb vs MySQL", International Journal of Modern Trends in Engineering and Research..

[17] Mishra R., Rathi S. (2020) Efficient and Scalable Job Recommender System Using Collaborative Filtering, ICDSMLA 2019. Springer, Singapore. https://doi.org/10.1007/978-981-15-1420-3_91

Role Of Fog Computing In Iot Based Applications

Manasi Kukarni, Mayur Panchariya, Damini Mahale, Ritesh Kulkarni, Bhushan Nandwalkar
Computer Engineering,Shri Vile Parle Kelvani Mandal's Institute of Technology,Dhule,India
manu221143@gmail.com, mayurpanchariya95@gmail.com, daminimahale234@gmail.com, riteshkulkarni2311@gmail.com, nandwalkar.bhushan@gmail.com

Abstract— **Services of Internet of things (IoT) have been accepted and accredited universally for the past few of years and have had increasing interest from researchers. Requirement of Internet of Things, are mobility support and geo-distribution in addition to location awareness and low latency. We express that a new platform is needed to meet these requirements; we call it the Fog Computing platform. The Cloud Computing paradigm to the sting of the network was extended by Fog Computing, thus enabling a replacement breed of applications and services. The potent idea of fog computing is currently attracting many researchers because it brings cloud services closest to the end-user. The aim of this paper is to spotlight the role of fog computing in IOT based applications.**

Keywords— Fog computing, Cloud computing, Edge computing, IoT applications, Fog with IoT

I. INTRODUCTION :

This template, The Strong concept of fog computing is now attracting many researchers as because it brings many cloud services closer to the end-user.

Over the previous couple of years, Internet of Things (IoT) has gained significant attention, because it provides various IoT services in most fields of life and Technology. IoT is an interconnected network of huge numbers of IoT devices, each having the capacity or power of sensing and communication, through which they report their sensed data to the most server. This permits the center, supported received data, to require decisions intelligently like small wireless devices utilized in S-band sensing technique, IoT uses small sensor devices. The rise in usage of IoT devices has led to requirement of resource and computing paradigms which can work efficiently together with IoT environment. The main prototypes are Fog computing, Cloud computing, and Edge computing. This paper will mainly special in Fog computing with IoT services.

Before that specialize in Fog computing allow us to know the technologies used before it, that are Cloud Computing and Edge Computing.

A. IOT:

The Internet of Things, or Iota, refers to the billions of physical devices round the world that are now connected to the web, all collecting and sharing data. because of the arrival of super-cheap computer chips and therefore the ubiquity of wireless networks, it's possible to show anything, from something as small as a pill to something as big as an aeroplane into a neighborhood of the IoT. Connecting up of these different objects and adding sensors to them adds A level of digital intelligence to devices that might be otherwise dumb, enabling them to speak real-time data without involving a person's being. the web of Things is making the material of the planet around us more smarter and more responsive, merging the digital and physical universes..

B. Cloud Computing:

The use of hardware and software to deliver a service over a network or we will say that internet is understood as Cloud Computing. another word we will say that Cloud computing is that the on-demand availability of computer Resources of systems, especially data storage and computing power, without direct active management by the user. The term is generally wont to describe data centers available to many users over the web .

Cloud didn't managed various requirements of IoT efficiently such as: privacy, scalability, enormous bandwidth requirements, efficiency in network computations, energy consumption, and delay-sensitive communication

C. Edge Computing

The computational processing of sensor data far away from the centralized nodes and shut to the logical fringe of the network, toward individual sources of knowledge is understood as Edge computing. The technologies involved network nodes storing static cached media information at locations closest to end-users. Only partial sets of data processed and analyzed by edge computing. And it Only delete the remainder of the records. thanks to its proximity to the users, latency in edge computing is usually less than in cloud computing. Edge Computing cannot Support the Multiple IOT Devices.

D. Fog Computing

Fog computing could also be a replacement technology paradigm to reduce the complexity, scale and size of the data actually rising to the cloud. Pre-processing of data beginning of the sensors and IOT devices is important and it's an efficient way to reduce the load of the large data on the cloud. Fog computing connects the gap between the cloud and end devices (e.g., IoT nodes) by enabling computing, storage, net-working, and data management on nodes network within the close vicinity of IoT devices

II. FOG COMPUTING IN IOT

Internet of Things (IoT) needs to be operate on a fast network topologies that provides end-to-end connection and real-time responses. For instances the frequent dis-connections and reconnections by the devices, or notifications of a disaster or an imminent collapse of the system. In many cases, decisions must be taken in a short time and it is necessary to be able to rely on a reliable connection between the customer and the corresponding servant who performs complex tasks. In many situations, especially dictated by the overload of communications in multi-hop WAN networks, these qualities aren't guaranteed by the Cloud. Because cloud computing isn't reliable for several Internet-of-Things

applications, fog computing is usually used. As Its distributed approach addresses the needs of IoT and industrial IoT, the immense amount of data smart sensors and IoT devices generate, which would be costly and time-consuming to send to the cloud for processing and analyzing. Fog computing reduce the bandwidth needed and reduce the back-and-forth communication between sensors and therefore the cloud, which may negatively affect IoT performance.

Although latency could also be annoying when sensors are a part of a gaming application, delays in data transmission in many real-world IoT scenarios are often life-threatening -- for an instance, in vehicle-to-vehicle communications systems, smart grid deployments or telemedicine and patient care environments, where milli-seconds matter. Fog computing and IoTs uses the cases also include smart rail, manufacturing and utilities. Hardware manufacturers, such as Cisco, Dell and Intel, are working with IoT analytics and machine-learning vendors to create IoT gateways and routers that support fogging. In the November 2015 the Open Fog Consortium was founded by members from Cisco, Dell, Intel, Microsoft, ARM and Princeton University; its mission is to develop an open reference architecture and convey the business value of fog computing.

III. LITEATURE SURVEY

Our Literature survey is based on previous technologies i.e. Cloud computing and Edge computing with reference to some technical parameters.

A. Geographically Distributed:

As per the author the cloud computing system is not geographically distributed whereas as per survey of authors the edge computing and fog computing are distributed systems. We can say that the edge computing and fog computing are interface between server and the IoT application. [1][2]

We can say that edge computing is partially geographically distributed system because it repeatedly needs a reference of cloud server to response to smart devices whereas fog computing is completely distributed system because to distribute data to move it closer to the end user which eliminate or decrease the latency.[3]

B. Real time application:

The author cloud computing does not support real time application due to some reasons like it is not geographically distributed system.[1]

Also author says that the edge computing supports small real time applications as follows the repetitive query system.[2]

Whereas the fog computing fully supports the real time applications as it is geographically distributed system, its latency is low .[3]

C. Large scale Application support:

As per the author the cloud computing does not support large scale application because it's bandwidth cost is high as compared to edge computing and fog computing [2][3]

Edge computing is suitable/supports large scale application but drawback of this system it is a partially distributed system and it follows repeated query system due to which response time is increased.[2] Whereas author claims that fog computing is perfect or complete solution for real time as well as large scale application because it is a distributed system as it distributes the data in many small modules nearer to the end user due to which its bandwidth cost is low and response time is high. [3]

D. Server and Storage:

As per the author in case of server cloud computing is centralized server system whereas edge computing is partially centralized server system and fog computing has decentralized server because it is a completely distributed system i.e. it has multiple remote locations same as its application like smart city or smart watch has. In case of storage cloud computing has a big storage as compared to edge and fog because edge computing act as a interface between server and end user. So we can say that it is a small bridged storage whereas fog computing has a multi storage or multiple storage due to its distributed format . [1][2][3]

E. Security:

After studying the data security problems in Fog-IoT network, authors considered different security protocols that could be used in Fog-IoT network for data security. The key focus of this model is data security during data travelling from client to Fog nodes.

They selected Shibboleth security protocol after its critical analysis in Cloud-IoT environment. Shibboleth provides system with integrity, authentication and privacy.

In Shibboleth, access control mechanism compares attributes, issued by identity provider. Moreover, metadata is the trust basis between Shibboleth providers. Considering the substantial of this protocol, authors added new layers of Shibboleth protocol between client and Fog node.

The purpose is to secure data access, authentication, authorization and user's privacy from Service Providers (SP). Later, they used High-Level Petri Net (HLPN) for system modeling and Z3 constraints solver for automated SMT solution and formal verification. In addition, it also provided the correctness of system.

So we can say that if we compare all three systems then cloud and edge computing are less secured than fog computing because:

- As cloud is centralized and edge is partially centralized system so it can get easily affected by any malware attacks like DOS attack , MITM attack ,etc.

IV. ARCHITECTURE OF FOG COMPUTING:

Let us understand architecture of fog computing as a part of our basic fog computing tutorial. The Fog computing architecture consists of physical and logical elements within the type of hardware and software to implement IoT (Internet of Things) network. As shown in figure-2, it's composed of IoT devices, fog nodes, fog aggregation nodes with the assistance of fog data services, remote cloud storage and native data storage server/cloud. allow us to understand fog computing architecture components.

A. Fog computing architecture:

- IoT devices: These devices are connected to the IoT network through various wired and wireless technologies. These devices produce data regularly in huge amount. There are numerous wireless technologies utilized in IoT which include Zigbee, Zwave, RFID, 6LoWPAN, HART, NFC, Bluetooth, BLE, NFC, ISA-100.11A etc. IoT protocols used which include IPv4, IPv6, MQTT, CoAP, XMPP, AMQP etc.
- Fog Nodes: Any device with computing, storage and network connectivity is understood as fog node. Multiple fog nodes are spread across larger region to supply support to finish devices. Fog nodes are connected using different topologies. The fog nodes are installed at various locations as per different applications like on floor of a factory, on top of power pole, in conjunction with of railway track, in vehicles, on oil rig then on samples of fog nodes are switches, embedded servers, controllers, routers, cameras etc. Highly sensitive data are processed at these fog nodes.
- Fog Aggregate Nodes: Each and every fog nodes have their aggregate fog node. It analyzes data in seconds to minutes. IoT data storage at these nodes are often of duration in hours or days. Its geographical coverage is wider. Fog data services were implemented to implement such aggregate node points. they're wont to address average sensitive data.
- Remote Cloud: All the mixture fog nodes are connected with the cloud. Time insensitive data or less sensitive data are processed, analyzed and stored at the cloud.
- Local server and cloud: Although architecture of fog computing uses private server/cloud to store the confidential data of the firm. These local storage is additionally useful to supply data security and data privacy.

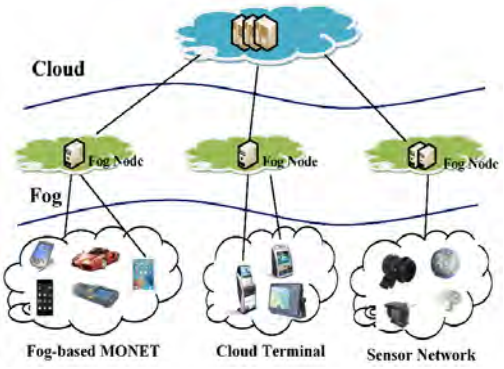


Fig 1. Architecture of Fog computing

B. Advantages :

- Low Latency: Fog computing has a low latency.
- Geo-distribution: The structure of fog computing is geographically distributed.
- Real-time interactions: Response time is low with the IoT applications.
- Location Awareness
- Support for mobility: It has a flexible mobility due to which its response time is low.

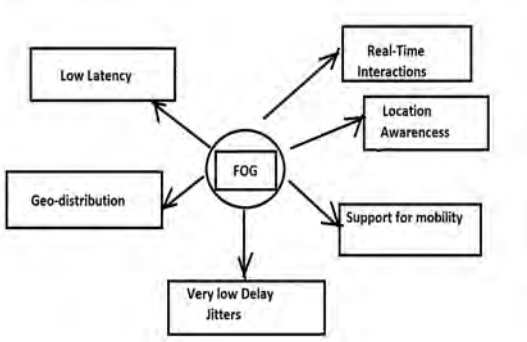


Fig 2.Advantages of Fog computing

V. APPLICATION BASED EXPLANATION:

SMART CITY

A. Smart city with Fog Computing:

Large cities face challenges from traffic jam , public safety, high energy use, sanitation and in providing municipal services. These challenges are often addressed within one IoT network by installing a network of fog nodes.

A lack of broadband bandwidth and connectivity may be a major issue in establishing smart cities. Where atest cities have one or more cellular networks providing adequate coverage, these networks often have capacity and peak bandwidth limits that hardly meet the wants of existing subscribers. This leaves little bandwidth for the advanced municipal services envisioned during a smart city. Deploying an architecture of fog computing allows for fog nodes to provide local processing and storage. This optimizes network usage.

Smart cities also working hard with safety and security, where time-critical performance requires advanced, real-time analytics. Municipal networks may carry sensitive traffic and citizen data, also as operate life-critical systems like emergency response. Fog computing addresses security, encoding and distributed analytics requirements.

Smart cities can see the subsequent benefits through fog computing:

- A minimal amount of knowledge sent to the cloud.
- The central goal of fog computing is to form big data smaller and more manageable. It's estimated that the quantity of knowledge captured by connected devices will exceed 79 zettabytes by 2025 consistent with IDC's 2019 forecast. Fog computing is capable of reducing this vast amount of data through the appliance of intelligent sensing and filtering, which enable the transmission of only useful information supported the knowledge available locally at a given fog device.

- Low data latency Fog nodes are ready to process and onboard data without sending it to remote cloud servers and delivering the results back. This makes it easier to save lots of time considerably when the info is traveling and to receive responses in real time. Immediate processing will only become more essential for smart city systems, especi
- ally when decisions or actions got to be made quickly: for instance , lives might be saved by having the ability to suddenly change traffic lights to green when emergency vehicles are moving through the town .
- Reduced bandwidthTransmitting and processing data requires a huge amount of bandwidth, which may be limited within the case of cloud computing. However, this is often not a problem when it involves fog computing seeing as all of the info is distributed between local devices and isn't sent wirelessly. this enables for a big decrease within the network bandwidth consumption.
- Enhanced data security Another critical driver behind smart cities turning their resources over to fog computing is known as data security. It keeps the more sensitive and confidential data out of reach from the vulnerable public networks, thereby preventing any cybercriminals from easily gaining access there to Fog computing has benefit that it allows for mal-ware and infected files to be found at an early stage in their cycles at the device level long before they even have the chance to infect the entire network.

.

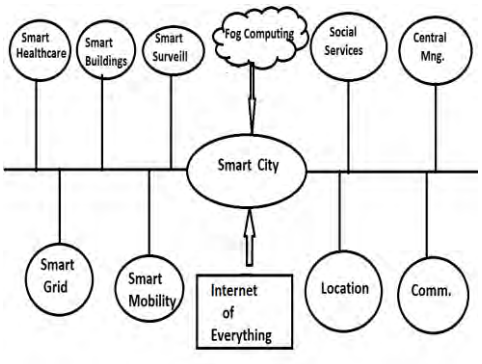


Fig 3. Services can be provided in Smart City using Fog computing

References:

[1] Rajaputhri Maharaja1 • Prashant Iyer1 • Zilong Ye1. A hybrid fog-cloud approach for securing the Internet of Things. Cluster Computing (2020) 23:451–459

[2] D. S. Park. Future computing with IoT and cloud computing The Journal of Super computing (2018) 74:6401–6407.

[3] Ping Zhang1·Mimoza Durresi2·Arjan Durresi1 Multi-access edge computing aided mobility for privacy protection in Internet of Things . Computing(2019)101:729–742 .

[4] Rida Zojaj Naeem1 · Saman Bashir1 · Muhammad Faisal Amjad1 · Haider Abbas1.Hammad Afzal1. Fog computing in internet of things: Practical applications and future directions. Peer-to-Peer Networking and Applications (2019) 12:1236–1262.

[5] SalvatoreVenticinque1 · Alba Amato1 . A methodology for deployment of IoT application in fog. Journal of Ambient Intelligence and Humanized Computing (2019) 10:1955–1976.

B. Smart city with cloud Computing:

The Problem is Vendor lock-in along side the shortage of control on the situation where applications run and data are stored. this is often a crucial barrier to cloud-based smart city solutions, especially when applications manage personal data and therefore the provider has legal obligation for securing data privacy. thanks to the massive overhead of smart city device data cloud computing suffers from time interval inefficiency.

C. Smart city with edge Computing:

As compared to the Fog computing, Edge computing is a smaller amount scalable. Also, edge computing supports only little interoperability, which can be make IoT devices incompatible with certain cloud services and operating systems. Also, IoT devices and cloud performed multiple tasks and operations which can't be extended by Edge Computing .

Smart city requires multiple iot devices but edge computing cannot support multiple iot devices.

VI. CONCLUSION

Fog Computing aims to scale back processing burden of cloud computing. Fog computing is bringing processing , networking, storage and analytics closer to devices and applications that are performing at the network's edge. that's the reason Fog Computing today's trending technology mostly for IoT Devices.

We have outlined the visions and key defined characteristics of Fog Computing, a platform to deliver an up-scale portfolio of latest services and applications at the sting of the network.The motivating examples prepared through-out the discussion range from conceptual visions to existing point solution prototypes. We have an envisioned that the Fog to be a unifying platform,rich sufficient to deliver this new breed of emerging services and enable the event of latest application.

E-Voting System Using Blockchain

Waseem Ansari, Siddesh Sharma, Yatish Chaudhari, Mayuri Kulkarni
Department of Computer Engineering, Shri Vile Parle kelvani Mandal,Institute of Technology,Mumbai Agra Highway, behind Gurudwara, Dhule, Maharashtra 424001
waseemansari9746@gmail.com , siddeshsharma1999@gmail.com , chaudhariyash29@gmail.com

Abstract- **In second Decade of 21st century,Latest technology is coming up with positive impacts on social life. And this all-time globally connected network enables us to access a variety of resources easily. One such revolution is Blockchain. With its special characteristics like immutability,decentralized architecture, many services are moving towards it. A potential application of blockchain can be found in electronic voting schemes. It has been a challenge since a long time to build an e-voting system which satisfies all legal requirements of lawmakers. Distributed ledger technologies can offer an infinite range of applications. This paper discusses various e-voting system frameworks conceptualized by different teams. Blockchain will give its benefits on e-voting systems including authentication, immutability of votes,system security, updation of votes in global ledger with not depending upon number of nodes in the network.**

Keywords—**Blockchain,E-Voting,ashing,Decentralized technology, Distributed System**

I. INTRODUCTION

Blockchain definitely became one of the most trending computational technologies.Blockchain is originally a continuous list of blocks, growing continuously, where each block is linked to next using cryptography. Every block consists of a hash, a timestamp and the data of transactions occurred. The generated hash is a cryptographic and according to the developer hash can be generated in many different ways. Blockchain consists of a system of recording all transactions efficiently and most importantly, in a verifiable manner and permanent way, known as distributed ledger. The main feature of blockchain is that its data can‟t be modified once it‟s been added to ledger.

I) E-VOTING

Electronic voting is a term that surrounds several different types of voting, embracing both electronic means of casting a vote and counting votes. an electronic voting (E-Voting) system is a voting system in which the election data is recorded, stored and processed primarily as digital information.Currently, various researches are conducting in - order to make a secure and reliable voting system while tackling issues of anonymity, fairness, reliability, and availability. Through the use of Blockchain, the focus is on making the Voting Process fair and without any third party intervention

While surveying these papers we see various concepts and ideologies,Various frameworks like Ethereum, Sawtooth are defined and concept of Smart Contracts using Solidity language. Various concepts like receipt free voting, E2E, Third-Party verifiable systems with development of various protocols relating to blockchain.

Ethereum

Ethereum is a decentralized, open source, and distributed vast computing platform that enables the creation of smart contracts and decentralized applications, also known as dapps. Ethereum is the largest cryptocurrency by market capitalization after Bitcoin.

Hyperledger Sawtooth

Hyperledger Sawtooth is an open source project of the Hyperledger umbrella, works as an enterprise level blockchain system which is used to create and operate distributed ledger applications and networks particularly for use by enterprises.

II) CHALLENGES OF VOTING

• Privacy: There shall be no third party intervention of any kind regarding Election. Only Voter is allowed to view his/her details and to whom they voted. The only disclosed information in election is total votes to candidates as well as in the entire election.

• Lack of Evidence: Although privacy with anonymity can ensure safeguards against electoral fraud. There is no way to ensure that votes are being casted under effect of bribes or any form of electoral fraud. This issue has roots from the beginning.

• Fraud-Resistance: Each eligible voter should be able to vote exactly once and no other voters should be able to vote. The system must verify the identity of each potential voter and check their status, but must not allow this information to become associated with their vote.

• Ease-of-Use: Elections must serve the entire public. It must be designed in such a way that it can be used with minimal training and some technical skills.

• Scalable: Election is a means to serve a large population. It must be flexible enough to work at large scale also.

• Speed: In this Computer Driven era, It must be ensured that results are declared within a few hours of election procedure ends.

II. LITERATURE REVIEW

PREVIOUS WORK HAS BEEN CONSIDERED WITH VARIOUS DIFFERENT SCHEMES AND THEIR IMPORTANT ASPECTS ARE DISCUSSED BELOW.

Baocheng Wang et al(2017) proposes a free-receipt electronic voting scheme on blockchain where the head node pushes contract in blockchain and uses smart contract to record, manage, calculate and inspect transactions. then further discusses the idea of one-time ring signature and encryption to protect the privacy and free-receipt of voters.

In their paper Robbie Simpson et al (2018) says about the Receipt free, voter verifiable system and how it gives assurance to voters about their vote. And also this system scheme prevents voters from being bribed or coerced. This paper discusses the idea of isolating voters in a secure polling booth to vote and then using a document as a witness to voters. This witness provides voters the confirmation that their vote has been counted correctly.

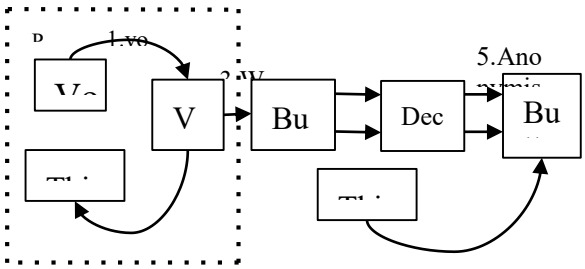


Fig.Third party reciept free system

In their paper on secure electronic voting system using blockchain technology ashish singh and kakali chatterjee(2018) suggests To Design more Secure and robust Electronic voting System.they done literature work on e voting system.and they have found many security issues which are very common in the e voting system to fulfill the security problem they have designed an e voting system which will mitigate all the possible threats and attacks. This system is based on the blockchain technology, which removes all the threats from the communication link.and the fact finding is It is a decentralized system, contain hashing and encryption concept for providing the security. system ensures that only registered and eligible voters are able to give their own votes. Once any voters completed her/his vote, the block will be created, which will be publicly verifiable and spread over the network. After completion of the blockchain no one will do any modification into the block. If an attacker wants to do any modification into the block, the hash value of THE block will change and the effect of the modification will reflect into the whole blockchain. The voter has the facility to register only once into the system. The voter ID is used for unique verification and checking the eligibility of the user. Thus, the model ensures that one voter gives only one vote, no one will allow two votes. The system security analysis shows that the system is more robust and secure against existing attacks.

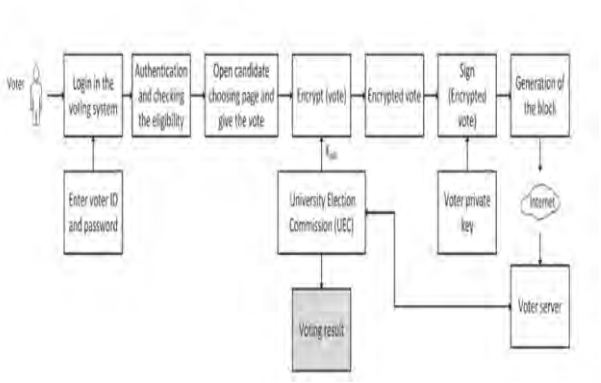


Fig.Framework of proposed electronic voting system

amish khandelwal [4] discusses various e-voting systems framework and conceptualized by different team.they have tackled the difficulties in the centralized voting system to make it more anonymous,reliable and secure while preventing any kind of frauds. through a decentralized system, focus is drifting towards making the voting process simple, secure and anonymity in the hand of the public.

In their paper vijayalakshmi v and vimal s [5] discusses the big number of problems in the existing voting system , which leads to political distraction of a nation. A transparent balloting system plays a vital role in a democracy to regulate free and fair elections. They propose a system that overcomes most of the disadvantages of the traditional system by providing voter privacy, security and also transparency where the user is allowed to verify their votes. In addition to this, the system allows the users to vote from their nearby location, thus achieving maximum voting percentage. the proposed system is cost-efficient when compared to the traditional electronic voting machines..The implementation of this system addresses most of the issues faced in the balloting scheme and it is used to avoid proxy roleplay and recasting and is also used to achieve maximum of the vote.

Kanika garg ET al(2019) reviewed on the paper and the techniques used to tackle voting challenges.to understand various methodology in voting system Comprehensive view on e-voting in thematic basis like e-voting with iot and fingerprint, e-voting with blockchain and Aadhar verification, etc.they have also reviewed existing voting method based on various factors (Authentication,Platform,Anonymity,Voter Verification,Decentralised, etc).

III. ANALYSIS

Serial Number	Year Of Publication And Author	Title	Main Focus/Concept
1	IIKI 2017 Baocheng Wanga,, Jiawei Suna, Yunhua	LARGE SCALE ELECTION BASED ON BLOCKCHAIN	INVESTIGATE AN ELECTRONIC VOTING SCHEME BASED ON BLOCKCHAIN
2	IEEE 2018 ASHISH SINGH,KAKALI CHATTERJEE	SECURE ELECTRONIC VOTING SYSTEM USING BLOCKCHAIN TECHNOLOGY	IMPLEMENTING MORE SECURED E-VOTING PROTOCOLS COMPARED TO OTHER SOLUTIONS
3	2018 ROBBIE SIMPSON, TIM STORER	THIRD-PARTY VERIFIABLE VOTING SYSTEMS	THIRD-PARTY VERIFIABLE VOTING SYSTEMS.
4	IEEE 2019 Amish Khandelwal	Blockchain implementation on E-voting system	AUTHOR BRIEFLY COMPARED VARIOUS PAPERS AND CONCLUDED HOW POSSIBLY AND EFFICIENTLY WE CAN IMPLEMENT BLOCKCHAIN TECHNOLOGY ON E VOTING
5	IEEE 2019 Vijayalakshmi v and Vimal s	A NOVEL P2P BASED SYSTEM WITH BLOCKCHAIN FOR SECURED VOTING SCHEME	AUTHOR PROPOSED COST EFFICIENT AS WELL AS RELIABLE SYSTEM WITH ENHANCED SECURE PROTOCOLS
6	IEEE 2019 Kanika Garg , Pavi Saraswat , Sachin Bisht	A Comparative analysis on E-voting system using blockchain	DETAILED COMPARISON ON ASPECTS LIKE PROS AND CONS OF TRADITIONAL VOTING AND ERROR FREE E-VOTING IS PROPOSED BY AUTHORS

The above analysis of various papers shows different regions of focus for achieving the same result. Different research uses different concepts and ideologies to move forward and which have been written in a few words as the main focus above.

IV. CONCLUSION

In this paper, an empirical review has been performed to understand various methodology in voting system. All the related papers are taken from thesis and literature and have been studied.

The existing voting system is having a big number of problems, and leads to political distraction of a nation. Transparent balloting system plays a vital role in a democracy to regulate free and fair elections. The proposed system overcomes most of the disadvantages of the traditional system by providing voter privacy, security and also transparency where the user is allowed to verify their votes.

Voter-verifiable systems like the Receipt free verification system have complex procedures for voters and demand a certain level of understanding at various points. Changing the whole system of voting introduces various security and reliability issues and concerns. On the other hand, changing the whole system should be the first thing according to the type of problems available in current implementation. We see how the revolution of blockchain can change various security systems and the amount of flexibility it brings with itself. Along with transparency to the system, it removes all the unnecessary steps required by the voter to give a vote and makes the process easier. Schemes and designs are still being proposed with different perspectives and each of them covers an aspect important to e-voting. New applications on blockchain technology are continuously being built and these developments would lead to more future scope of this system.

References

[1] Baocheng Wanga,, Jiawei Suna, Yunhua Hea, Dandan Panga, Ningxiao Lua . Large-scale Election Based On Blockchain .2017 IIKI .

[2] Ashish singh, Kakali Chatterjee ,"Secure electronic voting system using blockchain technology'', International Conference on Computing , power and communication Technologies (GUCON) Galgotias university Greater Noida, UP, India.(Sep28-29,2018).2018 IEEE

[3]Simpson, R. and Storer, T. (2018) Third-party verifiable voting systems: addressing motivation and incentives in e-voting. Journal of Information Security and Applications, 38, pp. 132-138.

[4] Amish Khandelwal, ''Blockchain implementation on E-voting system'', International Conference on Intelligent Sustainable system(ICISS 2019).2019 IEEE

[5] Vijayalakshmi v and Vimal s ,2019 Fifth International Conference on science and technology Engineering and Mathematics (ICONSTEM).2019 IEEE

[6]Kanika Garg , Pavi Saraswat , Sachin Bisht , Sahil kr. Aggarwal , sai Krishna Kothuri , Sahil gupta , ''A Comparative analysis on E-voting system using blockchain'', AKTU , Uttar pradesh , India .2019 IEEE

Voice And Text Based Natural Language Query Processing

Ashwini Kulkarni, Pranali Pawar, Mayuri Khairnar, Shital Patil, Tukaram Gawali
Computer Engineering, SVKM's IOT, Dhule,India
ashu798coolkarni@gmail.com,pranalipawar0110@gmail.com,mayurikhairnar2310@gmail.com,shitalpatil@gmail.com,t.gawali@gmail.com

Abstract—**Traditional query languages needs manual query writing to retrieve a result. In this paper we are going to propose system that aims to develop for users who do not know the database. Language like SQL is facing challenging and difficult situations while accessing or retrieving data. In these system natural language accepts a user as an input in natural language text or via voice input then extracts the necessary information needed for the formation of a query. They sing of natural language processing mapping the query in the English language to SQL after receiving query output generated as a table format. Which predicts which type of query is demand by the user. The information from this output is given to the final query and then it is given to the user on the interface.**

Keywords— SQL, Natural Language Processing, Mapping

I. INTRODUCTION

The main objective of Natural Language Processing is to communicate between humans and computers. This helps users who do not know the structured query language.it means a computer understands the human language used by humans. It is a branch of Artificial intelligence (AI) that is a retrieval machine translation and linguistic analysis. The NLP is an interface to a database system that application accepts a SQL query is creates as SQL query and executes is to retrieved data from relational databases. The results of the retrieved database are a stream of elements. Speech recognition is a machine or program to recognize words or phrases. Which is either spoken or word text.it is better than an excel sheet to stored and retrieved data. User one can view all the table. They can enter a query and the entered query and the query in natural language the query will be executed in SQL query. It describes natural language and query based on a probabilistic context-free grammar to the relational database. These systems are used for placement cell officers who work. on student databases to extract data .and also tourism. Railway reservation, chat robot voice, or textual interactions. Using our current system, we can predict which query the users have requested for select, update, delete, and any other query for that matter. And these system focuses on the resolution of problems arising in the analysis or natural language text or speech, such as syntactic and semantic analysis for a compilation of dictionaries and grammar necessary for such analysis. After this, it will be formatting final SQL query based on its type and execute it.

II. FIGURES AND TABLES

Nowadays data is increasing rapidly. There are lot of new database tools and technologies are growing, hence we can store large data, but the problem is that the technology or an interface which can process data and display the data as per the user demand is not familiarized with many of the people. The user will give voice input, which will be recognized and then converted into text format. The system takes input as spoken query language and dispatch it to the speech recognition engine. The output will be the into query text extracted from the speech. The accurate input query is extracted and sent to token tokenization. Perform preprocessing on text converted from a voice in tokenization. The sentences are broken down into tokens and remove the unwanted tokens. And then text translator translates the main content which is required for query processing. After the execution of the query, the expected output will be displayed on the system in the form table. User who wants to access a database but does not have any knowledge about the database language facing difficulties. Hence there is a requirement for a system that enables the users to retrieve the information in the database. This project aims to develop such a system using NLP by giving structured natural language question as input and receiving SQL query as the output. This project gives a use the regular expression to formation of the query in the natural language such as English to SQL. The system accepts the user input as natural language text or via voice input.

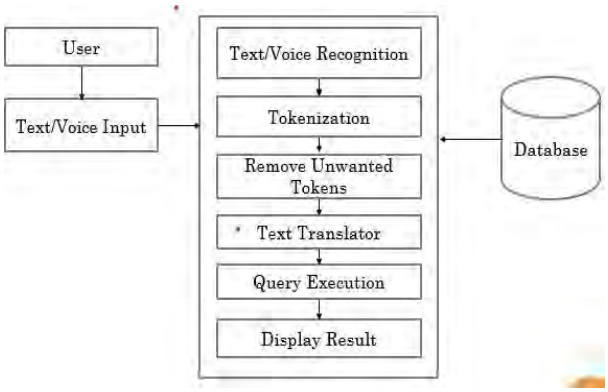


Figure 1. System Architecture

The main goal of this system is to allow communication between the database and its human users using natural language. The use of Natural Language brings ease for any human being. This system will help T&P officer to easily retrieve and manage data from student database using their natural language such as English language. There is no need for the user to learn complex query syntax to retrieve data. The facility to accept the input in speech format makes the system user friendly. For example, the Name of the department that has the maximum students count? So, the query for the above

sentence is Select department from student where MAX (student count) And the output for the above query is:

Table 1. Department name and Student count

Dept Name	Student Count
Computer	68

III. CONCLUSION

The main aim of our system is to allow communication between database and it's human users using natural language. Use of natural language brings easy for any human being. This system will easily retrieve and manage data from database using simple language. There is no need for the user to learn the complex query syntax to retrieve the data. The facility to accept input in speech format and voice based makes system user friendly.

IV. ACKNOWLEDGMENT

We take this opportunity to express our hurtful gratitude towards the Department of computer Engineering, SVKM's Institute of Technology, Dhule that give us an opportunity for presentation of our Project in their esteemed organization.

References

[1] Puja Munde, Sayali Tambe, Afreen Shaikh, Pratiksha Sawant, Prof. Deepa Mahajan, 'Voice-Based Natural Language Query Processing', International Research Journal of Engineering and Technology (IRJET), Volume: 07Issue: 03| Mar 2020.

[2] Aditya Narhe, Chaitanya Mohite, Rushikesh Kashid, Pratik Tade, Santosh Waghmare, 'SQL Query Formation for Database System using NLP', International Journal of Engineering Research & Technology (IJERT), Vol. 8 Issue 12, December-2019.

[3] Uma V, Sneha V, Sneha G, Bhuvana, J, Bharathi B, 'Formation of SQL from Natural Language Query using NLP', International Conference on Computational Intelligence in Data Science (ICCIDS), 978-1-5386- 9471-8/19,2019.

[4] Yasir Ali Solangi, Zulfiqar Ali Solangi, Samreen Aarain, Amna Abro, Ghulam Ali Mallah, Asadullah Shah,' Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis', IEEE International Conference on Engineering Technologies & Applied Sciences, 9781-5386-7966-1/18, 22- 23Nov 2018.

[5] Barack Wamkaya wanjava, Lawrence Muchemi,' Automatic Semantic Network Generation from Unstructured Documents – The Options', IEEE 5th International Conference on Soft Computing and Machine Intelligence, 978-1-7281-1301-2/18,2018.

[6] Anum Iftikhar, Erum Iftikhar, Muhammad Khalid Mehmood, 'Domain-Specific Query Generation from Natural Language Text', IEEE international conference on innovative computing technology (INTECH 2016), 978-1-5090-2000- 3 /1 6 , 201 6

[7] Prasun Kanti Ghosh, Saparja Dey, Subhabrata Sengupta, 'Automatic SQL Query Formation from Natural Language Query', International computer journal of International Conference on Microelectronics, Circuits, and Systems (MICRO-2014), July 2014.

Sentiment Analysis of COVID-19 Tweets

Kartik Rawool, Anurag Tiwari, Rameshta Vishwakarma
Department of Computer Engineering,Thakur College of Engineering & Technology,Mumbai, India
kartikr2k133@gmail.com, anurag219tiwari@gmail.com, vinayakvishwakarma97@gmail.com

Abstract—**The widespread increase of the COVID-19 has witnessed a remarkable increment in the number of tweets about it. An insight into this tweet data would easily unveil how the general population is perceiving this crisis. This paper focuses on examining the sentiments behind the tweets, in an attempt to gauge whether the overall sentiments of a population are positive or negative. The approach utilizes scraping tweets from March to May, to perform sentiment analysis on them. The two distinct types of Natural Language Processing techniques - Lexicon based and Machine Learning algorithms, have been analyzed and compared to identify the best-suited method for sentiment analysis. In the absence of traditional training and testing datasets, the Sentiment 140 dataset from Stanford university was utilized to train the models to predict sentiments of the COVID-19 tweets. The algorithms utilized included Logistic Regression, Linear SVC, and Multinomial Naïve Bayes wherein the highest accuracy was exhibited by SVC. The overall findings hint towards improved accuracies by using machine learning approaches to classify tweets as either positive or negative. Conclusively, a substantial number of tweets were classified as negative with the people associating words such as "million, china, wear, surgical, people, died" with negative sentiments and "gloves, spread, money, wash hand, public health" with positive sentiments.**

Keywords— Twitter, sentiment analysis, Lexicon, Natural Language Processing (NLP), Linear Support Vector Classifier (SVC), Naïve Bayes, Logistic regression, COVID-19

I. INTRODUCTION

Bearing pertinence to the situation of the coronavirus across the world, there has been a surge in the usage of online platforms and social media sites. One such site, Twitter, most popularly known for its limited sized tweets, has been one of the major platforms where people, government, and a lot of others can communicate effectively. After the announcement of lockdown in India, people took it to Twitter to communicate their reactions and their stand on the same. Various agencies, have to keep in mind people's opinions and gain insight into their reactions to the various issues of national and international importance. The research paper bears resemblance to the same and tries to extract the overall sentiments of people about the COVID-19 and lockdown situation in India by analyzing their tweets.

The sentiment analysis of people after the extension of lockdown announcements is to be analysed with the relevant #tags on twitter and build a predictive analytics model to understand the behaviour of people if the lockdown is further extended. We intend to develop a twitter sentiment analysis model to understand the following: 1. Get to know people's sentiment towards the pandemic, 2. Understand the sentiments of people in the with the government's decision to extend the lockdown

For this project, we shall be utilizing the various Natural Language Processing Techniques. These techniques are majorly two distinct types – Lexicon based and Machine Learning based. We intend to use both the techniques in order to calculate a sentimental score and assign a tweet to a specific class between positive and negative. The research here aims to build a model that can accurately disseminate an understanding of the sentiments of people through only the tweet data. We intend to not only compare and contrast the existing methods but also aim to explore the Word2Vec method, in the future, which has shown many improvements in sentiment analysis lately [1].

II. LITERATURE SURVEY

Algorithms discussed in [2] for SA are supervised ML algorithms like maximum entropy, SVM, Naïve Bayes, KNN, and unsupervised ML algorithms such as HMM, Neural network, PCA, ICA, SVD. Paper [3] has likewise concentrated around sentiment analysis of a YouTube video by removing and breaking down the YouTube comments. Survey paper [4] presented briefly regarding many recently proposed algorithms enhancements and various SA applications. To enrich the feature set for sentiment analysis, in [5] the enlargement of the intersection set between the SentiWordNet and the corpus vocabulary is discussed. Authors in [6] have built up their own specific manner of handling sentiment analysis through building their own database which incorporates grammatical features, positives, and negative words which gave us the motivation to build up our own thought. The authors of [7] have targeted to solve the inaccuracy and discrepancies present in the current text and image classifiers for geo-spatial sentiment analysis of disaster-related data objects.

Researchers of [8] use a single approach upon a variety of topics in order to fetch the sentiments of tweets pertaining to specific topics. The Stanford NLP classifier has been applied once tweets are fetched from Twitter. The major highlight of the application of sentiment analysis is observed in [10] wherein tweets for KFC and Mcdonald's are fetched and sentiment analysis on the same are performed to find out the more popular place among the two. The paper utilizes supervised ML techniques. For lexically analyzing a Twitter message, each token needs to be identified on a case by case basis before preprocessing the tweets dataset. The author of [11] investigates various techniques for lexical normalization of Twitter data and presents the findings as the techniques are applied to process raw data from Twitter. The main goal of the paper [12], is to extract, refine, and analyze the tweets further visualizing the user's tweet in a particular area according to the location information. In [13], artificial intelligence-based real-time disaster response system Disastro, which assists the volunteers by identifying the relevant tweets from the real-time twitter data and classifying them under the domains "rescue" and "donation", has been proposed to solve the same problem. The adaptive training method is proposed to address this problem [14].

In this method, non-textual features are also extracted from tweets for training the algorithm, which classifies the tweets of different topics as positive, negative, neutral. A survey of different approaches of clustering with respect to sentiment analysis is presented in [15] and a way to find relationships between the tweets on the basis of polarity and subjectivity is discussed as well. The power of public sentiments on predicting the success of movies is discussed in [16]. The data is taken from Youtube and IMDb for categorizing the comments as positive or negative and thereby predicting the success of movies. The paper [17] evaluates the people's sentiment about a person, trend, product, or brand, conducting visualization in the form of histograms and pie charts.

We can observe a different highlight in [18] as the paper focuses on Sentiment analysis, Feature-based Sentiment Classification, and Opinion Summarization, as a whole. The fact that the analysis is done in real-time is the major highlight. Positive or negative sentiment on Twitter posts is done in [19] using a well-known machine learning method for text categorization. In addition, manually labeled tweets are used to build a trained method to accomplish a task. A useful application of sentiment analysis is highlighted in [20] where twitter posts about electronic products like mobiles, laptops are analyzed and the classification accuracy of the feature vector is tested using different classifiers like Naive Bayes, SVM, Maximum Entropy, and Ensemble classifiers. Performing sentiment analysis on the Twitter dataset, the researchers in [21] are able to predict the current affairs and user's behavior as well as opinion. The method used is the N-grams model which is a new algorithm developed for the system that will combine Unigram, Bigram, and Trigram data.

III. Methodology

A. Dataset Description

For the problem undertaken, a requirement to extract the dataset in the form of tweets from Twitter is needed. The set of Tweet IDs from [22] was selected to get the tweets as it includes the tweets pertaining to COVID-19 from January to May. Now, in order to extract data from these Tweet IDs, we used the Twarc command-line tool. This tool is used to use Twitter API through Python with ease. By hydrating the tweet IDs various details of the tweets were gathered that were stored in a CSV file. After utilizing the Twarc command-line tool the received the dataset contained the following attributes. One tweet had over 34 features as follows: created_at, id, id_str, full_text, truncated, display_text_range, entities, source, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, in_reply_to_screen_name, user, geo, coordinates, place, contributors, retweeted_status, is_quote_status, quoted_status_id, quoted_status_id_str, quoted_status_permalink, retweet_count, favorite_count, favorited, retweeted, lang, extended_entities, possibly_sensitive, quoted_status, withheld_scope, withheld_copyright, withheld_in_countries. The extracted dataset contains 90540 tweets.

B. Preprocessing

After the implementation of data collection, the next crucial step was to pre-process the data. From the data collected, we needed to remove the various inconsistencies and irrelevant features from the dataset. This needed to be performed in order to ensure that the results obtained from our algorithm were appropriate as well as plausible. As an algorithm is constituted, we need to keep in mind the type of data it requires as the input. In our case, as there were 35 columns of various features, and the text of the tweet was not present in the format required, therefore, the following pre-processing steps were performed. 1. Listing down the features and checking for null values among them, 2. Selecting only the relevant features, 3. Filtering out tweets from the English language, 4. Removing irrelevant text from the tweets, 5. Creating a final data frame with concise data of tweets.

Upon the first retrieval of data from the original dataset, we obtained a set of 35 features where over 20 features had missing values or contained null values. Since no imputation technique can be applied here, the apparent choice was to drop these features. From a perspective of sentiment analysis, these features were not relevant as well. As a natural tendency, we reduced these features in order to make our dataset more concise and consisting only of relevant tweet's data. This ensured that our model had a focus only on those features which really depicted sentiments of the people through their tweets and no other data which would essentially be a reason for poor performance or extremely long operations in the algorithm. The filtered tweet's dataset consists only of the full_text of the tweets and 90540 rows over which sentiment analysis is to be conducted.

	full_text
0	@charliekirk11 Agree too much coverage of Coro...
1	RT @TheRickyDavila: With the Coronavirus clear...
2	RT @BharatD55138223: #Secrets_Of_Vedas\nReside...
3	RT @misayeon: Dahyun has donated 50 million wo...
4	RT @xx_Y4YA: So the coronavirus is in Houston ...

Fig. 1. Tweet's dataset with a relevant feature only

As a preliminary step, our first instinct was to filter out tweets from the English language only. Once we filtered out the relevant features along with the tweets only in English, the following preprocessing to the tweet's data was performed:

- Converting the tweet's text to lowercase – to standardize the tweet's text across all the rows.
- Removing punctuations – any punctuations do not add improvement in extracting the sentiments of the tweet.
- Removing hash symbols – hashtags contain information about the latest ongoing trends but the hash symbols in themselves can be filtered out to make the tweet's text more concise.
- Eliminating the twitter handles mentioned in the tweet beginning with '@' – the tweets which mention other accounts or reply to another tweet often use this, which in our case seems irrelevant in extracting the tweet's sentiments.

- Eliminating the 'rt' i.e. retweet status from the tweets – retweeting on twitter means sharing a tweet on your wall and this information as well does not add any value to the sentiment score, we would be calculating later on.
- Removing URLs from the tweet – any URL or a link to another page is irrelevant to the process of analyzing sentiments and therefore would be required to be eliminated.

As an example, we had the following tweet in the dataset.

"RT @treasonstickers: .@realDonaldTrump when are you going to visit and embrace the people in the US with Coronavirus? https://t.co/lEvq3S4iâ€¦"

As we can clearly see, the twitter handles, @realDonaldTrump and @treasonstickers are just holding up more space in the text of the tweet and do not have any resemblance or usage in gauging the sentiment of the tweet. The link at the end, along with the "RT" in the beginning also do not constitute anything relevant to the tweet which could be helpful in processing a sentimental score. Therefore, after performing the preprocessing steps enlisted above, the following new tweet was rendered.

"when are you going to visit and embrace the people in the us with coronavirus"

This looks much more sensible as it has a clear, concise meaning to it. The tweet can be easily used for sentiment analysis as it has a limited number of characters after all the irrelevant ones have been filtered out. Therefore, all the tweets in the dataset were pre-processed in the same manner and we obtained the following data frame as the result.

	tweets
0	s korea is giving free tests to citizens and ...
3	when are you going to visit and embrace the p...
4	glamorchina delicate ottelia blooms in lugu l...
6	s korea is giving free tests to citizens and ...
7	italy repos a 50 increase in confirmed coronav...

Fig. 2. Final processed dataset with compact tweets

As a next preprocessing step, we were required to remove stop words from the tweet's text. In natural language processing, useless words are referred to as stop words. A stop word is a commonly used word (such as "the", "a", "an", "in") that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. The tweets yielded do not also require these words as they hold little relevance in gauging the sentiment and context of the meaning of the word. However, we needed to be extremely careful in choosing the right stop words to eliminate. This is due to the fact that some words can completely change the meaning of a sentence even though at a first glance they do not look relevant. We, therefore, defined a list of our own stop words which we required to eliminate. These are the defined stop words: "i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself","yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself","they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these","those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do","does", "did", "doing", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "while","of","at", "by", "for", "with", "about", "between", "into", "through", "during","before","after", "above", "below", "to", "from", "up", "down", "in", "out","on", "off", "over", "under", "again", "further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each","few", "more", "most", "other", "some", "such", "only", "own", "same", "so","than","too", "very", "s", "t", "can", "will", "just", "don", "should", "now".

Defining our stop words was a critical task since our entire tweet's sentiment gauging depended upon the choice of words which was being fed to the natural language processing model and upon which it was analyzing the tweets. The words which were included in the stop words are:

1) Pronouns: Pronouns included 'I, me, myself, ourselves, themselves, her, his' and many others. These pronouns don't necessarily add much value in a tweet's sentiments or were helpful in any way. So, it is a very safe choice to eliminate these pronouns as they are only taking up more space and computational time and are not adding any value to the dataset.

2) Prepositions: words such as 'above, below, on' are known as prepositions. A preposition is a word placed before a noun or pronoun to form a phrase modifying another word in the sentence. Therefore, a preposition is always part of a prepositional phrase. It acts in modifying another noun or adjective but is not necessarily useful in calculating a sentiment from a tweet. As an example, we can consider the following tweet: *"impoantly both the canadian amp australian isolates from the above screenshot have recent travel history to iran 2 2".* After cleaning and preprocessing the tweet we got the following new tweet: *"impoantly canadian amp australian isolates screenshot recent travel history iran 2 2".* Even though we have removed the preposition 'above' from the tweet, we can clearly identify that the tweet has not entirely changed the meaning and is still very meaningful with the context it wants to communicate.

3) Conjunctions: A conjunction joins words, phrases, or clauses, and indicates the relationship between the elements joined. These include the words "and, but, or, so" and others. It is quite relevant to see that eliminating these words is not going to change the contextual meaning of the tweet as they are essentially just connectors. Connectors are used to combine and join, two or more sentences or ideas, and therefore, it is a prudent choice to eliminate them as they are not adding or changing the tweet's meaning.

4) Articles: Articles include "a, an, the". Articles are most definitely used to refer to nouns and state them qualitatively. The articles don't necessarily give us any

useful meaning in terms of a tweet's sentiments as they are more focused upon building the noun. Therefore, we can eliminate and remove them without the doubt of losing contextual meaning from the tweets.

Words which we have not eliminated include: "no", "not", "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't". This is done because eliminating these words can flip the entire meaning of the sentence leading to a sentiment that is wrongly classified. To understand better, we can consider the following tweet: "*new england journal of medicine coronavirus could be no worse than flu".* The tweet essentially conveys that coronavirus isn't as dangerous as it seems and is comparable to flu. The phrase "no worse than" is comparing flu to coronavirus and is not in any way portraying a negative sentiment with respect to coronavirus. If someone had to manually classify this tweet then it is likely to be classified positive as it eases out a tension and compares coronavirus to flu.

However, if we had removed the word "no" then the entire tweet changes to: *"new england journal of medicine coronavirus could be worse than flu".* There remains no possibility that this tweet is not negative now. It's entire meaning as well as sentiment has changed entirely. If someone had to classify this tweet then there is a 100% chance that they shall classify it as negative. As we can clearly see that the words, we choose to eliminate can have drastic impacts on the classification of tweets. Hence, we have manually created the stop words list in order to eliminate these inconsistencies from the data.

C. Training and Testing Dataset

The scraped data was completely unlabelled, meaning there were just the tweets present but no score on the sentimental value. This meant that we could not use a traditional training testing approach for the model to be used. For machine learning methods, we required a model that had been previously trained on a dataset to classify our new tweets. Firstly, the test dataset will be used to check the accuracies of the lexicon-based as well as machine learning-based models. The model with considerable accuracy was then used to classify sentiments on the scraped dataset. Before beginning to train a new dataset, we had to precisely choose a dataset consisting of tweets which were already labeled. We chose the "Sentiment140", which originated from Stanford University. It consisted of 1.6 million tweets labeled positive (4) or negative (0).

As stated earlier, there was a need to preprocess this dataset in order to extract only the tweet's text and their target values. The raw data is majorly not useful as our model shall focus only on the sentimental analysis of tweets and wouldn't procure to classify it any further. Cleaning essentially means we remove URLs, hashtags, stop words, Unicode encoding characters, punctuations, and convert the scale of the target tweets to 0 and 1 indicating negative and positive tweets respectively. The cleaned dataset obtained was as follows.

	text	target
0	awww that bummer you shoulda got david carr of...	0
1	is upset that he can not update his facebook b...	0
2	dived many times for the ball managed to save ...	0
3	my whole body feels itchy and like its on fire	0
4	no it not behaving at all mad why am here beca...	0

Fig. 3. Cleaned dataset for accuracy testing

It is a matter of significance to check the dimensions of our dataset again after cleaning to see whether any of the values have been changed completely or have been eliminated to the point of introducing missing values in our dataset. We, therefore, checked for missing values in our dataset. As we had expected some tweets went missing. This possibly happened during the cleaning of the tweets where characters, punctuations, and other strings were removed. However, in a dataset consisting of 1.6 million values, 4 thousand values are not so significant. Therefore, they are dropped from the dataset giving us a final clean dataset with a shape of (1596041,2).

A word cloud for the negatively classified tweets was then plotted. A word cloud essentially gives a visual representation of the different words used in a particular class whose size is proportional to their frequency in the data.

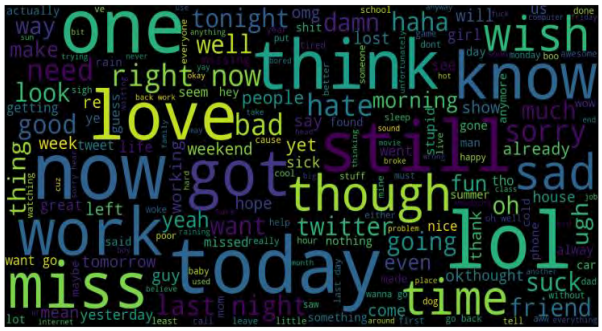


Fig. 4. Word cloud for negative words

As we can identify from the word cloud, the words most frequently used were "know, love, though, wish, today, miss" and others. These words represent the ones that were most frequently occurring in the class of negative tweets. A word cloud for the words associated with the positive sentiments plots the following figure.

Fig. 5. Word cloud for positive tweets

As we can observe the words here include "love, thank, wow, nice, haha". These words are clearly associated

with positive sentiments and we can observe that our dataset has classified them correctly.

D. Model Selection

As a primary step, we shall be using natural language processing techniques in order to get scores of the tweets. We identified 2 major approaches wherein we could explore the lexicon based methods. 1. Text Blob, 2. VADER. In order to begin the actual score calculations for our original dataset, the first decision was to check upon the accuracies of the two methods on a predefined and already classified dataset. This was done to understand the most suitable method which could give us reliable sentimental scores for our tweets which could act as a training dataset for us. The dataset which we chose to compare and analyze the different approaches was first pre-processed to remove stop words, punctuations, and make all the text to lowercase. The final dataset was as in figure 4. Here positive tweets stand for 1 whereas negative is represented by 0. We then checked for the balance of classes with respect to their classification of positive or negative tweets.

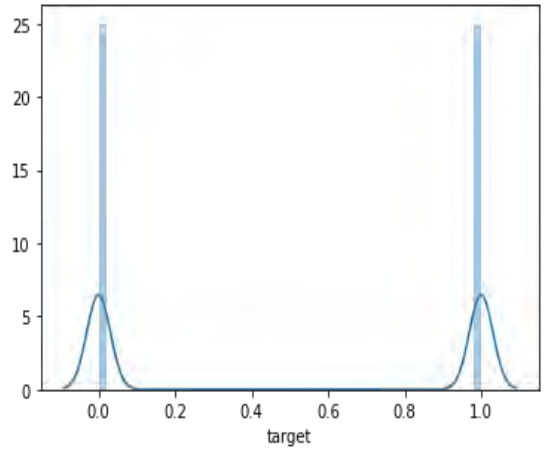


Fig. 6. Distribution of positive and negative classes

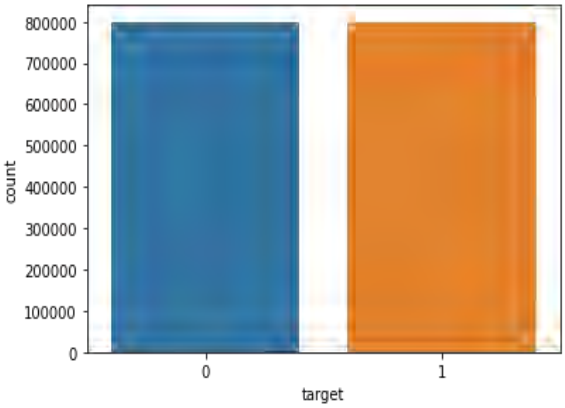


Fig. 7. Equally distributed classes of sentiments

We concluded that the dataset we had extracted was balanced. Therefore, we applied the two mechanisms as discussed earlier.

Text Blob is a python library for processing textual data. 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. We utilized the sentiment analysis part from the Text Blob library in order to test its accuracy on the dataset. Text Blob gives us a range of sentimental scores instead of a binary value. This range lies between -1 and 1 where the negative values are tweets with a negative sentimental impact and the tweets with a positive value are tweets with a positive impact. However, for our dataset, the only acceptable values were either 0 or 1. Therefore we had to change the values received by using the following function. If the score was less than zero then it returned zero and else it returned one. After utilizing the above function on the calculated values given by Text Blob, we finally compared the values with the actual data and plotted the following graph. As it can be seen that the correctly classified results were close to 1000000 out of 1600000, we calculated the actual accuracy and it turned out to be 62.59%.

Therefore, using Text Blob for labeling our training dataset would yield an accuracy that could be close to this number, if not exact.

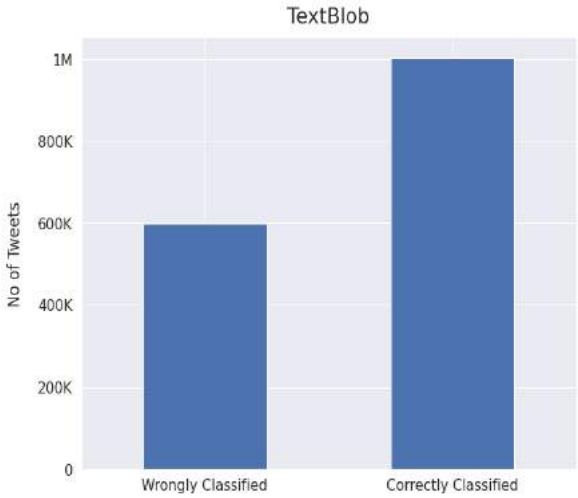


Fig. 8. Result classification by TextBlob

VADER Sentiment Analysis is a powerful open-source tool designed for analyzing the sentiment expressed in social media. VADER stands for Valence Aware Dictionary and sentiment Reasoner. It is a simple lexicon and rule-based model for general sentiment analysis. A key difference between the two is that VADER was designed with a focus on social media texts. This means that it puts a lot of emphasis on rules that capture the essence of text typically seen on social media — for example, short sentences with emojis, repetitive vocabulary, and copious use of punctuation (such as exclamation marks). It is therefore much more suitable and applicable for raw data than a complete refined and pre-processed dataset.

TABLE I. VADER CLASSIFICATION REPORT

Algorithm	**Precision**		**Recall**		**F1 Score**		**Accuracy**
	Negative	***Positive***	***Negative***	***Positive***	***Negative***	***Positive***	

VADER on the Preprocessed dataset	0.81	0.61	0.44	0.90	0.57	0.73	0.67
VADER on raw dataset	0.81	0.61	0.43	0.90	0.56	0.73	0.66

According to VADER classification, we obtained an accuracy of 67% which indicated better classification than TextBlob. We expected a better accuracy on VADER when the dataset applied to it was rawer and less preprocessed since it has been designed for slangs and text on the internet. To test our hypothesis, we performed VADER on the raw data as well and obtained the results which falsified our hypothesis. It was not entirely correct as there is not a considerable variation in the accuracy and f1-score of the model when it has been utilized on raw data instead of preprocessed one.

In order to utilize machine learning models on the dataset, a traditional training dataset was required. To test the accuracies of machine learning models in sentiment analysis, we used our test dataset and gathered accuracies for selected models. Therefore, the next step was to build and test models on this training dataset. Before we can train any model, we first consider how to split the data. Here we chose to split the data into two chunks: train and test. The ratio to split the data is 98/2, 98% of data as the training set, and 2% for the test set. The rationale behind this ratio comes from the size of the whole data set. The dataset has more than 1.5 million entries. In this case, only 2% of the whole data gives us more than 30,000 entries. This is more than enough to evaluate the model and refine the parameters.

The next step for textual data is to vectorize the data and check the frequency of the words. There are two methods that can vectorize the data – count vectorization and TFIDF vectorization. To use text in machine learning algorithms, it needs to be converted to numerical representation. One of the methods is called the bag-of-words approach. The bag of words model ignores grammar and order of words.

With the count vectorizer, we merely count the appearance of the words in each text. For example, let's say we have 3 documents in a corpus: "I love dogs", "I hate dogs and knitting", "Knitting is my hobby and my passion". If we build vocabulary from these three sentences and represent each document as count vectors, it will look like the pictures below.

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

Fig. 9. Count Vectorization example

But if the size of a corpus gets big, the number of vocabularies gets too big to process. With the 1.5 million tweets, if we build vocabulary without limiting the number of vocabularies, we will have more than 260,000 vocabularies. This means that the shape of the training data will be around 1,500,000 x 260,000. This sounds too big to train various different models. So, we decided to limit the number of vocabularies. TFIDF is another way to convert textual data to the numeric form and is short for Term Frequency-Inverse Document Frequency. The vector value it yields is the product of these two terms; TF and IDF. Relative term frequency is calculated for each term within each document as below.

$$\text{TF(t, d)} = \frac{number\ of\ term(t)\ appears\ in\ document(d)}{total\ number\ of\ terms\ in\ document(d)} \quad (1)$$

Next, we need to get Inverse Document Frequency, which measures how important a word is to differentiate each document by following the calculation as below.

$$\text{IDF(t, D)} = \log\left(\frac{total\ numbers\ of\ documents(D)}{number\ of\ documents\ with\ the\ term(t)\ in\ it}\right) (2)$$

The number of occurrences for the non-topic bearing terms will be significantly higher than any other term while using count vectorization. This will force them to have the highest weight in the model simply due to their high occurrence and will skew our model.

The way to combat this problem is to use TFIDF. TFIDF balances out the term frequency (how often the word appears in the document) with its inverse document frequency (how often the term appears across all documents in the data set). This means that common words will have very low scores as they'll appear in all documents in our set. TFIDF will give higher scores to the distinct words and thus they'll be the ones that the model identifies as important and tries to learn on the training dataset. Therefore, we utilize TFIDF vectorization for all the models which we evaluate. The next step before we start training and evaluating models will be to see whether we train the models with unigram, bigram or trigram texts. In other words, n-grams are simply all combinations of adjacent words or letters of length n that you can find in our source text. In [23], the logistic regression accuracy for all three are performed and the results obtained are plotted below.

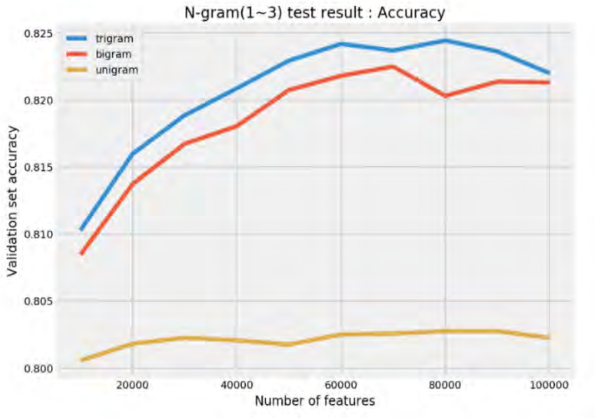


Fig. 10. Validation set accuracy graph on Unigram, Bigram, and Trigram

According to [23], the results obtained were unigram: 80,000 & 90,000 features at validation accuracy 80.28%, bigram: 70,000 features at validation accuracy 82.25%, trigram: 80,000 features at validation accuracy 82.44%. Therefore, we can conclude that we shall be using trigrams for the model training along with TFIDF vectorization.

IV. Result And Discussion

Once we are done in deciding the preprocessing and vectorization required for the training dataset, we move on to training different algorithms on the dataset and evaluate them comparatively. Based upon their accuracies, we shall be choosing the best model and using the same, we shall be predicting the sentiments on our tweet's dataset. The following algorithms for comparisons have been utilised.

- Logistic Regression with count vectorization
- Logistic Regression with TFIDF vectorization
- Linear Support Vector Machine (Linear SVC)
- Linear Support Vector Machine with L1 regularization
- Multinomial Naïve Bayes

We have intentionally avoided the use of algorithms such as Random Forest or Decision Tree Classifier because, with a dataset of such gigantic dimensions, these algorithms won't be computationally effective or feasible. The classification report for Logistic Regression with count vectorization and TFIDF vectorization is first checked.

As we can check the classification report, we can conclude that both the methods are giving us exactly the same results with one or two percent difference in precision and recall. However, the accuracy parameter is the same for both, and contrary to our hypothesis the count vectorization performed as good as TFIDF. Considering the feasibility as well as the functionality of TFIDF in vectorizing the uncommon words and scaling them to a suitable count, we shall be utilizing and preferring it over count vectorization.

TABLE II. Classification Report of LR with Count and TFIDF Vectorization

Algorithm	Precision		Recall		F1 Score		Accuracy
	Negative	*Positive*	*Negative*	*Positive*	*Negative*	*Positive*	
LR Count Vectorization	0.83	0.81	0.80	0.84	0.82	0.83	0.82
LR TFIDF	0.84	0.81	0.80	0.85	0.82	0.83	0.82

Next, we check the accuracy of the Support Vector Machine classifier with L1(Lasso) and L2(Ridge) regularization. The key difference between these techniques is that Lasso shrinks the less important feature's coefficient to zero thus, removing some features altogether.

As we can observe, the accuracy for both the classifiers do not differ much and are almost comparable. This means that both the L1 and L2 penalties work effectively on our training dataset. As a final step, we check the Multinomial Naïve Bayes classifier.

The accuracy of the Naïve Bayes is comparatively lower by 2% than the other algorithms. This is quite a contrasting result as Naïve Bayes is generally and very popularly known to perform well on textual data. We can visualise the accuracies of the different models by the table and graph below.

As we can see that Linear SVC and LR is giving us almost comparable results. We need to choose between the two. Therefore, we chose our model as SVC because it has more flexibility in hyperparameter tuning in order to increase its accuracy. This would prove critical in the later stage if we shall be using the tuning for getting the absolute optimum parameters of c, kernel and gamma.

TABLE III. Accuracies of Different Algorithms

Serial No.	**Machine Learning Algorithms**	**Accuracy**
1	LR Count Vectorization	82%
2	LR TFIDF	82%
3	Linear SVC (L2)	82.33%
4	Linear SVC (L1)	82.40%
5	Multinomial NB	80.21%
6	TextBlob	62.5%
7	Vader	69%

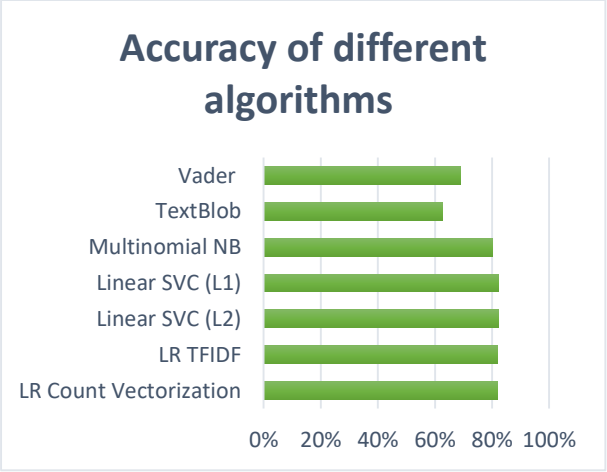


Fig. 11. Accuracies of different algorithms

	text	sentiment
0	agree too much coverage of prep taking time fr...	0
1	with the clearly spreading in the us now know ...	0
2	secrets of vedas residents of america iran chi...	0
3	dahyun has donated million won to help prevent...	1
4	so the is in houston methodist hospital but th...	0

Fig. 12. Tweet's scores Classified with SVC

From the analysis we conclude that we need to apply the SVC model to classify our sentiments. Further we applied our SVC model which was trained on the earlier dataset imported. Based on the prediction this is the result obtained in the new dataset. Once the tweets are classified as positive and negative the words present in the tweets were analysed. We now print a word cloud again for the negative and positive tweets in order to understand the type of data that has been classified in either of the two classes.

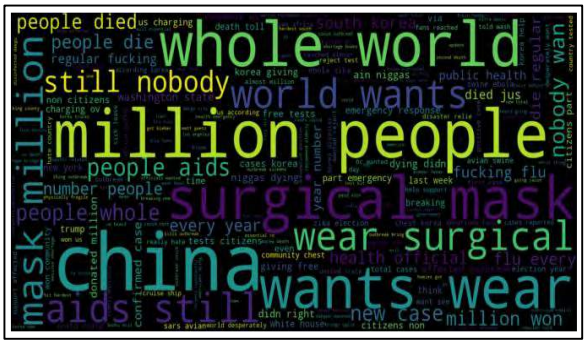


Fig. 13. Negative Word Cloud

The words "million, china, wear, surgical, people, died" are clearly highlighted. This indicates that tweets which were more closely associated with negative sentiments are prone to using these set of words more often. Next, we plot the positive word cloud.

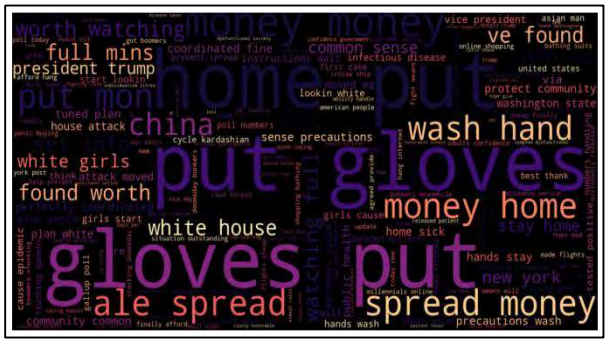


Fig. 14. Positive Word Cloud

In the positive word cloud, the words such as "gloves, spread, money, wash hand, public health" are evident. These words have been used more frequently in the tweets which have been classified as positive. As a human instinct, it is quite obvious as well that tweets which increase public awareness for example to wash hands or wear gloves are having a positive sentiment in their delivery.

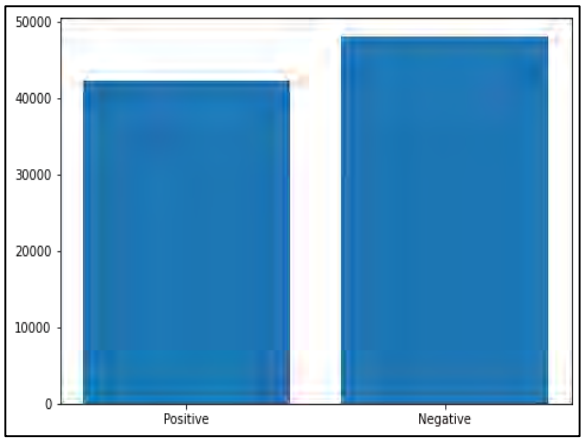


Fig. 15. Distribution of positive and negative tweets

The graph clearly highlights that majority of the tweets were negative. This means that the people's opinion with respect to the topic "COVID-19" upon which the data was scraped, is in general more negative. Therefore, the model is predictably displaying the sentiments of people through their tweets by classifying them as positive and negative.

V. Conclusion

We created a twitter sentiment analysis model in this project which aimed to classify a given set of tweets as positive or negative. In order to begin, we had to undergo several variations in the training and testing dataset as we had no clearly predefined dataset for this problem available which can be used directly. Therefore, after using a training dataset that was provided by the Stanford library, we implemented different machine learning models and trained them. Upon a general comparison of two lexicon-based methods and several machine learning-based models, we concluded that the machine learning models were more accurate and thus we implemented SVC upon our dataset. We finally concluded with analyzing sentiments of the tweets wherein we found out that the sentiments of people towards the COVID-19 crisis displayed by their tweets were in general more negative than positive.

VI. Future Scope

We realize that no calculation can't be 100% exact in foreseeing the outcomes. In our venture, we have attempted and tried distinctive algorithms to expand the precision so our model can be increasingly productive in investigating the sentiments of a text. The future extent of the venture is to recognize the sarcastic tone in a text. Many algorithms are unable to do this. It's still under process. We have explored the idea of utilizing the word2vec library which changes over a word into a 2-dimensional vector. This is an idea that we are anticipating to take a shot in examining sentiments of text. The thought about expanding this project to a commercial scale where companies can use this to predict customer satisfaction is also to be explored.

References

[1] Bai Xue, Chen Fu, Zhan Shaobin, "A Study on Sentiment Computing and Classification of Sina Weibo with Word2vec", 2014 IEEE International Congress on Big Data

[2] Supriya B. Moralwar and Sachin N. Deshmukh, "Different Approaches of Sentiment Analysis," International Journal of Computer Sciences and Engineering Vol.-3(3), PP(160165) Mar 2015, E-ISSN: 2347-2693

[3] A. Manjula1, A. Rama Mohan Reddy, "Sentiment Analysis on Social media", International Journal of Computer Engineering In Research Trends, 6(11):pp1-6, November 2019

[4] Walaa Medhat, Ahmed Hassan and Hoda Korashy," Sentiment analysis algorithms and applications: A survey," Ain Shams Engineering Journal Volume 5, Issue 4, December 2014, Pages 1093-1113

[5] Eissa M.Alshari, Azreen Azman∗ , Shyamala Doraisamy, Norwati Mustapha and Mostafa Alkeshr, "Effective Method for Sentiment Lexical Dictionary Enrichment based on Word2Vec for Sentiment Analysis", 2018 Fourth International Conference on Information Retrieval and Knowledge Management

[6] Biswarup Nandi, Mousumi Ghanti, Souvik Paul, "TEXT BASED SENTIMENT ANALYSIS", Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017) IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

[7] Abdullah Alfarrarjeh, Sumeet Agrawal, Seon Ho Kim, Cyrus Shahabi, "Geo-spatial Multimedia Sentiment Analysis in Disasters", 2017 International Conference on Data Science and Advanced Analytics

[8] C.J. Hutto, Eric Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text", Association for the Advancement of Artificial Intelligence 2014

[9] Miss. Shital Anil Phand, Mr. Jeevan Anil Phand, "Twitter Sentiment Classification using Stanford NLP", 978-1-5090-4264-7/17/$31.00©2017 IEEE

[10] Sahar A. El_Rahman, Feddah Alhumaidi AlOtaibi, Wejdan Abdullah AlShehri, "Sentiment Analysis of Twitter Data", 978-1-5386-8125-1/19/$31.00 ©2019 IEEE

[11] Bilal Ahmed, "Lexical Normalisation of Twitter Data", Science and Information Conference 2015 July 28-30, 2015 | London, UK

[12] Neha Garg, Rinkle Rani, "Analysis and Visualization of Twitter Data using k-means Clustering", International Conference on Intelligent Computing and Control Systems ICICCS 2017

[13] KRISHNA KANTH A, ABIRAMI S, CHITRA P, GAYATHRI SOWMYA G, "Real Time Twitter based Disaster Response System for Indian Scenarios", 2019 26th International Conference on High Performance Computing, Data, and Analytics Workshop (HiPCW)

[14] K Lavanya, C Deisy, "Twitter Sentiment Analysis Using Multi-Class SVM", 2017 International Conference on Intelligent Computing and Control (I2C2'17)

[15] Shreya Ahuja, Gaurav Dubey, "Clustering and Sentiment Analysis on Twitter Data", 2017 2nd International Conference on Telecommunication and Networks (TEL-NET 2017)

[16] Quazi Ishtiaque Mahmud∗ , Asif Mohaimen, Md Saiful Islam‡ , Marium-E-Jannat, "A Support Vector Machine mixed with statistical reasoning approach to predict movie success by analyzing public sentiments", 2017 20th International Conference of Computer and Information Technology (ICCIT), 22-24 December, 2017

[17] Prakruthi V, Sindhu D, Dr S Anupama Kumar, "Real Time Sentiment Analysis Of Twitter Posts", 3rd IEEE International Conference on Computational Systems and Information Technology for Sustainable Solutions 2018

[18] M.Trupthi, Suresh Pabboju, G.Narasimha, "SENTIMENT ANALYSIS ON TWITTER USING STREAMING API", 2017 IEEE 7th International Advance Computing Conference

[19] Peiman Barnaghi and John G. Breslin, Parsa Ghaffari, "Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment", 2016 IEEE Second International Conference on Big Data Computing Service and Applications

[20] Neethu M S, Rajasree R, "Sentiment Analysis in Twitter using Machine Learning Techniques", IEEE - 31661

[21] Dhruvi K. Zala, Ankita Gandhi, "A Twitter Based Opinion Mining to Perform Analysis Geographically", Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)

[22] E. Chen, "echen102/COVID-19-TweetIDs," GitHub. [Online]. Available: https://github.com/echen102/COVID-19-TweetIDs. [Accessed: 27-Aug-2020].

[23] R. Kim, "Another Twitter sentiment analysis with Python - Part 5 (Tfidf vectorizer, model comparison...," Medium, 13-Jan-2018. [Online]. Available: https://towardsdatascience.com/another-twitter-sentiment-analysis-with-python-part-5-50b4e87d9bdd. [Accessed: 27-Aug-2020].