Predicting User Movement Frequent Patterns in Mobile Service Environment

Rajkumar K[1], Nageswari[2]

[1]Lecturer & Head, School of Computer Science and Information Technology,
DMI-St John the Baptist University, Malawi, Central Africa.
[2]Lecturer, School of Computer Science and Information Technology,
DMI-St John the Baptist University, Malawi, Central Africa.
Email: [1]rajkumarengg2020@gmail.com, [2]nageswari50@gmail.com

Abstract— Mobile users can be attracted towards the services through their mobile gadgets by means of "Information Service and Application Provider (ISAP)" anytime around the world. When customers move within the cell network, their carrier requirements based totally on the locations use to be tracked in a central mobile user transactional database. The systems that provide mobile service offer users valuable data through mobile gadgets. Depending on volatile user movement behavioral patterns, mobile provider structures have the potential of successfully mining a unique request from abundant information. In this paper, consumer movement and conduct patterns are studied with respect to the hassle of predicting identical cell access styles based on joining the subsequent three forms of characteristics: location (L), timestamp (T), and services (S). Traditional cell carrier structures are inadequate in handling delicate consumer movement conduct pattern without taking L, T, and S. In proposed system of this paper, FP-Growth set of rules is used to locate the frequent styles. It stores the frequent patterns in a tree and hash statistics structure to calculate the prediction ratio. Prediction ratio is used to locate the match of asked services.

Keywords - Mobile Transaction Database, Movement Behavior Patterns, Mining, Prediction Ratio.

1. INTRODUCTION

In general, most of the users are moving and accessing wireless services for the activities, which they are currently using in a mobile environment. We have proposed an idea to intricate activity for characterizing continuously changing complex behavioral patterns of the movable users. For this purpose of data management, an intricate activity is modeled for a series of location movement, service requests, co-occurrence of location and service, or the interleaving of all above. An activity may be collection of sub-activities. Different activities may reveal dependencies that affect the user behaviors. We argue that the intricate activity concept provides a more precise, rich, and detail description of user behavioral patterns, which are invaluable for data management in mobile environments.

Proper searching of user activities has the potential of providing much higher quality and personalized services to individual user on the right place at the right time. We, therefore, propose new methods for complex activity incremental maintenance, online recognition and proactive data management supported user activities. Especially, we devise pre-fetching and pushing techniques with cost-sensitive control to facilitate predictive data provision.

Mobile users can request services through their devices through Information Service and Application Provider (ISAP) from anywhere, at anytime. This business model can be termed as Mobile Commerce (MC) that gives Location-Based Services (LBS) through mobile phones. It provides the cellular network composed of several base stations.

The communication coverage of every base station names a cell as a location area. When a user moves within the mobile group, its location and repair requests are stored during a centralized mobile transaction database.

To achieve a fast reply from the system, data processing is used in many applications, which is one among the foremost promising technologies that want to fulfill a dynamic service request.

2. BASIC CONCEPTS AND PROBLEM DEFINITIONS

The match is made between the mobile access patterns and the user movement database. This database is utilized to pivot raw data into useful knowledge.

Definition 1- [Location]: The generic location may be a collective term for 1 or more interesting locations, and therefore the interesting location may be a subset of generic locations. A generic location can be defined as L $=\{l_1,l_2,l_3,\ldots,l_j\}$, where each element lj represents any random generic location. The interesting location can be defined as the user u_i present at a location l_i longer than the maximum duration.

Definition 2-[Timestamp]: The timestamp Tm is assumed to have the service accessing.

Definition 3-[Services]: S= $\{s_1, s_2,\ldots, s_n\}$ is a set of services requested by the movable users. Each element represents an individual service id. In addition, an optimum time set for each service is requested. If the mobile users use the acquired service longer than the best time, then the service is regarded as an interesting or useful information service.

3. RELATED WORK

It has been known for some time that user behavioral patterns are important for effective mobile computing [3], [4], [5], [6]. The first stage of research represents the acquisition of user behavior. Among the different methods for learning user behaviors, data mining is probably the most broadly used technique [7]. It is well suited for discovering hidden patterns from large volumes of data such as transaction records. The mining of mobility patterns has been the focus of many previous works [1], [8], [9], [10], [11]. The term “activity” is also used for mobility prediction in [16]. The WiQoSM model collectively considers mobility, user-generated traffic, wireless technologies, and the QoS models for user performance and trace generation, which can be used for the simulation of wireless data network protocols [4]. In spite of its importance, mobility patterns tell us from the time when users don’t simply move around without any reason in mind. They are considered as parallel to the location-only activities in our paper. We need to consider about the users, who are projected beyond mobility.

The mining of sequential patterns takes a different loom by discovering frequent sequences such as, in sequence item purchasing behavior [12], general sequential data [13], [14], multidimensional sequential data [15], in order patterns for interval-based data [16], and sequential mobile access patterns [17]. In general, the key focus of mining sequential patterns is the relationship between data items and their time, i.e., it is considered as parallel to the service-only activities in our paper. However, the association between in order patterns and user mobility has not been investigated systematically.

It was only until both spatial and temporal relationships are considered together in mining sequential patterns [2], [18], [19]. Li and Li took a step further to combine movement and access pattern analysis for enhanced services in the cellular systems [3]. However, our concept of activities is still much more general with consideration of the structure and communication between sequential service invocations and mobility patterns.

While user activities’ mining is interesting enough by itself, it is not the primal premise. Our primary goal is to explore the user performance for high-quality information services of devices in movable environments. Activity mining techniques are used to consider user performance patterns. The key point is to build an effective mechanism for data management based on user activities. Among the previous works on mobility or one-by-one pattern mining, very few of them have in depth conversation on how to employ the discovered patterns for data management and information services. Peng and Chen developed data allocation algorithms based on the classification of user moving patterns for achieving local and global optimization in terms of likelihood of local data access [16]. Wu and Chang proposed the use of the user schedules for active replica management, which improves local availability of data [5]. The MoDA scheme proposed by Yamasaki et al. employs the information of user trajectories to determine whether the replica of data are copied and transferred among movable nodes.

4. PROPOSED SYSTEM

The generic location may be a collective term of one or more interesting locations, and therefore the interesting location may be a subset of the generic location. The common patterns of location movement may be considered for geographic relationships between locations or service allocation. The regularity in commission invocation may come from the dependencies between services or proximity of service providers. It’s potentially beneficial to seek out mobility and repair patterns to facilitate network and data management.

In this paper, we introduce three different parameters like location, time stamp, and services to be used. With the help of these parameters and the database, we can easily predict the new user behavior.

By using the database, the new user can get the services in an easy and effective manner. The user can request services at anytime and anywhere. Based on the user request, the services are yet to be provided. Some users can frequently access the services in a particular location.

User static profile is analyzed and their services can be located accordingly fast access. The dynamic behavior of user movement pattern is logged and mined, and suitable service structure could be restructured. The location updating can be done in a proactive manner or in a reactive manner along with locating appropriate services.

Mobile access pattern generation is proposed, which has the capability to generate strong patterns among three different parameters of mobile location, time and service. The proposed approach generates the whole pattern by joining these parameters and finds the uppermost predicted patterns. The proposed approach is very helpful in the mobile service environment to predict the latest services and improve the existing one. By using the user movement behavioral patterns, we have to implement FP-Growth algorithms with the input of generated patterns.

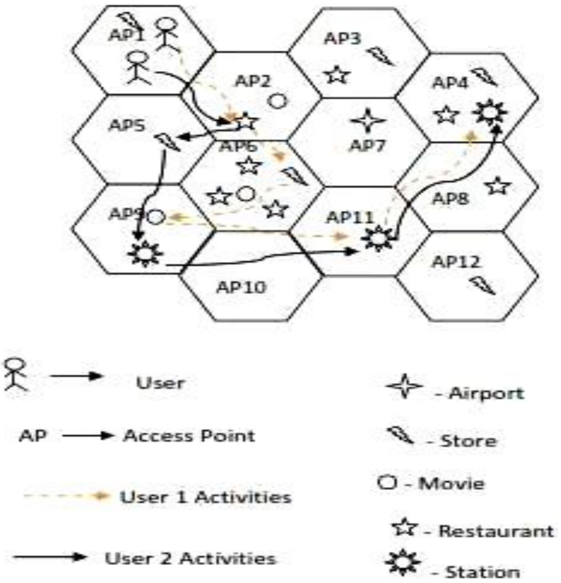


Fig.1. User Behaviors Patterns and the Activity concept

5. SERVICE ARCHITECTURE

Here in this architecture, mobile users are connected through wireless link to the network through the base station. The base stations in the areas are arranged into clusters and are connected to a data server. The data server is responsible for coordinating the base stations in order to provide information services to the mobile users. Both the servers and stations are connected with each other through high-speed static network links.

A mobile user can avail any information at anywhere in any time. Even if the requested data is not available in the neighborhood base station, it is used to be accessed from the remote server or base station. We didn't assume the occurrence of replicas. However, once the client or base station cache has a local copy, no remote access is essential until the next update. In such case, the cached local copy acts as a replica.

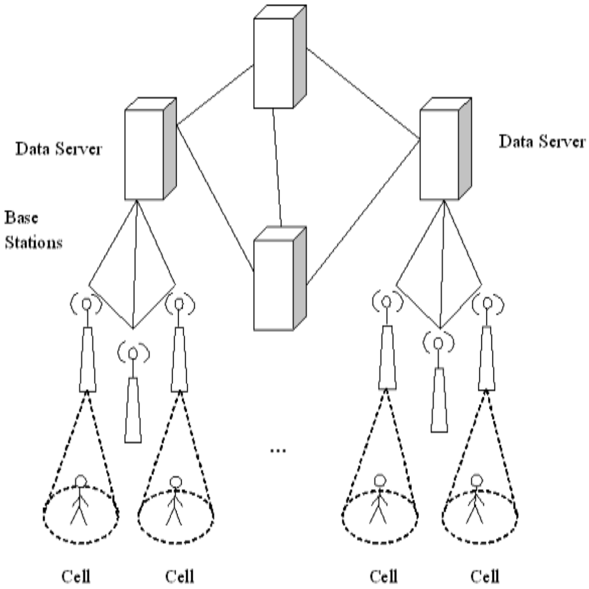


Fig.2. Service Architecture

The data storage available for data management is termed with names, such as, the base station cache, client hoard. Public accessible data like news clips and weather forecasting information are assumed to exist in the data servers, while location-dependent facts like traffic condition and local attractions are kept on the corresponding base stations.

6. GENERATING FREQUENT PATTERNS

Algorithm1 extends the FP-Growth with the special emphasis on activity mining. FP-Growth algorithm does not have any candidate generation like Apriori algorithm. It takes less time to generate the tree construction with this activity set.

Algorithm 1. FP-Growth Algorithm

This algorithm is used to generate frequent patterns without candidate set generation.

Input: Database D (contains timestamp, location, service id)

Output: Generate Frequent Patterns

Procedure:

(1) Calculate minimum support, min_sup = total number of frequencies in the database / total number of behavior items in the database**;**

(2) Find the behavior items which has the equal and above min_sup and remove the infrequent behavior items;

(3) Re-order the frequent items in the database;

(4) Construct tree with the re-order frequent items from the database;

(5) The FP-Tree construction has the following steps:

   a. Create the root of the tree and assume it as null;
   b. Read the transactions one by one from the database and provide counter for each frequent item;
   c. Insert the frequent items in the tree and increment the counter with the help of node;
   d. While reading each transaction, check the frequent items that are already inserted in the tree; if it matches the already inserted frequent items then just increment the counter only;
   e. If it does not match then create the new node for the frequent item;
   f. Repeat the procedure until all the frequent items are to be inserted;

(6) Frequent patterns are generated in the following steps:

   a. Check the frequent pattern prefix from the least min_sup frequent item;
   b. Bottom-up approach is used to search the frequent patterns;
   c. Search the frequent item prefix in the tree until it reaches that there is no frequent item has to be checked.

(7) Finally, the frequent patterns are generated.

Table 1. Mobile User Transaction Database

Transaction id	**Behavior Items**
6	1 K 1 $, 7 T 1 $, 12 Q 1 $, 4 B 2 S11 S19, 25
5	4 B 2 S11 S19, 1 K 1 $, 15 L 4 S10 S19 S15 S1, 17 D 3 S6 S4
8	9 J 2 S15 S1, 14 Y 1 $, 1 K 1 $, 18 V 2 S11 S13, 21 Q 3 S6 S18 S6, 22 Q 4 S11
5	2 A 3 S18 S12 S1, 3 J 2 S2 S12, 7 G 1 $, 16 R
4	6 U 4 S1 S9 S3 S6, 8 O 1 $, 15 L 4 S10 S19 S15 S1, 4 B 2 S11 S19
3	14 Y 1 $, 22 Q 4 S11 S19 S6 S10, 9 J 2 S15 S1
6	8 O 1 $, 26 X 2 S15 S3, 16 R 2 S16 S5, 7 G 1
4	4 B 2 S11 S19, 1 K 1 $, 9 J 2 S15 S1, 2 A 3 S18 S12 S1

Example for FP-Growth algorithm with minimum support = 2

In this example, all behavior items contain timestamp, location, number of services, and services i.e., in the following behavior item 4 B 2 S11 S19, 4 – Timestamp, B - Location, 2– Number of services, and S11 S19 – Services.

Table 2. Extract Frequent Behavior Item Sets from the FP-Tree

Behavior Item	**Frequent Patterns Generated**
26 X 2 S15	{ 22 Q 4 S11 S19 S6 S10, 26 X 2 S15
16 R 2 S16	{ 7 G 1 $, 16 R 2 S16 S5 : 2 }
14 Y 1 $	{ 9 J 2 S15 S1 , 14 Y 1 $:2 }, { 22 Q 4 S11 S19 S6 S10, 14 Y 1 $: 2 }, { 9 J 2 S15 S1 , 22 Q 4 S11 S19 S6 S10, 14 Y 1 $: 2 }
8 O 1 $	{ 4 B 2 S11 S19, 8 O 1 $: 2 }
15 L 4 S10 S19 S15 S1	{ 4 B 2 S11 S19, 15 L 4 S10 S19 S15 S1 : 2 }, {1 K 1 $,15 L 4 S10 S19 S15 S1 : 2 }
9 J 2 S15 S1	{ 1 K 1 $, 9 J 2 S15 S1 : 2 }
1 K 1 $	{ 4 B 2 S11 S19, 1 K 1 $: 3 }

Above Table 1 and Table 2 indicates that the input and output of the FP-Growth Algorithm

7. PREDICTION RATIO CALCULATION

The frequent patterns are stored in a structured manner like tree, graph, etc. In the proposed system, tree and hash is used to store the frequent patterns. With the help of the data structure, the services requested by the mobile user are searched. If it matches then the service is easily satisfied to that particular user. In the same way, the prediction ratio has to be calculated. Prediction ratio has to be achieved by matching the number of services in the structure with the requested services from the user.

8. EXPERIMENTAL RESULTS

The performance analysis of FP-growth algorithm after the thorough experimentation is shown in Fig. 3. From the experimental results, it is observed that FP-growth is an efficient method of the mining frequent patterns in a large Mobile user Database. This method uses a highly compact FP-tree, divide-and-conquer method usually for mining the data. FP-tree is a data structure used to store the compressed, critical information about frequent patterns, aiming to find out a complete set of patterns.

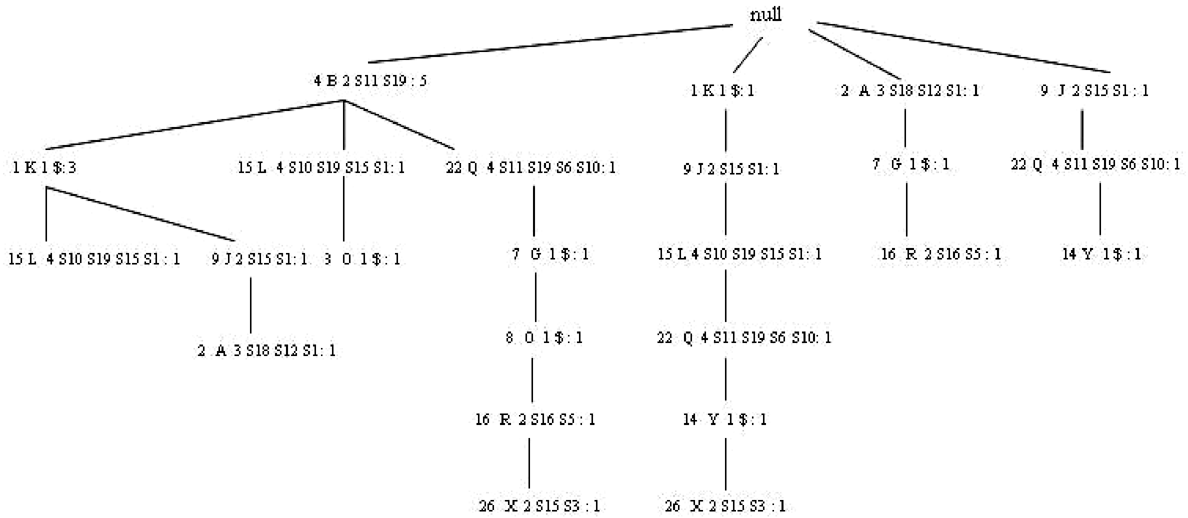


Fig.3. FP-Tree Construction

9. CONCLUSION

In this paper, we propose a data mining method for predicting user movement behavioral patterns by using the parameters user-id, location, timestamp, and services. From this work, it was found that the FP-Growth algorithm works faster to predict the frequent patterns in user movement behavior. With the help of the predicted patterns, the mobile users have the high response for their requested services in less time. By using the tree and hash data structures the predicted frequent patterns are stored to calculate the prediction ratio. It is very useful to find the the exact match of requesting services.

REFERENCES

[1] W. C. Peng and M. S. Chen, “Mining User Moving Patterns for Personal Data Allocation in a Mobile Computing System”, Proceedings of 29th International conference Parallel Processing, pp. 573-580, August 2000.

[2] C. H. Yun and M. S. Chen, “Mining Mobile Sequential Patterns in a Mobile Commerce Environment”, IEEE Trans. Systems, Man, and Cybernetics, Part C, vol. 37, no. 2, pp. 278-295, March 2007.

[3] X. Li and Q. Li, “User Pattern Analysis in Cellular Systems”, Proceedings of Seventh IEEE International conference Mobile Data Management, 2006.

[4] G. Resta and P. Santi, “WiQoSM: An Integrated Qos-Aware Mobility and User Behavior Model for Wireless Data Networks”, IEEE Trans. Mobile Computing, vol. 7, no. 2, pp. 187- 198, February 2008.

[5] S. Y. Wu and Y. t. Chang, “A User-Centered Approach to Active Replica Management in Mobile Environments”, IEEE Transactions Mobile Computing, vol. 5, no.11,pp. 1606-1619, November 2006.

[6] A. Yamasaki, H. Yamaguchi, S. Kusumoto, and T. Higashino, “Mobility-Aware Data Management on Mobile Wireless Networks”, Proceedings of IEEE 65th Vehicular Technology Conference, pp. 679-683, 2007.

[7] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proceedings of International conference Very Large Databases (VLDB), pp. 487-499, 1994.

[8] W. C. P. Jiun-Long Huang and M. S. Chen, “Exploring Group Mobility for Replica Data Allocation in a Mobile Environment”, Proceedings of 12th International conference Information and Knowledge Management, pp. 161-168, 2003.

[9] J. L. Huang and M. S. Chen, “On the Effect of Group Mobility to Data Replication in Ad-Hoc Networks”, IEEE Transactions Mobile Computing, vol. 5, no. 5, pp. 492-507, May 2006.

[10] W. Ma, Y. Fang, and P. Lin, “Mobility Management Strategy Based on User Mobility Patterns in Wireless Networks,” IEEE Trans. Vehicular Technology, vol. 56, no. 1, pp. 322-330, Jan. 2007.

[11] W. C. Peng and M. S. Chen, “Allocation of Shared Data Based on Mobile User Movement”, Proceedings of Third International conference Mobile Data Management, pp. 105-112, 2002.

[12] M. Sricharan, V. Vaidehi, and P. Arun, “An Activity Based Mobility Prediction Strategy for Next Generation Wireless Networks”, Proceedings of IFIP International conference Wireless and Optical Communication Networks, 2006.

[13] R. Agrawal and R. Srikant, “Mining Sequential Patterns”, Proceedings of 11th International Conference Data Engineering, pp. 3-14, 1995.

[14] K. Gouda, M. Hassaan, and M. J. Zaki, “Prism: A Primal- Encoding Approach for Frequent Sequence Mining”, Proceedings of Seventh IEEE International conference Data Mining, pp. 487-492, 2007.

[15] J. Pei, J. Han, B. Mortazavi - Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M. C. Hsu, “Mining Sequential Patterns by Pattern Growth: The Prefixspan Approach”, IEEE Transactions Knowledge and Data Engineering, vol. 16, no. 11, pp. 1424-1440, November 2004.

[16] C. C. Yu and Y. L. Chen, “Mining Sequential Patterns from Multidimensional Sequence Data”, IEEE Transactions Knowledge and Data Engineering, vol. 17, no. 1, pp. 136-140, Jan. 2005.

[17] S. Y. Wu and Y. L. Chen, “Mining Non-ambiguous Temporal Patterns for Interval-Based Events”, IEEE Transactions Knowledge and Data Engineering, vol. 19, no. 6, pp. 742-758, June 2007.

[18] V. S. Tseng and K. W. Lin, “Mining Sequential Mobile Access Patterns Efficiently in Mobile Web Systems”, Proceedings of 19th International conference Advanced Information Networking and Applications, vol. 2, pp. 762- 767, March 2005.

Credit Card Fraud Detection using Anomaly Detection

Amanjot Kaur Shemar[1], Brahmaleen Kaur Sidhu[2]

[1]Scholar, Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India.
[2]Assistant Professor, Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India.
Email: [1]amanjotkaur184@yahoo.com, [2]brahmaleen.ce@pbi.ac.in

Abstract- Fraud is a malicious activity that causes financial loss. Fraud causes by Credit Card have costs consumer as well as banks. Nowadays fraudsters are implementing diverse methods to commit frauds. So there should be a system like fraud detection which has the capability to detect the fraud activities before occurring and also in an accurate way. This paper discusses the utilization of machine learning methods in credit card fraud detection. Anomaly detection is a decisive obstacle that has been examined in varied research zones and various application domains. For a particular domain, different anomaly detection methodology has been specifically developed. This paper will talk about anomalies and its various types. Different aspects and challenges are considered in anomaly detection. Training data and different techniques are used in anomaly detection to solve a particular problem. Moreover it will discuss how the different techniques of anomaly detection are applied to solve the fraudulent activities in credit card data and what factors should be consider to apply to get full accuracy in the detection of fraudulent tasks in credit card data analysis.

Keywords - Anomaly Detection, Challenges and Techniques, Credit Card Fraud Detection.

1. INTRODUCTION

Nowadays utilization of credit card to pay bills and for online transaction is on its peak. It has changed the way of payment. Money is transferred in the form of transaction. But, due to increase in the users of credit card, the fraud cases have also on arisen. Frauds in Credit card can be done in multiple ways that are as:

- **Application Fraud:** When fraudster accesses the sensible details of user like username and password and creates a fake account and gains control of application system. So fraudsters steal underpin documents in order to support fraudulent application. It mostly happens in relation to identify theft.

- **Manual Credit Card Imprints:** In this fraudster make use of the magnetic strip of the card to get the information.

- **Original Card Not Present:** When card is used without its actual physical possession because the fraudster has knowledge about its expiry date and account number.

- **Counterfeit Card Fraud:** In this dummy magnetic swipe card is made that hold all features of genuine card. This is done by the method of skimming. The counterfeit card is fully functional and used in transactions.

- **Mail Non Receipt Fraud:** Whenever a customer applies for a new card, all the procedural formalities take some time. So, Fraudster intercepts in the middle of delivery, they perform some by changing the name of user with their own name and make the purchases. So it is also called Never Received Issue Fraud.

- **Off track and Stolen Card Fraud:** In this case, the holder of card misplaced their card. Fraudster uses that card to make payments. It is not easy to do because pin number is needed but fraudster can use it to make online transactions.

- **False Merchant Sites:** It is same as phishing attack. Fraudsters create a website that is consisting of fake web pages, but it seems to be genuine. These web pages contains attractive designs and also provides some offers like huge discounts, buy one get one free etc. So customers often attracted to these offers. Once a transaction takes place, information related to all transactions is collected and fraudster uses it further to carry out fraudster exchanges.

Detecting a fraud is a tricky computational piece of work because it is very difficult to choose parameters. The success of fraud detection depends upon cluster and classification of parameters. A transaction is fraud or genuine is classified by the existing systems based on various patterns. So application of different techniques and algorithms of machine learning to solve these problems depends upon various factors. There is a technique of machine learning that can easily detect this types of frauds is well known as Anomaly Detection.

2. ANOMALY DETECTION

Anomalies are those values and patterns that do not belong to the normal behavior.

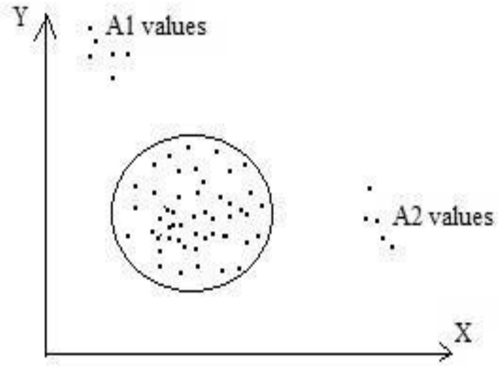


Fig.1. Anomaly Detection

In Figure 1. shown if A1 and A2 values that are very far away from the values of group or do not belong to any group. So say that A1 and A2 are Anomalies. So to find these patterns, we use a technique that technique is called Anomaly Detection.

So anomaly detection is the procedure or a method that identifies these items or patterns that do not confirm to other items present in the Dataset. These irrelevant patterns are known as anomalies, exceptions, outliers, observations, depending upon the problem. So anomaly detection is used in various applications to find these problems.

3. DIFFERENT ASPECTS OF ANOMALY DETECTION

3.1 Description of Input Data

The crucial characteristics of anomaly detection are the description of input data. Data may be of various instances like values, observations, records, points, vectors, objects, patterns etc. Each data instances are expressed applying a set of attributes like variable, characteristics, feature and dimension etc. These may have the values like binary, categories or continuous. Data instances may be of two types. First one is univariate data instance in which only one attribute is used. Second is multivariate in which multiple attributes might be of fusion of different data types or might be of same type. The identity of attributes determines that which anomaly detection technique is to apply. The data instance may be of different types like in sequence data, in spatial data, in graph data. Data in sequence is cramped. It means data is in sequence. Neighboring instance related with data, makes the Spatial data. In Sequence data type data is linearly ordered. For instance, in time series, data is in sequence. In spatial data, data is associated to its neighboring instances, such as traffic data and ecological data. Graph data in which data instance uses vertices and edge. Vertex represents the data instances and edges are used to connect with other vertices or with data instances.

3.2 Data Labels

Labels in the dataset provide normal or anomalous details of instance. So it is often extremely expensive and difficult to obtain labeled data that is accurate and represent all types of behaviors. No specified method is used for labeling, so Manual ways are used to label the data instances. This need an expert so it requires substantial efforts to obtain data set that is properly labeled. Moreover, an Anomalous behavior is frequently dynamic in nature according to various conditions. Like, latest types of anomalies might arise which don’t include labeled training data. So there are three categories of dataset used in anomaly detection.

A. Supervised Data - In this there is availableness of training data set that has labeled instance for normal as well as for anomalous class. There are two issues in this. First is- In training data, both normal and anomalous instance exists, but very fewer comparisons are made between normal and anomalous instance. It includes subjects that make an appearance due to imbalanced class distributions. Second is - datasets that is accurate and representatively labeled for the anomaly class is usually demanding. So to obtain a properly labeled dataset, various numbers of techniques are applied that makes artificial anomalies in the dataset.

B. Semi Supervised Data - In this both labeled and unlabeled data is used. Only normal class data sets in the training data have labeled instances. Anomaly classes do not require labels. So, they are more applied.

C. Unsupervised Data - This requires no training labeled data, so that’s why it is the most widely applicable. It uses implicit assumptions for normal instance in the test data...

3.3 Types of Anomaly

Various categories of Anomalies are point, contextual and collective.

A. Point Anomaly - if individual data instance is considered different from rest of data. For example, in Credit card amount spend greater than normal range.

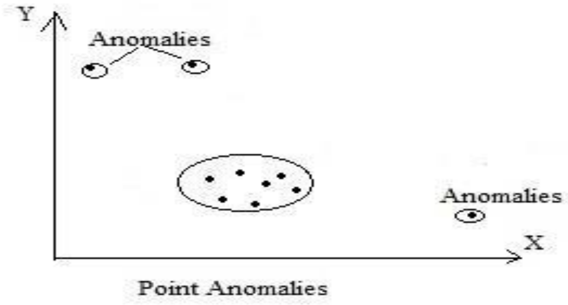


Fig. 2. Different Points Anomalies

In figure 2 different point's anomalies are shown.

B. Contextual Anomaly – if data instances are anomalous in a specified context.

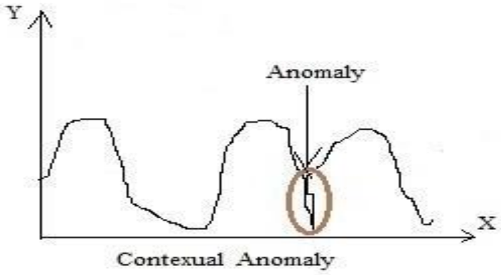


Fig. 3. Circled part is contextual anomaly

In figure 3: circled part is contextual anomaly. Data instances are defined by two attributes: First is Contextual Attributes, which provides the description about the context of instances. For instance-in spatial data set, contextual attributes are longitude and latitude of location. Second is Behavioral Attributes that give details about Non-Contextual properties of the instance. For instance- in spatial data set average rainfall of the entire world is behavioral Attribute.

C. Collective Anomaly- If the collective of connected data is divergent from the entire data set. If individual data instance is not anomalous, but their occurrence together is Anomalous.

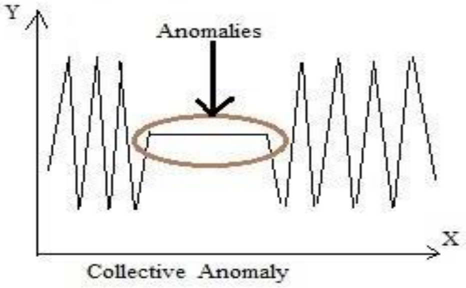


Fig. 4. Oval circle represent the collective anomaly

Figure 4 shows oval circle that represents the collective anomaly.

3.4 Consequence of Anomaly Detection

Output of particular problem in anomaly detection may be represented in different formats that may be scores, ranges or labels etc. But most of result is represented in two formats. First is Score. In test data, each instance is assigned a score and depending on degree, finds out that anomalous instance. So get a ranked list of anomalies in this. So use cut off threshold to select anomalies. Second are labels that labeling is used for normal and anomalies and assign to each data instance. It uses a domain specific threshold to choose anomalies. Approaches that use binary labels to examine instance may assign 0 as anomalies and 1 as normal and may be vice versa.

4. TECHNIQUES OF ANOMALY DETECTION

There are various techniques that are used in anomaly detection. Each particular problem uses different anomaly detection technique according to various factors like accuracy or fast result etc. Those techniques are:

A. Artificial Neural Network

It is the process that consists of interconnected nodes. Its working is same as a human brain. A weight value is assigned to each node. Inputs are provided to the network and output is generated by multiplying these input values with the weights. This technique do not need to be reprogrammed, because it can learn from past. Extraction rules from past helps in making predictions about future situations. Its accuracy is high and also high portability and high speed. But, it takes peak processing time for massive normal networks and also sensitive to data format. It needs excessive training.

B. Artificial Immune System

It is very complex system that is composed of a network of specialized tissues, organs, cells or chemicals molecules. Interrelation between the element and they act in a highly coordinate and specific manner that identify cells causing disease and eliminate them. It is easy to integrate with other system and also handle noise and fault tolerance. But, it is not very much expensive. It need peak training time and is poor in handling mislaid data.

C. Genetic Algorithms

In this technique stranger member of the population have longer chances to live and replicate offspring. We use fitness factor to select population. It works well with noising data. So it is easy to integrate with other system. We can combine into other technique to raise the performance. Easy to build and inexpensive and fast in detection. It is quite difficult to understand. Extensive knowledge about tool is required to set up and operate.

D. Hidden Markov Model

It is the technique that is applied to model that is much more complicated and to chiastic process. So it is double embedded stochastic process. It is fast in detection. It is highly expensive and has low accuracy. But, it is not applicable to data sets that are of larger size.

E. Support Vector Machine

It is supervised learning that is used in classification and regression for analyzing and recognizing patterns. Finding an optimal Hyper Plane is the main function of this technique. It separates intervals of two given classes linearly. Hyper plane was supposed to be located between same marginal instances called support vectors. It delivers a unique solution by selecting a suitable generalization grade. This technique is robust if training sample has same bias. It is poor in processing large database. It has low speed of database and medium accuracy.

F. Bayesian Network

It is expressive model that identifies conditional dependencies amid random variables. It is used to find unknown probabilities in the present of uncertainty. It is worthwhile when the basic knowledge is previously known but increasing data is uncertain or partially available. It has high processing & speed of detection is also very high.

G. Fuzzy Logic

It is based on fuzzy rules. It defines fuzzy sets in which uncertainty of input & output variables are addressed and linguistics variables are used to define values. Example - small, medium and large. It is very fast in detection. It has good accuracy. It is quite expensive and has very low speed in detection.

H. Expert Systems

Information is used by human expert to generate rules and that rules are kept in a rule or order based system as IF-THEIR rules. It is effortless to develop and construct the system and also simple to control. It has high degree of accuracy and its performance is better. It is very poor handling missing information or data values that are unexpected. Its processing is low in processing different data type. It is poor in building and integrating.

5. VARIOUS CHALLENGES IN ANOMALY DETECTION

- Describing a Normal Region- Region that create common behavior is not that any easy. The dividing line between the normal anomalous behaviors is often not defined precise. To the normal observations may be Anomalous and vice versa.
- When we use the Anomalies that are the fallout of malicious actions and we use them to define Normal observations that always generate wrong results and task to define Normal behavior becomes extremely arduous.
- In some domains, effectiveness and representation of common behavior may not be sufficiently done in the future.
- Availability of labeled data that is used for training and validation of models is usually a major subject.
- Noise in the data that is similar to anomalies is difficult to distinguish and remove.
- Different application domains use different AD techniques like In Medical domain-: fluctuations in the body temperature. In Market domain-: fluctuations in the value of a stock.

So applying one technique to another is not straight forward.

6. APPLICATIONS OF ANOMALY DETECTION

Anomaly Detection is used in various fields and to solve various types of problem. That is:

- **Intrusion Detection**: it is the technique uses that mention detection of malicious task, in a computer related system. An intrusion is dissimilar from common behavior of a system. So we apply anomaly detection to solve these problems.
- **Fraud Detection:** It is the technique that mentions detection of criminal activities that takes place in commercial organization that may be banks; credits card companies, stock market, cell phone etc. The users who want to perform malicious activities act as an actual customer or may be the person in the organization. The fraud happens when users consume the resource in an unauthorized way.
- **Medical and Public Health**: It functions with patient records. Anomalies can happen due to several causes as like abnormal patient condition or errors such as instrumentations errors recording errors.
- **Industrial Damage Detection:** In Industry resources are used. But due to continuous usage and simple wear and tear of these resources, the industries have to suffer a lot of damage. This should be detected untimely to present further rises and losses. So apply different techniques to detect such damages.
- **Image Processing:** It deals with image. In this, two activities are done. First is, changes are done in image to get a new type of image and second is to find abnormal regions in static image. The anomalies in this may be in

point or in regions like point and contextual anomalies. So we use detection technique to solve this problem.

- **Text data:** There is a gathering of documents or news articles. So in this sphere, detect some novel topics or events or new stories in that gathering that can cause anomalies. So we apply different techniques to solve and find these Anomalies.
- **Sensor Network**: In this domain, network collects the data. Anomalies in the collected data can have two meaning. First is that one or more sensors are defective. Second is that they are detecting events that are engaging for analysis. So Anomaly Detection technique in sensor network can catch both types of detects fault and intrusion detection.

7. RELATED WORK IN CREDIT CARD FRAUD DETECTION

In 2011, Raghavendra Patidar and Lokesh Sharma [4] have put forward a hybrid of Artificial Neural Network and Genetic Algorithm. Neural network is used to categorize the transactions and genetic algorithm optimizes the solution. Aleskerov *et al.* [5] established a neural network formed data mining system for the detection of credit card fraud. The system used three layers auto associative architectures. For training and testing the system they used a set of synthesized data. This shows extremely successful fraud detection rates. Krenker *et al.* [6] proposed a model on bidirectional neural networks for real time fraud detection based. A large dataset of cell phone transactions are provided by credit card companies. In terms of false positive rate the system outperforms the rule based algorithms. An Artificial Immune System based model for online credit card fraud detection was proposed by Brabazon *et.al.* [7] In this, three algorithms were carried out and a logistic regression model was used for standardization of their performance. These three algorithms were the unmodified negative selection Algorithm, the modified negative selection algorithm and the Clonal selection algorithm. Distance Value Metric was proposed for enumerating distance between records. This metrics is built on the probability of data occurrence in the training set. K. Rama Kalyani *et al.* [8] presented a model that is based on principles of genetic algorithm for credit card fraud detection. This approach developed a synthesizing algorithm that works in two phases. In first phase, it generates the test data and in second phase, applies that proposed algorithms to detect fraudulent transaction. A genetic programming based fuzzy system was developed by Bentley et al. [9] for the extraction of rules for classification of data examined on actual home insurance claims and credit card transactions. Bhusari et al. [11] utilized HMM for recognizing credit card frauds that has short false alarm. The present system was also applicable for processing huge number of transactions. Ghosh and Reilly [12] designed a model making use of Support vector machines and admired neural networks. In this, a three layer feed-forward RBF neural network is developed that detect fraudulent credit card transactions. Only two passes required to produce a fraud score in every two hours. Tung-shou Chen *et al*. [13] generated a binary support vector system (BSVS), in which genetic algorithms (GA) are used to choose support vectors. To acquire a high true negative rate, first self-organizing map (SOM) was first applied and BSVS was then applied to better train the data according their distribution. In 2015, J. Esmaily and R. Moradinezhad [14] in their paper presented a hybrid of artificial neural network and decision tree. They implemented a two-phase approach. In first phase a new dataset is generated from the classification fallout of Decision tree and Multilayer perceptron. In second phase, classification of data is done be feeding this datasets into multilayer perceptron. This model has very low false detection rate that promises reliability. Ezawa and Norton developed a four-stage Bayesian network [15]. They explained that other popular manners like regression, K-nearest neighbor and neural networks take too extensive time to be applicable in their data. In 2015, Tanmay Kumar and Suvasini Panigrahi[16] in their paper implemented a hybrid methodology using fuzzy clustering and neural network. It works in two phases. In phase one; they developed a c-means clustering algorithm for generation of suspicious score of the transaction. In second phase, suspicious transaction is feed into neural network to find out whether it was really fraudulent or not. Sam Maes[17] proposed Bayesian Networks and Artificial Neural Network, two machine learning techniques for detecting frauds in credit card. He also talked about that how Bayesian networks after a small training gave better consequences and the use of ANN to enhance speed. Thuraya Razooqi [18] made a system of fraud detection by applying Fuzzy Logic and Neural Network. They determine that accuracy of Artificial Neural Network was 33% more than fuzzy logic. For decision making, the data that exists in the system was used and a membership attribute was given to each data using fuzzy logic. Neural Network was used for the validation of results.

8. CHALLENGES IN CREDIT CARD FRAUD DETECTION

Solving the problem of anomaly detection, the outlier class of modeling can be senseless and unproductive. This needs to pay attention to the structure of the normal data and its distribution. Great fraud detection system makes it possible to identify fraud precisely. It should be capable of detecting fraud in the transit process. This should not term any authentic transaction as fraudulent and vice versa. Numbers of wrong classifications should be minimum. Unavailability of the complete data for analysis because neither banks nor customers reveal the details and information. Absence of fine and efficient evaluation pattern that are used to check the accuracy of the system. Need a technology that should be able to detect fraudulent transaction when it is occurring. It reduces cost. All the techniques never give the same result in all environments. So each deals with different properties. No one technique gives 100% of accuracy. So an integration of multiple algorithms can raise the accuracy of the final result.

REFERENCES

[1] Jain, Y., Namrata Tiwari, S., & Jain, S. A, “Comparative Analysis of Various Credit Card Fraud Detection Techniques”, International Journal of Recent Technology and Engineering (IJRTE), vol.7, pp. 402-407, 2019.

[2] Sharmila, V. C., Kumar, K., Sundaram, R., Samyuktha, D., & Harish,R, “Credit Card Fraud Detection Using Anomaly Techniques”, 1st International Conference on Innovations in Information and Communication Technology (ICIICT), pp.1-6. IEEE, 2019.

[3] Samaneh Sorournejad, & Zojaji, Zahra & Ebrahimi Atani, Reza & Monadjemi, Amir, “A Survey of Credit Card Fraud Detection Techniques”, Data and Technique Oriented Perspective, 2016.

[4] Raghavendra Patidar, Lokesh Sharma, “Credit Card Fraud Detection Using Neural Network”, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231-2307, vol 1, 2011.

[5] E. Aleskerov, B. Freisleben, B. Rao, “CARDWATCH: A Neural Network-Based Database Mining System for Credit Card Fraud Detection”, The International Conference on Computational Intelligence for Financial Engineering, pp. 220-226,1997.

[6] A. Krenker, M. Volk, U. Sedlar, J. Bester, A. Kosh, “Bidirectional Artificial Neural Networks for Mobile-Phone Fraud Detection”, Journal of Artificial Neural Networks, vol. 31, No. 1, pp. 92-98, 2009.

[7] A. Brabazon, et.al., “Identifying Online Credit Card Fraud using Artificial Immune Systems”, IEEE Congress on Evolutionary Computation (CEC), Spain, 2011.\

[8] K. Rama Kalyani, D. Uma Devi, “Fraud Detection of Credit Card Payment System by Genetic Algorithm”, International Journal of Scientific & Engineering Research, vol 3, No.7, 2012.

[9] Bentley, P., Kim, J., Jung., G. & Choi, J., “Fuzzy Darwinian Detection of Credit Card Fraud”, Proceedings of 14th Annual Fall Symposium of the Korean Information Processing Society, 2000.

[10] V. Bhusari, S. Patil, “Application of Hidden Markov Model in Credit Card Fraud Detection”, International Journal of Distributed and Parallel Systems (IJDPS), vol. 2, No. 6, November 2011.

[11] Cortes, C. & Vapnik, V, “Support-Vector Networks, Machine Learning”, vol. 20, pp. 273-297, 1995.

[12] Ghosh, S. & Reilly, D, “Credit Card Fraud Detection with a Neural Network”, Proceedings of 27th Hawaii International Conference on Systems Science, vol. 3, pp. 621-630. 2004.

[13] Tung-shou Chen, Chih-Chianglin, “A New Binary Support Vector System for Increasing Detection Rate of Credit Card Fraud”, International Journal of Pattern Recognition and Artificial Intelligence, vol. 20, No. 2 pp. 227–239, 2006.

[14] R. M. Jamail Esmaily, “Intrusion Detection System Based on Multilayer Perceptron Neural Networks and Decision Tree”, International Conference on Information And Knowledge Technology, 2015.

[15] K. Ezawa, S. Norton, “Constructing Bayesian Networks to Predict Uncollectible Telecommunications Accounts”, IEEE Expert, vol 11, No. 5, pp. 45–51, 1996.

[16] S. P. Tanmay Kumar Behera, “Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering and Neural Network”, International Conference on Advances in Computing and Communication Engineering, 2015.

[17] K. T. B. V. Sam Maes, “Credit Cards Fraud Detection using Bayesian and Neural Networks”, vol 4, No.7, August 2002.

[18] P. K. D. K. R. D. A. A. Thuraya Razoogi, “Credit Card Fraud Detection using Fuzzy Logic and Neural Networks”, Society for Modeling and Simulation International(SCS), 2016.

[19] Varun Chandola, Arindam Banerjee, and Vipin Kumar, “Anomaly Detection: A Survey”, ACM Computing Surveys, vol 41

[20] , No.3, Article 15, 2009.

Transformed Puzzle for Preventing Selfish Mining: A Non-Viable Way to Defend Zero Block Algorithm

Anjaneyulu Endurthi[1], Boya Dastagiri[2], Shaik Janipasha[3], Durgam Santhosh[4], Akhil Khare[5]

[1]Assistant Professor, Department of Computer Science and Engineering,
Rajiv Gandhi University of Knowledge Technologies, Basar, Nirmal, Telangana, India.
[2,3,4] Scholar, Department of Computer Science and Engineering, Rajiv Gandhi University of Knowledge Technologies, Basar, Nirmal, Telangana, India.
[5] Professor, Department of Computer Science and Engineering, MVSR Engineering College, Hyderabad, India.

Email: [1]anjaneyuluendurthi@gmail.com, [2]dastagiri1471@gmail.com, [3]janipasha1121@gmail.com, [4]durgamsanthosh141@gmail.com, [5]khare_cse@mvsrec.edu.in

Abstract- Blockchain has become a one-stop solution to achieve data integrity. Blockchain is the series of blocks that contains the transactions or data of any kind. All of these blocks (data) is guarded by using cryptographic principles. The block will be added once the cryptographic puzzle is solved. The main aim of miners (people who solve the cryptographic puzzle) behind adding a new block is to achieve reward. But to achieve more rewards, miners keep the generated block private and these miners will generate more blocks based on the block which was generated previously by the corresponding miners. These miners will publish more blocks at a time to get a big reward. It is called as selfish mining. Moreover, a solution is being proposed known as the zero block algorithm. This paper presents a modified zero block algorithm in such a way that it can resist selfish mining. The original zero block algorithm generates dummy zero blocks whereas the modified zero block algorithm will not generate any dummy zero blocks and is more efficient than the zero block algorithm.

Keywords-- Puzzle , Target Hash , Selfish Mining , Maximum Acceptance Time.

1. INTRODUCTION

Blockchain has brought revolution across many sectors where there is some kind of data is to be stored. Blockchain is well known for its irreversible transactions and also it maintains transparency, security by making use of distributed ledger. Blockchain has many applications that are used to transform society such as Internet-of-Things[1], Banking and Finance, supply chain management etc. Along with its applications, there are many new-age complex attacks that cause unrecoverable damage to the Blockchain. The well-known attacks are categorised as attacks based on peer-to-peer networks, attacks based on Consensus & Ledgers, attacks based on Smart Contracts, Wallet-based attacks [2],[3],[4]. One of the attacks that have created more destruction to the block chain network [5] is selfish mining which comes under Consensus Mechanism and Mining-based Attack. By selfish mining, the miners can increase their revenue by dynamically releasing the blocks into the network. According to the block chain protocol once the miner has found the block then they have to release the block right away so that other miners will start working the newly generated block so as to create one more block on top of it. But selfish miners, instead of releasing the generated block, they may maintain their private branch of blocks. These selfish miners, after generating a private chain they release all the blocks at once to get a bigger reward. As per the Blockchain protocol, the network accepts the longer chain of blocks. So, once the miner reaches certain condition they immediately broadcast all the blocks into the network. This situation affects wasting all the resources and the computational power of the other miners in the network. In this case honest miners lose their resources without getting any benefit.

This paper presents a non-viable way to defend selfish mining by a modified zero block algorithm. The organization of the paper is as follows: section 2 discusses about existing techniques to overcome/reduce selfish mining. Proposed algorithm is introduced in section 3 along with the flexibility of proposed algorithm and its advantages over other algorithms and conclusion is provided in section 4.

2. LITERATURE REVIEW

Very few strategies have been implemented to reduce the percentage of selfish mining. Each strategy tried to reduce selfish mining considering some parameters. Some strategies degrade selfish mining based upon the computational power, by using timestamps and acceptance time [6]. Eyal and Sirer [7] proposed interesting facts that encourage selfish mining. According to this, Even though the selfish miner has not won many of the block races still they can obtain more reward if the selfish miner has more computational power(Around 33%). Another possibility of selfish mining is that, if the selfish miner won in the block races more than the honest miners then the selfish miners can be allowed to receive more than their fair share. In 2014 Ethan Heilman [8] introduced a new strategy called freshness preferred. This approach reduces the percentage of selfish mining by making use of unforgivable timestamps. As per this strategy, it accepts the block which has a very recent valid timestamp. Siamak Solat, Maria Potop-Butucaru [6,10] also implemented an algorithm to have command over the selfish mining.

This algorithm makes use of one time constraint called Maximum Acceptable Time (MAT) to accept the generated block. Once this time is exceeded then the dummy block will be added to the Blockchain and then the honest miners will not accept broadcasted blocks. The disadvantage of this algorithm is that, due to too many dummy blocks, the efficiency of Blockchain is reduced. A new approach is coined in this paper to remove the flaw present in their approach.

3. PROPOSED ALGORITHM

This section discusses about modified zero block algorithm and how it works. The main modification is as follows. Every block must be generated within the MAT. If none of the miners are able to generate a block, then the blockchain network will change the target hash. The new target hash is different from the previous target hash, containing different nonce. Thus calling it as “**Transformed puzzle**”. Then except the selfish miners, all the other miners get to know that the target hash has been changed by calculating the difficulty again. Then the honest miners will mine the block accordingly, but the selfish miners may still continue to work on the old target hash as they might know that the target hash is changed.

The notations used in the algorithm, like Target hash (T), Information Propagation Time (IPT) [9], Block Generation Time (BGT), Difficulty, Maximum Acceptance Time (MAT) are described in [6] along with their formulae.

Modified Zero Block Algorithms:

```
index = 0
mat[index] = 0
bgat = block generation average time
localChain = Genesis Block
NewBlockFlag = False
nonce = 0
HashPrB = 0
T = target hash
newChain = Null
seconds_counter =0
ansPoW = 0
while (True) do
    if (NewBlockFlag == False) AND (mat[index] != 0) then
        T = new Target Hash
    endif
    index = index + 1
    mat[index] = mat[index-1] + ( bgat + ipt)
    while (seconds_counter <= mat[index]) do
        newChain = checkNewBlock()
        if (newChain != Null) then
            HashPrB = SHF(getHead(localChain))
            if (FHF(HashPrB, newChain.ansPoW) <= T) then
                localChain = newChain
                newChain = Null
                NewBlockFlag = True
                break
            end if
        end if
        if (seconds_counter < bgat) then
            if (NewBlockFlag == False) then
                HashPrB = SHF(getHead(localChain))
                if (FHF(HashPrB, nonce) <= T) then
                    ansPoW = nonce
                    localChain = join new block mined
                    BroadcastBlock(localChain, ansPoW)
                    NewBlockFlag = True
                    nonce = 0
                    break
                end if
                nonce = choose random nonce
            end if
        end if
    end while
end while
```

In the above algorithm, Seconds counter will be activated once the miners start finding the hash for the block. It will be on until the MAT is completed.

3.1 Flexibility of Modified Algorithm

Original version of the algorithm has the flaw of facing too many sequences of dummy zero blocks, we can avoid it by changing the puzzle once the Maximum Acceptance Time is over instead of adding the dummy zero block.

In our algorithm, we are calculating Maximum Acceptable Time (MAT). If no new block is received by the miner within MAT then, instead of adding a dummy zero block to the local chain the network will generate a new Target Hash (T). All honest miners will solve the proof-of-work based on this new Target Hash. The honest miner will never accept the selfish miner’s block because selfish miner will not able to generate correct nonce by using proof of work, as the network has generated a new Target Hash.

Below mentioned are the all the possible scenarios, in which we can defend selfish mining by using Modified Zero Block Algorithm.

Scenario 1: In this case, neither honest nor selfish miner discovers nonce for the puzzle during the MAT. So, the network will change the puzzle (new Target Hash). The honest miners know that the puzzle has been changed and they try to find the target hash of the new puzzle (puzzle2). The news about the changing of the new puzzle is unknown to the selfish miners and they keep working on the old puzzle target hash (puzzle1). So, even if they find the nonce for the old puzzle it will never be accepted by the network.

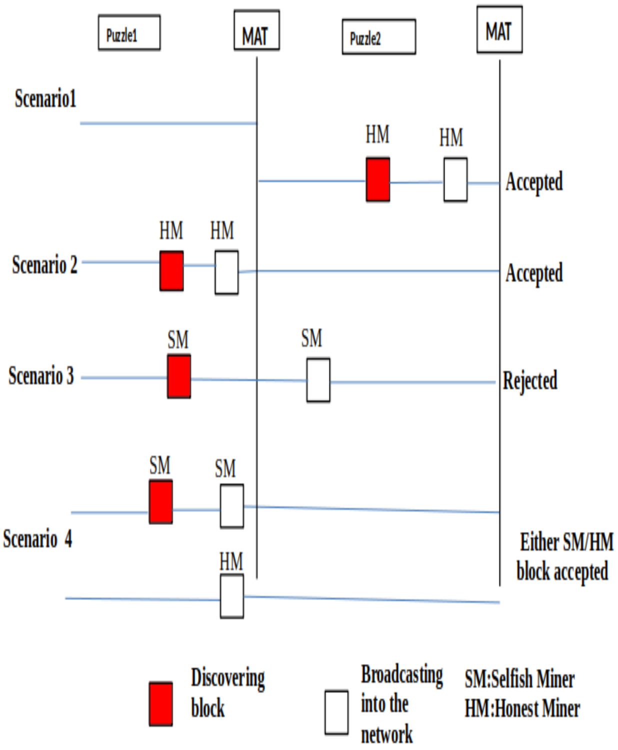


Fig.1. Different scenarios in Modified Zero block algorithm

Scenario 2: In this case, an honest miner discovers a nonce for the puzzle within the MAT then he/she broadcasts the block to the network and starts mining the next block. The block has been discovered by the honest miners within the MAT, so there is no need to change the target hash of the network.

Scenario 3: In this case, the selfish miner discovers a nonce for the puzzle within the MAT then he/she keeps the block private and will not broadcast block. Since within MAT no block is received, network will change the target hash of the puzzle (puzzle2) and the honest miners starts working on the new puzzle (puzzle2). Once the MAT is over the selfish miner broadcast the block but the block will not be accepted because all the miners are working on the new puzzle.

Scenario 4: In this case, Both selfish and honest miners discovers a new block within the MAT and immediately broadcasts the block. So, at this time block race happens and either the honest miner's or the selfish miner's block gets accepted.

3.2 Advantage of Modified Algorithm over Existing algorithm

The advantage of proposed algorithm over existing algorithm is that the existing algorithm produces large number of zero blocks when the miners are not able to solve the puzzle within MAT, which results in accumulation of more number of zero block thus results in memory wastage. Whereas proposed algorithm will not results in creation of unnecessary dummy zero blocks.

4. CONCLUSION AND FUTURE DIRECTIONS

This paper presents a modified Zero block algorithm which generates a new target hash whenever a new block is not generated within MAT. By this approach, we can overcome the selfish mining without the need to create dummy zero blocks. Thus the modified algorithm defends a major problem of selfish mining. Selfish miners are creating new approaches by to get more rewards at the cost of honest miners. More efficient and stringent mechanisms should be proposed to defend selfish miners.

REFERENCES

[1] Malak Alamri, NZ Jhanjhi, Mamoona Humayun, "Blockchain for Internet of Things (IoT)Research Issues Challenges & Future Directions: A Review", IJCSNS International Journal of Computer Science and Network Security, Vol. 19, No. 5, 2019.

[2] Aruba Marketing, Blockchain and New Age Security Attacks: https://blogs.arubanetworks.com/solutions/10-blockchain-and-new-age-security-attacks-you-should-know/, 2019.

[3] S. Sayeed, H. Marco-Gisbert and T. Caira, "Smart Contract: Attacks and Protections", IEEE Access, vol. 8, pp. 24416-24427, 2020.

[4] S. Sharkey and H. Tewari, "Alt-PoW: An Alternative Proof-of-Work Mechanism", IEEE International Conference on Decentralized Applications and Infrastructures (DAPPCON), Newark, CA, USA, pp. 11-18, 2019.

[5] Anjaneyulu Endurthi, Akhil Khare, “Certificate management system using blockchain”, 11th International Conference on Soft Computing and Pattern Recognition, 2020.

[6] Siamak Solat, Maria Potop-Butucaru, “Zeroblock: Preventing selfish mining in bitcoin”, [TechnicalReport] Sorbonne Universites, UPMC University of Paris, 2016.

[7] Eyal I., Sirer E.G. Majority Is Not Enough: Bitcoin Mining Is Vulnerable. In: Christin N., Safavi-Naini R. (eds) Financial Cryptography and Data Security. FC 2014. Lecture Notes in Computer Science, vol 8437. Springer, Berlin, Heidelberg, 2014.

[8] Heilman, Ethan., “One Weird Trick to Stop Selfish Miners: Fresh Bitcoins”, A Solution for the Honest Miner (Poster Abstract). vol. 8438. 161-162., 2014.

[9] C. Decker and R. Wattenhofer, “Information propagation in the Bitcoin network”, IEEE P2P 2013 Proceedings, Trento, pp. 1-10, 2013.

[10] Siamak Solat, Maria Potop-Butucaru, “Brief Announcement: ZeroBlock: Timestamp Free Prevention of Block-Withholding Attack in Bitcoin”, Springer Science and Business Media LLC, Chapter 25, 2017.

Automatic Vehicle Speed Control: An Assistance forAccident Avoidance

Saurabh Baviskar[1], Pushkar Kadhane[2], Ahmed Deshmukh[3], Saurabh Dake[4], Kapil Mundada[5]

[1,2,3,4]Scholar, Department of Instrumentation and Control Engineering,
Vishwakarma Institute of Technology, Pune, India.
[5]Assistant Professor, Department of Instrumentation and Control Engineering,
Vishwakarma Institute of Technology, Pune, India.
Email: [1]saurabh.baviskar15@vit.edu, [2]pushkar.kadhane15@vit.edu, [3]ahmed.deshmukh16@vit.edu,
[4]saurabh.dake15@vit.edu, [5]kapil.mundada@vit.edu

Abstract - Looking at the current scenario; increasing the speed of vehicles causes a lot of accidents and eventually a lot of deaths. The government has made it compulsory for all the transport vehicles in India to follow a maximum speed of 80 km per hour on highways. For this reason, the use of Speed Governors is made. Everywhere in the country, we find such zones that have certain speed limits. The objective of our research work is to develop an automatic speed control in which the speed of the vehicle in a particular zone will automatically be reduced to the specified speed of the concerned zone. We have detected the initial speed of the vehicle in the zone and the position of the accelerator. Upon detection, the position of the throttle valve is controlled. In turn the air to fuel ratio is controlled which controls the speed of the vehicle. The implementation of this prototype is not only possible on highways, but also on local roads to avoid accidents in the city as well.

Keywords - Speed Governors, Speed Limit Zone, Throttle Valve.

1. INTRODUCTION

The abstract mentions the major objective to formulate the idea of controlling speed. Thus for the formulation of the same, different speed controlling techniques are taken into consideration for knowing the various aspects behind the objective of this research. In our country, accident statistics are increasing day by day due to all the reasons mentioned in the abstract. As per the reports from the NHAI (National Highway Authority of India), The stats say that 1,20,518 accidents have taken place on state highways and 1,42,268 on national highways in the year 2015 [3]. Out of these, the maximum number of accidents has taken place due to excessive speeding of vehicles or due to geographical conditions[6]. Also, in the vicinity of some speed restriction zones, the idea of speed reduction can be implemented to avoid mishaps on the local roads. This technique can be used on signals to bide the people to follow signal rules and not break the signal. Taking into consideration the figures of the accidents and mishaps, the implementation of the idea of automatic speed control is necessary, especially in a developing country like India [4] Though various high-end cars already use the methods of speed reduction with the help of image processing, there are many factors in our idea which are capable of taking over the image processing methods of speed reduction [5]. Also, the other methods of speed reduction to avoid accidents are not pocket friendly. The prototype developed using the idea, can be used in any car and not only the high-end cars which all the people cannot afford.

From the table given below, we can see that the idea which we are trying to implement has many merits over any conventional assembly. Many a time, it is not always the fault of the driver. But this idea can be useful to avoid mishaps, whosever mistake it may be. This research can be implemented in many real-life applications to avoid major accidents and to save a life which is very important. The following given is the comparison of the existing mechanisms and the prototype proposed by us Table 1.

Table 1. Comparison- Existing & Proposed Mechanism

Comparison parameters	Existing mechanism	Proposed mechanism
Zone Detection	Digital Image Processing	Radio Frequency
Processing Time	More	Less
Power Consumption	High	Low
Cost	High	Low (Rs. 2500)
Drawbacks	Visibility, environmental conditions	Less than Existing

2. DESIGN OF PROPOSED SYSTEM

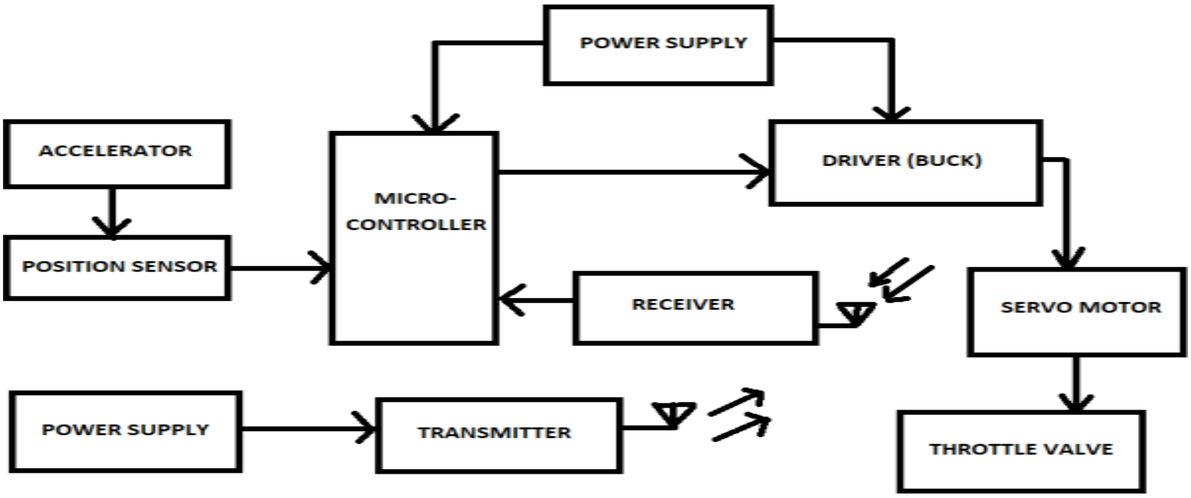


Fig.1. Block Diagram

3. SYSTEM DESCRIPTION

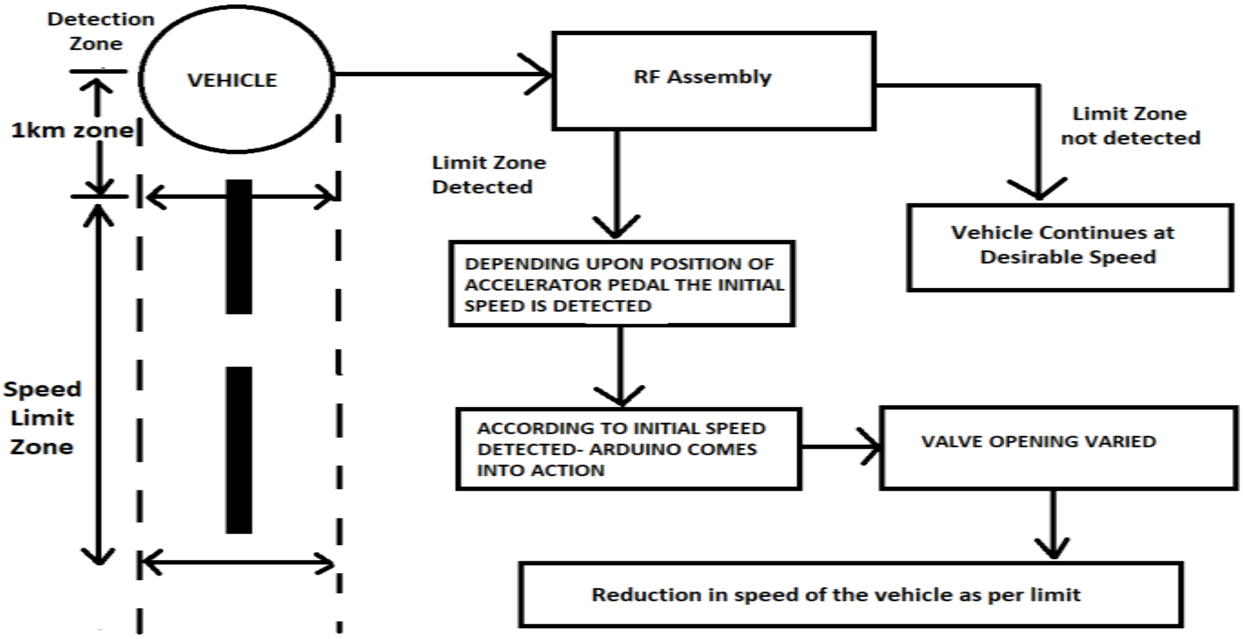

Detection Zone
VEHICLE
1km zone
Speed Limit Zone
RF Assembly
Limit Zone Detected
Limit Zone not detected
DEPENDING UPON POSITION OF ACCELERATOR PEDAL THE INITIAL SPEED IS DETECTED
Vehicle Continues at Desirable Speed
ACCORDING TO INITIAL SPEED DETECTED- ARDUINO COMES INTO ACTION
VALVE OPENING VARIED
Reduction in speed of the vehicle as per limit

Fig. 2. Schematic of the Prototype

As seen in the block diagram of the research work, we can see the micro-controller i.e. the Arduino. When the vehicle enters the zone which has a speed limit, the Transmitter- Receiver pair detects the vehicle. Subsequently, the initial speed of the vehicle and the position of the accelerator is detected. The position of the accelerator is detected using the Hall Effect sensor.

Depending upon the initial speed and the position of the accelerator, the Arduino comes into action. The Arduino provides a PWM signal to the buck driver. Depending upon the PWM signal, the driver gives the voltage to the servo motor and hence controlling the position of the throttle valve and thus controlling the air to fuel ratio.

We have used an ‘Arduino’ as an interfacing medium. An RF (Radio Frequency) Transmitter–Receiver Pair (Tx-Rx) is used to activate the speed limit zone and in turn the assembly when the car enters the speed limit zone. The Tx-Rx pair becomes high after the car enters the zone and the initial speed of the vehicle is detected according to the position of the accelerator. The code is written such that, the speed decreases 1 km before the limit zone. The initial speed of the vehicle is considered as 0-30 kmph, 30-60 kmph, 60-80 kmph and 80-100 kmph at different locations having a speed limit. According to the limit of the zone, the Arduino will take action and hence provide PWM signal to the buck driver. Depending upon the PWM signal, the driver gives the voltage to the servo motor and hence controlling the position of the throttle valve and thus controlling the air to fuel ratio. This in turn controls the speed of the vehicle.

Table 2. Accelerator position and voltage analogy

Accelerator opening %	O/P Voltage	I/P Voltage	Accelerator Position(cm)
0	2.44	5	4
10	2.46	5	3.6
20	2.47	5	3.2
30	2.48	5	2.8
40	2.51	5	2.4
50	2.54	5	2
60	2.59	5	1.6
70	2.68	5	1.2
80	2.81	5	0.8
90	3.11	5	0.4
100	3.23	5	0

4. SYSTEM EXECUTION

For the implementation of the prototype for testing it in the laboratory, a servo motor and a butterfly throttle valve are used as controls to control the air to fuel ratio. Thus following steps or algorithm explains the reduction in speed of a motor and position of the throttling valve. The proper flow of the code fed to the Arduino is also explained by a flowchart.

Steps of Execution:

1. The receiver pin is high that is the vehicle is in the zone.
2. The initial speed of the vehicle and the position of the accelerator is detected.
3. Depending upon the initial speed and position of accelerator, the arduino will detect the zone and respectively provide a PWM signal to the Buck driver.
4. Now depending upon the PWM signal, the driver will provide respective voltage to the servo motor.
5. According to the voltage provided to the servo motor, the position of the throttle valve is controlled and the air to fuel ratio is controlled.
6. When the car moves out of the zone, the control of the vehicle is shifted to the user.
7. Now the user may operate the vehicle according to his/her desired speed.
8. This cycle will continue as far as the speed zones are concerned.
9. The respective speed of the vehicle is thus controlled according to the initial speed and the speed limit of the zone with the help of the Arduino.
10. Hence, Arduino plays an important role in controlling speed.

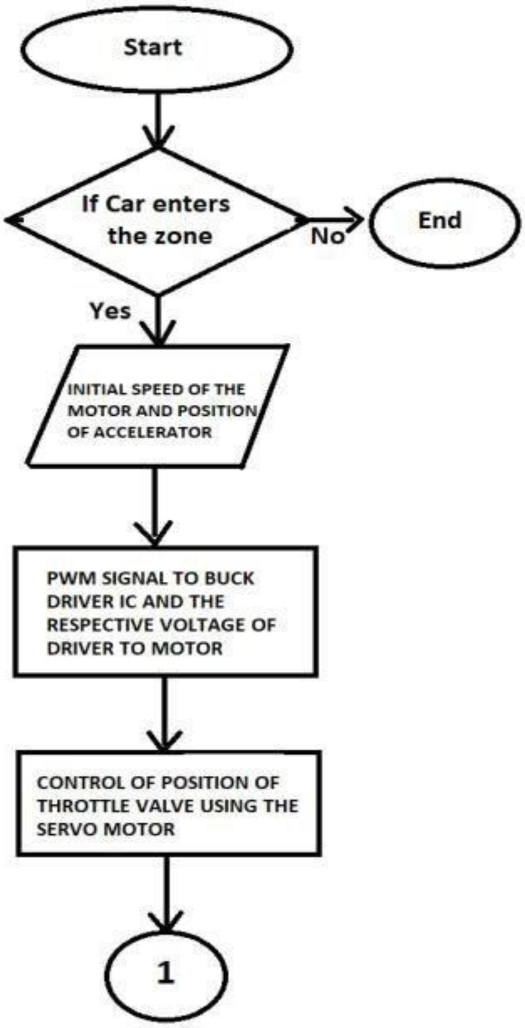

Start
If Car enters the zone
No
End
Yes
INITIAL SPEED OF THE MOTOR AND POSITION OF ACCELERATOR
PWM SIGNAL TO BUCK DRIVER IC AND THE RESPECTIVE VOLTAGE OF DRIVER TO MOTOR
CONTROL OF POSITION OF THROTTLE VALVE USING THE SERVO MOTOR
1

Fig. 3(a). Flow of Research Work

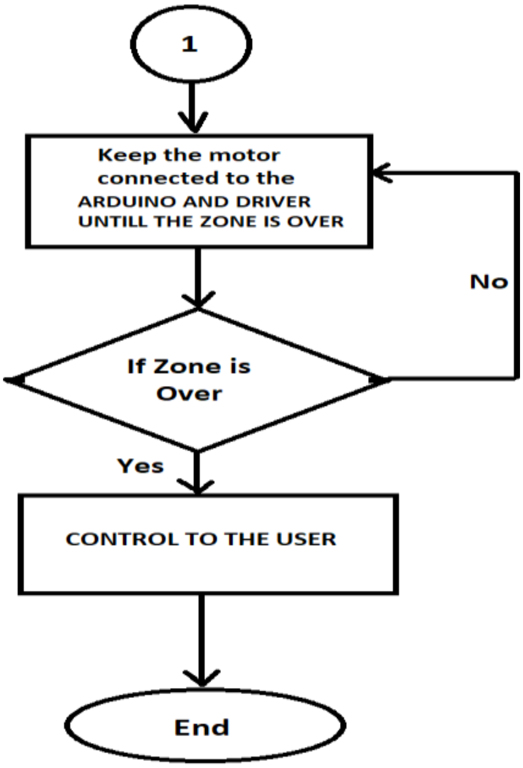

1
Keep the motor connected to the ARDUINO AND DRIVER UNTILL THE ZONE IS OVER
No
If Zone is Over
Yes
CONTROL TO THE USER
End

Fig. 3(b). Flow of Research Work

For the practical working of the same, the following table shows the analogy related to the required output of the research work.

- Voltage, Speed and PWM analogy

The table below states the different voltage values and the respective speed at the particular value of PWM.

Table 3. PWM Values and Corresponding Voltage and Speed

PWM values	Voltage	Speed (km/hr)
255	11.52	100
229.5	10.29	90
204	9.1	80
178.5	7.99	70
153	6.80	60
127.5	5.69	50
102	4.59	40
76.5	3.38	30
51	2.28	20
25.5	1.10	10
0	0	0

5. TESTING AND RESULTS

To prove the above-given flowchart, it is necessary to give a mathematical explanation of the prototype with the help of certain statistics and calculations. These calculations help to gain the proper values of speed, voltages, etc. necessary for the testing and working of this prototype. Therefore, the logic of Decrementation and the detection of initial speed are shown with the help of mathematical calculations and their respective values are tabulated for a proper understanding of the flow of the prototype.

A. Decrement Logic

The logic was to decrement the speed of the vehicles moving at different speeds at a different rate to avoid the collisions of the vehicle and that all the vehicles would reach the zone in the same time interval. This was achieved by doing calculations for the amount of negative acceleration to be provided for the vehicle so that the speed would reach to the desired value. A delay of some time is provided so that the reduction in the PWM is done after every delay of given time that is if in the program it has 5 loops then the speed of the motor will gradually decrease in 5 steps. Also, Table No. 4 shows that decrement logic and the negative acceleration required to stop the vehicle as well as the PWM conversion of it.

By Using Newton's Law:

$$\boldsymbol{v}^2 = \boldsymbol{u}^2 + \mathbf{2} \qquad (1)$$

Where, v = Final Speed,

u = Initial Speed, a = Acceleration, s = Distance.

Now, Initial Speed = 100 km/hr Final Speed = 30 km/hr

Distance = 1 km with a delay of 5 loops.

Therefore, From (1):

$$\mathbf{30}^2 = \mathbf{100}^2 + \mathbf{2}(\mathbf{a})(\mathbf{1})$$

$$= \frac{\mathbf{900} - \mathbf{1000}}{\mathbf{2}}$$

Hence, **a = - 4500 km/hr** ^{2}

Similarly, computing it for the initial speed of 90 km/hr, 80 km/hr, 70 km/hr, 60 km/hr, 50 km/hr and 40 km/hr:

Table 4. Speed Decrementation and corresponding Deceleration and PWM

Speed Decrementation	Deceleration	Relative Decrement in PWM
100 km/hr–30 km/hr	4500 km/hr^2	35.7
90km/hr – 30 km/hr	3600 km/hr^2	30.6
80 km/hr – 30 km/hr	2750 km/hr^2	25.5
70 km/hr – 30 km/hr	2000 km/hr^2	20.4
60 km/hr – 30 km/hr	1350 km/hr^2	15.3
50 km/hr – 30 km/hr	800 km/hr^2	10.2
40 km/hr – 30 km/hr	350 km/hr^2	5.1

B. Initial Speed Detection

The point was to detect the speed initially so as the Arduino would put the motor in the loop according to the decremental logic so basically, Arduino pin is incapable of reading the voltages above 5 volts so the 12 volt voltage to the motor was to be converted to the 5 words and further on in an analogous way for achieving that voltage divider circuit is used of which the calculations are as shown below.

Thus, **R2 = 76.45 KΩ**

The voltage Output corresponding to different speed values:

And, **R1 = 100 KΩ**

Therefore,

5 V = **11. 54**

The Voltage Conversion and Detection Analogy:

Using Voltage Divider Rule**:**

Now, **Vin= 11.54 V** and **Vout= 5V**

$$V_{out} = V_{in} \frac{(R2)}{(R1 + R2)} \qquad (2)$$

Thus, **R2 = 76.45 KΩ**

Table 5. Voltages corresponding to different speed values

Speed	Input Voltage	Output Voltage
100	11.54	4.58
90	10.29	4.14
80	9.10	3.66
70	7.99	3.16
60	6.80	2.73
50	5.69	2.31
40	4.59	1.80
30	3.38	1.37
20	2.28	0.92
10	1.10	0.47
0	0	0

5. CONCLUSION

Thus, according to the information, calculations, analogies, and different diagrams stated and explained above in detail; we can conclude by stating that this research is a useful product to avoid mishaps, accidents and to make people follow rules. This research work shall be implemented in many areas where speed should be restricted. Time is Money, but life is more important. Although this research work is implemented to abide by the rules, the traffic rules should be followed in all scenarios and is made to practice all the time. This in turn will guide our nation to be a developed one soon. Also, the advancement of this research brings some advantages to the prototype as compared to the previous one. The control of the air to fuel ratio will have more accuracy than the previous prototype.

REFERENCES

[1] Kenneth J. Ayala, “The 8051 Microcontroller Architecture, Programming and Applications”, Third Edition, Cengage India Publication, 2007.

[2] Mohammed Ali Mazidi and Janice Gillispie Mazidi, “The 8051 Microcontroller & Embedded Systems”, Second Edition, Pearson Publication, 2009.

[3] S. Naga Kishore Bhavanam, “Automatic Speed Control and Accident Avoidance of Vehicles Using Multi-Sensors”, International Conference on Innovations in Electronics and Telecommunication Engineering, July 2014.

[4] Fergus Tate, External Vehicle Speed Control, “Implementation scenarios”, University of Leads, October 1997.

[5] Manoharan, S., “An Improved Safety Algorithm For Artificial Intelligence Enabled Processors In Self Driving Cars”, Journal of Artificial Intelligence, vol. 1, No. 02, pp. 95-104, 2019.

[6] Sunil R. Kewate, S. V. Karmare, Nehal Sayankar and Siddharth Gavhale, “Automatic Speed Control System by the Color Sensor for Automobiles -An Innovative Model Based Approach”, International Journal of Advanced Mechanical Engineering, ISSN 2250-3234, Vol. 4, No. 2, pp. 223-230, 2014.

Classification of Phishing Websites using Extreme Learning Machine and Hybrid Bat Algorithm

Devika V[1], Thushara A[2], Manu J Pillai[3]

[1]PG Scholar, Department of Computer Science and Engineering,
T K M College of Engineering, Kollam, Kerala, India.
[2]Associate Professor, Department of Computer Science and Engineering,
T K M College of Engineering, Kollam, Kerala, India.
[3]Assistant Professor, Department of Computer Science and Engineering,
T K M College of Engineering, Kollam, Kerala, India.
Email: [1]devikavalsala@gmail.com, [2]tusharaa@gmail.com, [3]manujpillai@gmail.com

Abstract— Phishing is an electronic fraud through which an attacker can gain access to user credentials. Phishing websites are the one which mimics the legitimate websites and fraudsters evade their detection without much effort. The effect of phishing attack raises the necessity of anti-phishing mechanisms. Several approaches are there to recognize phishing websites such as whitelist, blacklist, machine learning and heuristic-based approach. This paper investigates extreme learning machine and hybrid bat algorithm for the classification of website phishing attacks. The features that distinguish the legitimate from the fraudulent website is selected using the random forest algorithm. The comparison of these techniques with other classification methods, for instance, support vector machine, logistic regression, decision tree and the random forest is also carried out. The results of the classification accuracy show that the hybrid bat algorithm achieves better classification accuracy compared to other classification methods.

Keywords - Phishing, Extreme Learning Machine, Neural Network, Hybrid Bat Algorithm.

1. INTRODUCTION

Phishing is becoming a progressively increasing threat to web security. Nowadays half of the emails received contain a spurious link that is redirected to phishing websites. These websites have a chance of being fraudulent ones where the attackers send fake emails or make use of fake websites that imitate the original ones, and through that, they can gain the user credentials quickly. There are also various methods to scam users such as VOIP, counterfeit websites and covert redirect. The attacker creates counterfeit websites which are built efficiently and look similar to the original website in its layout. The most challenging issue is that the forged website acts as legitimate sites, and it becomes tough to distinguish whether it is a phishing site or not. The phishers also exploit the limitations of the weaker web servers and ignorance of security policy by the end-users. Attackers also purchase web security certificates to convince the end-users that the phishing website is a legitimate one.

The official reports indicate that the phishing attacks are increasing for the past half decades. There are strict laws against such activities, and violations of this lead to the punishment, which includes fine, or imprisonment. The law enforcement authority should take quick action against these kind of violations as the fake websites may live only for a few hours, and after performing such illegal activities, the attackers disappear into the internet. To prevent this kind of violation, the Government of India passed the anti-phishing act under Section 43 of the Information Technology Act, 2000 [1]. Even though the act could prevent the phishing attack to a certain extent, the act is unable to stop the phishing completely. So to reduce the phishing attack, anti-phishing tools and heuristic-based solutions have emerged. There are various approaches to identify and avert phishing scams, and it should correctly predict the attack for safeguarding the victims.

To dodge the anti-phishing methods, the attackers adopt new methods, which make users challenging to identify the credibility of websites [2]. The attacker uses a massive amount of public information in methods such as social phishing and context-aware phishing. Different methods are there to prevent phishing, such as blocking the phishing sites and filtering the emails. The filtering can be done using different classification methods. In this work, machine learning approaches are used to discover the phishing websites.

Phishing detection is a classification problem where the data set containing the phishing information is classified into different classes. The machine learning algorithms such as Naïve Bayes, Support Vector Machine or Support Vector Networks, decision tree, random forest and neural networks are used to solve the classification problem. Although there are many classification methods, the neural network such as extreme learning machine

(ELM) and the hybrid bat algorithm is considered here. Extreme learning machine which helps in classification, compression, regression is a feed-forward neural network where connections are not in the form of a cycle. There are several variants of Extreme learning machine such as Incremental ELM, Error-minimized ELM, Online sequential ELM, Ordinal ELM, Voting-based ELM, and Symmetric ELM. The neural network which uses the hybrid bat algorithm focuses on parameter setting of the neural network. Compared to other classification methods like logistic regression, support vector machine, decision tree and random forest, the neural network that uses the hybrid bat algorithm acquires better classification accuracy. This study aims to classify and predict the phishing websites by using extreme learning machine and, hybrid bat algorithm. Comparison of other classification methods is also explored.

The remaining paper is organized as follows. Session 2 presents related works in phishing classification and prediction approaches. Section 3 shows the methodologies used in this study. Section 4 reveals the results and the conclusion and future work are given in section 5.

2. RELATED WORK

This section surveys the various anti-phishing approaches and methods adopted in establishing solutions to decrease the false-positive rate. The two main approaches used in the classification of phishing websites are whitelist and blacklist approaches.

The Whitelist approaches have a list of legitimate websites and their associated information. The IP address is included in the list, and it should be modified or updated if any change occurs. Websites that are not included in this list are suspicious and can be phishing or legitimate site. In the blacklist approaches instead of having the list of legitimate sites, list of phishing websites is included. Websites in this list are most vulnerable or considered as phished, and the websites which are not belonging to this list are safe from the attacker or phisher. In this approach, the requested URL is compared with phished old URLs. In [3], the authors used URL similarity checking techniques to prevent users from accessing phishing websites. The entry is given only after determining whether the site is safe or under the attacker's influence. In [4], a blacklist approach has been proposed. In this method, the URL is divided into multiple components by using a matching algorithm and finally compares it with the entries in the blacklist, which are considered as phishing websites. The method proposed in [5] named "CANTINA" detects the phishing websites by making use of term-frequency-inverse-document-frequency (TF-IDF) rates. The CANTINA is the abbreviation for 'Carnegie Mellon Anti-phishing and Network Analysis Tool'. A content-based technique is used in this work. By the use of TF-IDF, it determines whether the webpage is phishing or legitimate by checking the content of the webpage. A webpage is considered as a legitimate webpage when topmost search results have current webpage specified; otherwise, it is a phishing webpage. The main drawback of using this technique is that the number of search results is fixed and it is set to 30 in this work. When no result is found, or the browser returns zero, then the website is phished. To solve this problem, the TF-IDF combined more features such as the age of the domain, known images, suspicious URL, IP address, dots in URL, and forms are used. Another drawback of this approach is that it cannot identify legitimate webpages which contain only images.

The method employed in [6] is based on classification algorithms. In this approach, 27 different features have been selected from different websites. The gathered features belong to the three sets of values like genuine, legitimate, and doubtful. The experiments were conducted to evaluate the selected features using data mining techniques such as MCAR [7], C4.5 [8], and PRISM [9]. The outcome displayed a dominant relation between the two features named Domain Identity and URL. In [10], fuzzy data mining techniques have been used for the prediction of the phishing website. In this, 27 features are selected by the authors to construct a model to predict the type of websites. This model failed to specify the features extraction process, even though the authors categorized websites into very legitimate, legitimate, doubtful, phished and extremely phished they are not specifying the factor that makes the classes different.

Frederick Livingston [11] proposes the implementation of a random forest algorithm using weka tool. Because of the complexity of the algorithm, only the variable importance part was implemented using weka tool. According to his work, the random forest java application can implement the Breiman's algorithm completely and is suitable with the Weka's datasets. In [12], the author proposes a comparison of random forest and j48, which was implemented on many datasets. For the selection of the appropriate model, a baseline will be created using this work. In the cases of large datasets, random forest achieves increased classification performance and acquire results that are more accurate.

The work [13] discusses how artificial neural network is applied for phishing website classification problem. The paper includes the concept of artificial neural network and differences between artificial neural network, computers and biological neurons. The applications of artificial neural networks contain the prediction of thrift failures, prediction of the stock price index, OCR systems, industrial process control, data validation. The limitations are also listed in work, and it concludes by reminding that the need for artificial intelligence is rapidly increasing. In [14], an extreme learning machine with a single feed-forward network was proposed, and it tends to reach the smallest training errors. For feed-forward neural networks with several hidden layers, algorithms such as backpropagation which is a gradient-based learning algorithm, can be used. It runs extremely fast and gives better performance.

The method proposed in [15] presented the main concepts of swarm intelligence with a specific centre of attention on two swarm intelligence influenced optimization techniques. In addition to that, a comparison of swarm intelligence algorithm with an artificial neural network and the genetic algorithm has been carried out. In [16], bat algorithm frequency tuning has been used to raise the miscellaneousness of the solutions in the population is discussed.

The description and characteristics of bat echolocation are also discussed in this paper, and the algorithm makes use of echolocation behaviour of bats. There are several variants of bat algorithm like fuzzy logic bat algorithm, multiobject bat algorithm, k means bat algorithm, binary bat algorithm, chaotic bat algorithm, differential operator and levy flights bat algorithm and improved bat algorithm. In [17] an advanced form of bat algorithm proposed, named hybrid bat algorithm to upgrade bat algorithm for higher dimension problems. A differential evolution strategy is used for hybridizing bat algorithm, and the results are improved than that of the original bat algorithm.

3. METHODOLOGY

The extreme learning machine and the hybrid bat algorithm are the methods used here for the classification and prediction of a phishing website. The first step is the data collection, followed by the feature selection using random forest. Then the dataset is split into the training set and the testing set. The model is trained using the training set. Subsequently, the model is evaluated using the testing set. The proposed framework is shown in Figure. 1.

A. Data Collection

The dataset is obtained from the UC Irvine Machine Learning Repository and which contains 30 attributes and 11,055 data. The attribute or the feature values in the dataset are phishing (-1), legitimate (1) and suspicious (0). The labels are -1 for phishing and 1 for not phishing. The features are mainly divided into four groups [18], such as Address Bar based Features, Abnormal based Features, HTML and JavaScript-based Features, and Domain-based Features.

B. Feature Selection

Feature selection retrieves the most relevant features automatically or manually. The feature selection algorithm used in this work is the random forest, which automatically selects the features based on the feature importance. All the 30 features are assigned a feature importance score; out of this 30 feature, the most dominant 16 features are selected by the algorithm. The sample dataset is as given in Table I. This algorithm helps to improve classification and prediction accuracy by creating a large number of trees.

C. Extreme Learning Machine

Extreme learning machine is a feed-forward neural network model which utilizes activation function, rectified-linear -unit (ReLU). The count of hidden units in ELM is greater than other neural networks which are trained by utilizing the backpropagation algorithm. The weights from input to the hidden layer are arbitrarily generated in this approach. In order to find the output weights, which is a least-square error regression problem, the minimization of least square error between training labels and predicted labels are carried-out using the following solution.

$$\beta=\left(X^{T}X\right)^{-1}X^{T}y \tag{1}$$

Eq. (1) [19] indicates error minimizing between predicted labels and training labels represented as β where X is the input to hidden layer matrix and y is training labels. A function is also added to predict the output. ELM has faster learning capability than other neural networks and is also scalable. Extreme learning machine has application in various areas including image processing, speech application, computer vision, medical or biomedical applications and system modelling and prediction. This extreme learning machine algorithm performs better than the other conventional neural network learning algorithms. In a traditional neural network, a large number of training parameters need to be set whereas in case of ELM [20] it generates an optimal solution by fixing the number of nodes in the hidden layer.

D. Neural Network Using Hybrid Bat Algorithm

Neural network maps a given input to the desired output. The features selected by the random forest algorithm are taken as the input.

Table 1. Sample dataset after feature selection

	SSL Final State	URL of Anchor	Web Traffic	Having Sub Domain	URL Length	Result
0	-1	-1	-1	-1	1	-1
1	1	0	0	0	1	-1
2	-1	0	1	-1	1	-1
3	-1	0	1	-1	0	-1
4	1	0	0	1	0	1

The parameters of the feed-forward neural network that are set by a hybrid bat algorithm and the activation function, ReLU, is the same as in ELM for input and hidden layer. The model is trained and evaluated. Since there are 16 features, there are 16 neurons in the input layer and only one output neuron is required. The output layer uses the sigmoid function as the activation function. The network parameters can be enhanced with the help of the hybrid bat algorithm. The initial population is represented as vectors as given in equation (2) [21]

$$x_i^t = \left(x_{i,0}^t, \dots \dots x_{i,n}^t\right), \text{for i=0,...p-1} \qquad (2)$$

where, x_i^t represents the individuals and p represent the population. Then the real values are mapped with respect to equation (3), equation (4) and equation (5) [21].

$$\mathrm{e} = \mathrm{x[i]} * 100 + 1 \qquad (3)$$

$$\mathrm{l} = \frac{\mathrm{x[i]}}{10} \qquad (4)$$

$$b = \frac{x[i]*100}{2} + 10 \qquad (5)$$

where e is the number of epochs, l is the learning rate and b is the batch size, and the highest integer values of e and b are taken. The swarm intelligence [21] family includes algorithms that mimic biological features of few special animal species such as their movement and behaviour for decision making purpose. The behaviour of micro-bats is adopted in the bat algorithm, which is a part of the swarm intelligence family is proposed. The main steps in the bat algorithm are as initialization, generalization of the solution, local search step, evaluation of the result, conditional savings of best result and finding the best result. In hybrid bat algorithm (HBA), the local search step from bat algorithm is eliminated, and mutation strategy included. The parameter tuning of the neural network is an optimization problem where this hybrid bat algorithm can be used. The parameters which are considered by the neural network are batch size, epochs, learning rate and neurons in the first hidden layer. The hidden layer also uses ReLU activation function, and the number of neurons is fixed. The cost function used here is binary cross-entropy and can be represented as in equation (6) [22]:

$$CE = -\sum_{i=1}^{C=2} t_i \log s_i \qquad (6)$$

CE represents the cross-entropy loss and t_i and s_i are target value and the output scores respectively.

Instead of classical stochastic gradient descent method, the optimizer used in this study is Adam [23]. This optimizer sets single learning rates to update the weights and maintains a per-parameter learning rate. It needs only first-order gradients because from evaluation of that gradient which includes first and second moments. This procedure computes individual learning rates. The Adam provides a combination of advantages of two other algorithms named adaptive gradient algorithm (AdaGrad) and root mean square propagation (RMSProp) both maintain per-parameter learning. Adam is computationally well organized, and it has only little memory requirements. So, this method comprises of neural network which used hybrid bat algorithm optimized parameters for tuning neural networks. The parameters optimized in order to easily acquire the highest accuracy when classifying phishing websites.

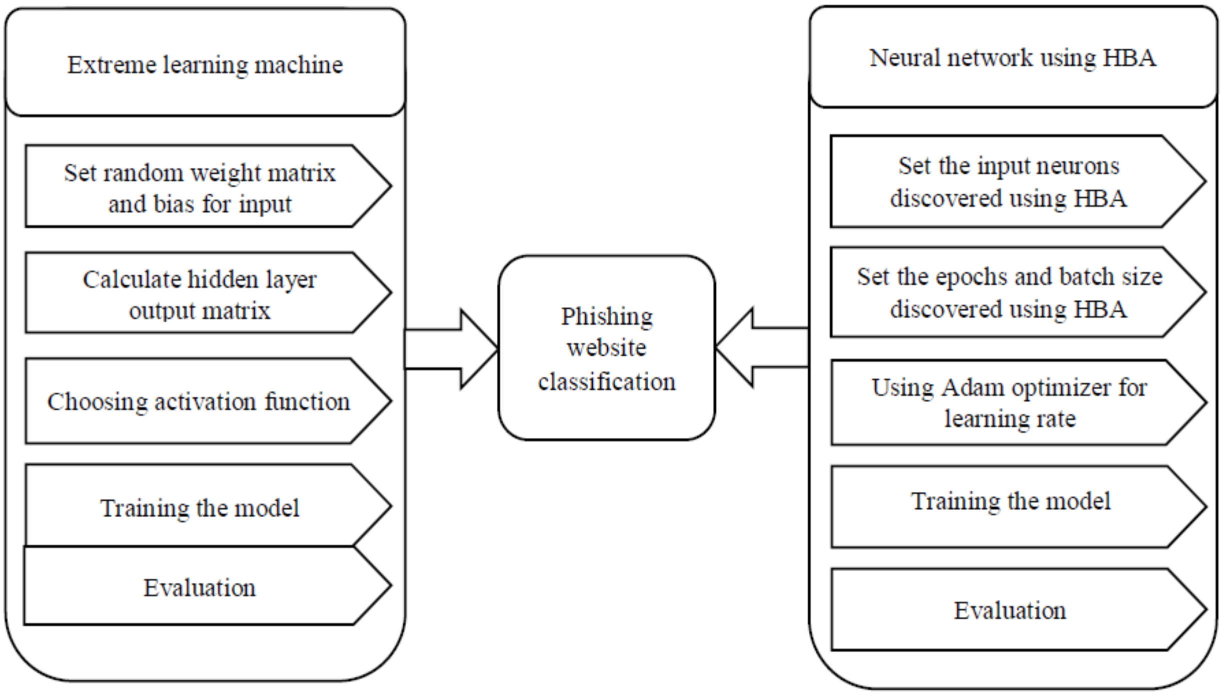

Extreme learning machine
Set random weight matrix and bias for input
Calculate hidden layer output matrix
Choosing activation function
Training the model
Evaluation
Phishing website classification
Neural network using HBA
Set the input neurons discovered using HBA
Set the epochs and batch size discovered using HBA
Using Adam optimizer for learning rate
Training the model
Evaluation

Fig.1. The Framework of The Proposed System

4. RESULTS AND DISCUSSION

The experiments are conducted on the dataset obtained. The python programming language along with its libraries such as are Keras, NumPy, Matplotlib, scikit-learn and pandas is used to implement this proposed work. The random forest algorithm is applied to the dataset, and features are selected based on feature importance. The feature importance score of the 30 features which are listed in Table III is shown in Figure 2. The seventh feature, SSL Final State, has the highest score (0.338535), and the top 16 features are selected from the 30 features.

The observation of dataset is given in Table II, where the websites are marked as phishing or legitimate. The data set is split into 70% for training, and 30% for testing. In extreme learning machine, the number of hidden layers is set to 1000, and the weights between the hidden layer to the output layer is computed using the error function. Then the activation function, ReLU, is applied to the model, and the model is trained for the training data set. Then the testing data set is evaluated on this trained model. This model classifies 3098 samples and classification accuracy of 93.39% are achieved.

Table 2. The observation of dataset

Index	Features
0	Having the IP Address
1	URL Length
2	Shortening Service
3	Having At Symbol
4	Double Slash Redirecting
5	Prefix Suffix
6	Having Sub Domain
7	SSL Final State
8	Domain Registration Length
9	Favicon
10	Port
11	HTTPS Token
12	Request URL
13	URL of Anchor
14	Links in Tags
15	SFH
16	Submitting to Email
17	Abnormal URL
18	Redirect
19	On Mouseover
20	RightClick
21	PopUpWidnow
22	Iframe
23	Age of Domain
24	DNSRecord
25	Web Traffic
26	Page Rank
27	Google Index
28	Links Pointing to Page
29	Statistical Report

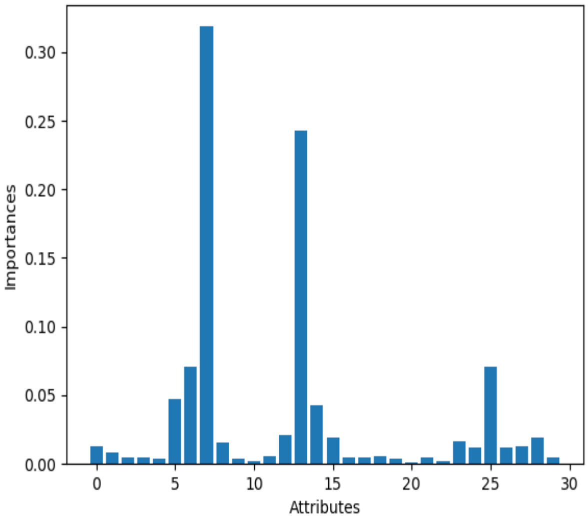


Fig. 2. Feature importance using the Random Forest Algorithm

The neural network tuned with hybrid bat algorithm used the same selected features for classification. The input layer of this neural network contains 40 neurons, and the activation function for the input layer is ReLU with the input dimension set to16 features. The hidden layer consists of 30 neurons, and the same activation function, ReLU is applied. The output layer has only one neuron, and the activation function for the output layer is the sigmoid function. The neural network parameters such as epochs, batch size and learning parameter are optimized by using the hybrid bat algorithm. The neural network, which is tuned by the hybrid bat algorithm achieved an accuracy of 96.71%. The neural network which used the hybrid bat algorithm achieved more accuracy than the extreme learning machine with the same number of features as input. The accuracy with respect to epochs is plotted and is shown in Figure.3.

Table 3. Features in dataset

	Class	Number of observations
0	1	6157
1	-1	4898

The accuracy of the extreme learning machine, hybrid bat algorithm and other classification algorithms such as logistic regression, support vector machine, decision tree and random forest are compared. The metrics used to compare the classification models are accuracy, precision, recall, f1-score and support. The mathematical representation of these metrics is shown in equation (7), (8) and (9).

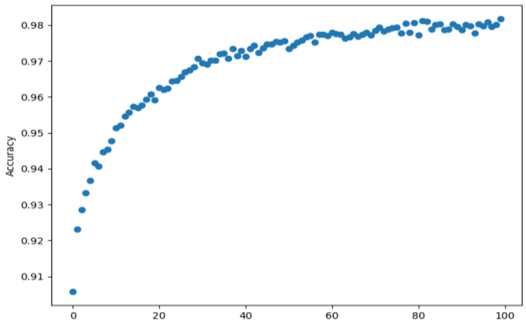


Fig.3. accuracy plot of neural network with optimized values

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \quad (7)$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \quad (8)$$

$$\text{Accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (9)$$

Where, the number of true positives, true negatives, false negatives and false positives is TP, TN, FN and FP, respectively. The f1 score is the weighted average of precision and recall, and the support is the actual number of occurrences of the class in the given dataset. The confusion matrix of all classification methods is generated.

Table 4. Performance metrics of Random Forest

Precision	Recall	F1-score	Support	
0.97	0.95	0.96	1520	-1
0.95	0.97	0.96	1797	1
0.96	0.96	0.96	3317	avg/total

Table 4 shows the values of the metrics obtained for the random forest classification method. In comparison with all other classification methods, the neural network with optimized parameters has the highest accuracy. The comparison of the accuracies of different classification methods is shown in Table 5.

Table 5. The comparison of the accuracies

Algorithm	Accuracies
Logistic regression	91.64%
Extreme learning machine	93.39%
Support vector machine	94.09%
Decision tree	95.41%
Random forest	95.93%
Neural network using HBA	96.71%

5. CONCLUSION

In this paper, two techniques are used to classify phishing website. The analysis of extreme learning machine and neural network using the hybrid bat algorithm for parameter setting is performed on the selected attributes. The hybrid bat algorithm with enhanced parameters performed better than the extreme learning machine and other classification algorithms. The limitation of this work is that it is difficult to distinguish between adversarial and legitimate website. As the future work, the classification model can be improved to handle the adversarial attack, and the prediction accuracy of the phishing websites can be improved. Also, the classification model can be validated using other datasets.

REFERENCES

[1] Yogesh Prasad Kolekar, Protection of data under Information Technology law in India, Social Science Research Network, 2015.

[2] Kang, J. & Lee, D., "Advanced white list approach for preventing access to phishing sites", International Conference on Convergence Information Technology, Gyeongju, Korea, 2007.

[3] Prakash, P., Kumar, M., Kompella, R.& Gupta, M., "PhishNet: predictive blacklisting to detect phishing attacks", IEEE INFOCOM, San Diego, USA, 2010.

[4] Y. Zhang, J. Hong and L. Cranor, "CANTINA: A Content-Based Approach to Detect Phishing Web Sites", in Proceedings of the 16th World Wide Web Conference, Banff, Alberta, Canada, 2007.

[5] Aburrous, M, Hossain, M. A., Dahal, K. and Fadi, T, "Predicting Phishing Websites using Classification Mining Techniques," Seventh International Conference on Information Technology, Las Vegas, Nevada, USA, 2010.

[6] F. Thabtah, C. Peter and Y. Peng, "MCAR: Multi-class Classification based on Association Rule", The 3rd ACS/IEEE International Conference on Computer Systems and Applications, 2005.

[7] J. R. Quinlan, "Improved use of continuous attributes in c4.5", Journal of Artificial Intelligence Research, pp. 77-90, 1996.

[8] M. Aburrous, M. A. Hossain, K. Dahal and F. Thabtah, "Intelligent phishing detection system for e-banking using fuzzy data mining", Expert Systems with Applications: An International Journal, pp. 7913-7921, December 2010.

[9] J. Cendrowska, "PRISM: An algorithm for inducing modular rule", International Journal of Man-Machine Studies, pp. 349-370, 1987.

[10] Frederick Livingston, "Implementation of Breiman's Random Forest Machine Learning Algorithm", 2005.

[11] Jehad Ali, Rehanullah Khan, Nasir Ahmad, Imran Maqsood, "Random Forests and Decision Trees", September 2012.

[12] Manish Mishra, Monika Srivastava, "A View of Artificial Neural Network", August 2014.

[13] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew, "Extreme Learning Machine: A New Learning Scheme of Feedforward Neural Networks".

[14] Xin-She Yang, Xingshi He, "Bat Algorithm: Literature Review and Applications", International Journal of Bio-Inspired Computation, Vol. 5, No. 3, 2013.

[15] Iztok Fister Jr, Du san Fister, and Xin-She Yang, "A Hybrid Bat Algorithm", Elektrotehniski Vestnik, Vol. 80, pp. 1–7, 2013

[16] Hazem Ahmed, Janice Glasgow, Swarm Intelligence: Concepts, Models and Applications, February 2012.

[17] Rami M. Mohammad,Fadi Thabtah, Lee McCluskey, "Phishing Websites Features", Neural Computing and Applications, vol. 25(2), pp.443-458 , August 2013.

[18] Mark D. McDonnell, Migel D. Tissera,Tony Vladusich Andre van Schaik and Jonathan Tapson, "Fast, simple and accurate handwritten digit classification by training shallow neural network classifiers with the 'extreme learning machine' algorithm", Neural and Evolutionary Computing, 2015.

[19] Shifei Ding, Han Zhao, Yanan Zhang, Xinzheng Xu, Ru Nie, "Extreme learning machine: algorithm, theory and applications, Artificial Intelligence Review", Vol. 44, No. 1, 2015.

[20] Grega Vrbancic , Iztok Fister Jr., Vili Podgorelec , "Swarm Intelligence Approaches for Parameter Setting of Deep Learning Neural Network: Case Study on Phishing Websites Classification", International Journal on Artificial Intelligence Tools, 2019.

[21] Raul Gomez, Understanding Categorical Cross-Entropy Loss, Binary Cross-Entropy Loss, Softmax Loss, Logistic Loss, Focal Loss and all those confusing names, https://gombru.github.io/2018/05/23/cross_entropy_loss/. 2018.

[22] Diederik P Kingma, Jimmy Lei Ba, "Adam: A Method for Stochastic Optimization", The 2nd International Conference on Learning Representations (ICLR), 2015.

Calculation of BMI using Voice and MLP

Shravani Bhosle[1], Aiman Saba Ansari[2], Seema Redekar[3], Vaishna Shetty[4]

[1,2,4] Scholar, Department of Information Technology, SIES Graduate School of Technology, Navi Mumbai, India.
[3]Professor, Department of Information Technology, SIES Graduate School of Technology, Navi Mumbai, India.
Email: [1]shravani.bhosle217@siesgst.ac.in, [2]aimansaba.ansari217@siesgst.ac.in,
[3]seema.redekar@siesgst.ac.in, [4]vaishna.shetty217@siesgst.ac.in

Abstract - Overweight and stoutness are significant dangers to numerous well being maladies just as it can cause perpetual mental impact. With the assistance of Body Mass Index prominently otherwise called BMI is utilized to tell whether an individual is having typical weight, overweight or having stoutness. BMI have a few blemishes now and again like overestimates in competitors, belittles in old, and when an individual experiences Oedema. Recently the specialists have been centering after deciding the connection between individual's voice and their BMI. This paper centers around producing BMI utilizing discourse signal with the assistance of Mel Frequency Cepstral Coefficients (MFCCs) and preparing will be completed on these examples utilizing Multi Layer Perceptron (MLP).

Keywords - BMI, Speech Signals, Feature Extraction, Classification, Obesity, Multilayer Perceptron, MFCC, Keras.

1. INTRODUCTION

The reason behind obesity is hereditary as well as environment and diet and exercise schedule. Obesity and Overweight is not just a superficial issue. It's a well-being peril. A person who is 40% overweight is twice as liable to pass away than a normal weight individual. This is on the grounds that weight and overweight has been lead to a lot of genuine ailments like heart illnesses, stroke, and diabetes, osteoarthritis, gout and gallstone and pregnancy issues. Modification in diet, increased exercise and behavior changes can help a person to lose weight. Prescribed medicines, healthy lifestyle and exercise are some alternatives for treating obesity.

The meaning of overweight and stoutness is having an excess of muscle versus fat to such an extent that it is threat to well-being. A solid method to decide if an individual has an excess of muscle to fat ratio is to ascertain the proportion of their weight to their tallness squared. This proportion, called the weight file (BMI), clarifies that the way that taller individuals have more tissue than shorter individuals, accordingly they have more weight. BMI is important as it is broadly regarded that chances of having a longer and healthier life are enhanced if a person have a healthy BMI. BMI is checked regularly by doctors to check whether a person is having a healthy weight depending upon BMI reading. BMI categories are mentioned below[3]:

Table 1. BMI Reading

BMI Category	Description
Below 18.5	To consume more food in order to gain weight
Between 18.5-25	Normal weight
Between 25-30	Exercise more to lose weight preventing the risk to obesity
Over 30 BMI	Dietary regime's and weight loss programs are proposed and are to be followed

The downsides of standard indicative, the specific weight isn't known by certain patients during assessment of BMI in far off areas where estimation is done through long strategy, likewise there is trouble in increasing patient data because of absence of correspondences gear in all likelihood inclined to blunder at the hour of conclusion of BMI.

This paper proposes on generating BMI using speech signal. For this, the dataset is generated by collecting voices of people with different age group tentatively between 18 to 50 and then with the help of librosa library in python 3.6; Mel-frequency cepstral coefficients (MFCCs) features will be extracted from that voice signals. Training will be carried out on these samples using Multilayer Perceptron (MLP); thus BMI will be calculated.

Need of Project- "Calculation of bmi using voice and mlp" can be used in medical sector for determining the BMI of patient. If the BMI of the patients is not appropriate then that person can be considered as unhealthy and accordingly medication and treatment shall be given. Also it can be used by senior citizen for determining their BMI at home for their routine check-up rather than going at diagnostic centre. The basic motivation behind this idea is entrenched in the gyms where it finds it's primary use. So, in order to tailor the diet and gym routine of the customers analysis of the BMI using this method proves to be cost-effective and time efficient. Apart from this it can be used at schools for determining the BMI of students as a of part of yearly check-up. Insurance companies can use it as a part of their health check-up's for giving customer their premium health policies.

2. LITERATURE SURVEY

Chawki Berkai et. al[1], "Estimation of BMI status via speech signals using short-term cepstral features," this paper include categorises of two different speech feature that are extracted for calculating the BMI. Methodologies that were used are KNN and PNN for doing the classification.

C. Berkai et.al[2],"Prediction of body mass index using speech signals: A review," describes previous work on estimating BMI using voice signals. Different methodologies which were used previously used are the GMM, the ANN and the HMM. Also the author highlights issues that were encountered.

C. Tai et.al[3], "A Framework for Healthcare Everywhere: BMI Prediction Using Kinect and Data Mining Techniques on Mobiles," built BMI prediction application using Kinect and data mining technologies. BMI was predicted based on facial feature which was the data stored in database and verified the compliance using visualization technique.

C. Yadav et.al [4], "Using Brain MRI Images to Predict Memory, BMI Age," predicted memory, age and BMI from Brain MRI images using 3D CNN architecture. The model developed by this author only predicts mean value for all parameters.

Also heatmaps are visualized. H. A. Rahman et.al[5], "Analysis of correlation between BMI and human physical condition using resonant field imaging system (RFI)," used resonant field imaging for examining the correlation between BMI and person's health. Also the author concluded that between 18.5-25 are categorised as healty as compared to compared to other categories.

R. M. Simply et.al[6],"Diagnosis of Obstructive Sleep Apnea Using Speech Signals From Awake Subjects," proposes a method using speech analysis for OSA detection and also estimates when the subject is preferably awake than asleep.

G. N. Jayabhavani et.al[7], "Silent speech generation using Brain Machine Interface," proposes a method to generate silent speech using brain machine interfacing (BMI) system. This system is efficient for people who are disabled and paralyzed.

M. Abdollahian et.al[8], "The Impact of Body Mass Index on Low Birth Weight," proposes a simple and multi-regression model to evaluate the mother pre-pregnancy BMI, age, gestation age at the time of delivery on the new born weight for low birth weight babies.

3. EXISTING SYSTEM

In the existing system KNN and PNN were used and during analysis it was concluded that PNN was more efficient and overpowered KNN. With the proposed system MLP will be used and it is more efficient, lightweight and it requires less time to compute.

A. K-Nearest Neighbour (KNN)

The KNN algorithm is a process about a solid procedures for perceiving objects design dependent on the pre-trained element information space [19, 20]. KNN, depending on labelled input data tends to learn the function that creates an output when new unlabeled input is given. The object is classified by calculating the distance of object which is close to individuals and vote of its neighbors, with the subject being given single class also, this may be generally partaken in its k nearest-neighbors (k is a positive whole number).For the KNN calculation, the structure of a new test include vector is relied upon the class of its KNNs. For this situation, the KNN calculation is executed utilizing Euclidean separation measurements to restrict the closest neighbors. The quantity of neighbors appointed (i.e .... k ...) used to sort the new test vector is differed from 1 to 10.[1]

A. Probabilistic Neural Network(PNN)

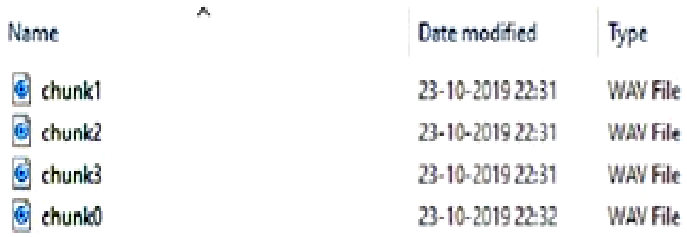
Probabilistic Neural Network maps any input data to a number of classifications, where it very well may be required into a more general function estimated, furthermore, PNN is a feed forward neural network in which operations are performed into multilayered network consist of four layers these includes: input layer, pattern layer, summation layer, output layer. In spite of its multifaceted nature, PNN just has a single training parameter. This is a subdue parameter of the probability density functions (PDFs) which are used for the stimulation of the neurons in the pattern layer. Along these, the training procedure of PNN exclusively requires a solitary input-output signal pass so as to compute network response. However just the ideal estimation of the smoothing parameter gives the chance of accuracy of the model's reaction as far as speculation capacity. The estimation of must be assessed based on the PNN's order execution which is generally accomplished in an iterative way[1].

4. PROPOSED SYSTEM

Objective:

1. To calculate BMI using voice.
2. To increase the accuracy.
3. To make it lightweight and would require less computational power.
4. To make it of minimal cost equivalent to zero.

The arrangement comprises of 5 stages, as portrayed in "Fig. I": framework contains five significant units: Data collection, Pre-processing, Feature extraction, Classification and Results/Output[1].



Name	Date modified	Type
chunk1	23-10-2019 22:31	WAV File
chunk2	23-10-2019 22:31	WAV File
chunk3	23-10-2019 22:31	WAV File
chunk0	23-10-2019 22:32	WAV File

Fig. 2. Data set

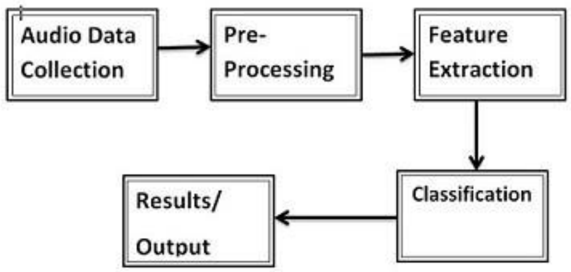


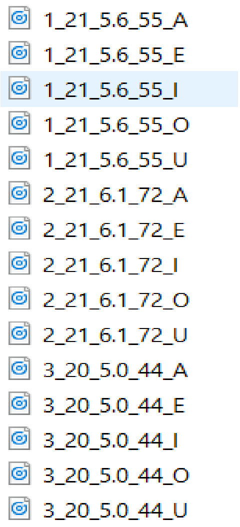
Fig. 1. Schematic Illustration

A. Audio Data Collection

For full filling the need of the data, we have collected data from multiple locations from nearby vicinity. The total numbers of voice samples are 250.The attributes of this data-set are age, weight and height. The voice samples that are collected was between the age group of 18-50.

B. Pre-Processing

The speech signal was recorded at frequency of 44.1 KHZ in m4a format. Initially, there are 5 recordings of all the vowels present in english language such as "A A A A A", "E E E E E" and so on. Later on these recordings are divided into a single alphabet as "A" "A" using python library pydub and the format was changed to wav using subprocess library. **ffmpeg** command was used for conversion. For splitting the audio files we considered minimum silence time as 500 milliseconds and for calculating the minimum silence threshold a parameter called dBFS(Decibels relative to full scale) was used and threshold was -16 dBFS. In digital system, it is used for measuring the amplitude level. This is typically used when maximum peak levels are defined. Also, level of 0 dBFS is set to the maximum possible digital level.



1_21_5.6_55_A
1_21_5.6_55_E
1_21_5.6_55_I
1_21_5.6_55_O
1_21_5.6_55_U
2_21_6.1_72_A
2_21_6.1_72_E
2_21_6.1_72_I
2_21_6.1_72_O
2_21_6.1_72_U
3_20_5.0_44_A
3_20_5.0_44_E
3_20_5.0_44_I
3_20_5.0_44_O
3_20_5.0_44_U

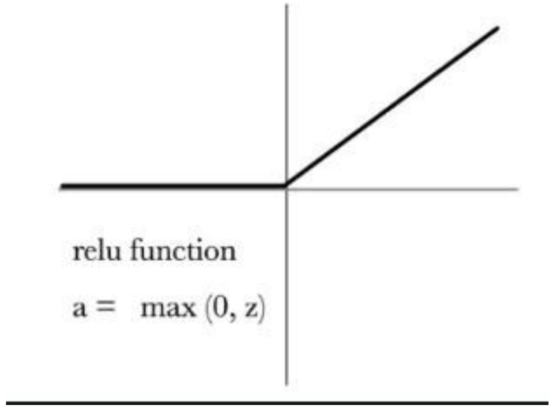
Fig. 3. Audio Chunks

C. Feature Extraction

The original sampling frequency of speech samples is 44.1 KHZ and speech signals (a,e,i,o,u) are tested at the same frequency. Short-term cepstral based feature extraction method named MFCC i.e Mel Frequency Cepstral coefficient is used in speech signal processing applications to extract known features. MFCCs containing of certain focal points from which to catch the acoustic features of audio recognition and has a decent limit with regards to separation and the possibility of noises and perceive the language, and other features and by applying basic method. Cepstral coefficient is a value obtained by taking the Inverse Fourier Transform of the logarithmic value extracted from the spectrum of the audio signal.

D. Classification

The algorithm used here is Multi-layer perceptron, commonly known as MLP which is a feed forward mechanism. The two important fundamentals in MLP are weight and bias. The equation of MLP is given below[4]:



Equation 1: MLP Function

The audio samples are processed and is classified according to different classes of BMI as mentioned in the table earlier. Also equation for BMI is mentioned below [5]:

Imperial English BMI Formula

weight (lbs) x 703 ÷ height (in^2)

Metric BMI Formula

weight (kg) / height (m^2)

Equation 2: BMI Function

5. NETWORK ARCHITECTURE

Multi-layer perceptron(MLP) is a perceptron with one or more layers. It comprises of input layer, hidden layer and out- put layer as described in the following Network Architecture of MLP. Each neuron compute an activation function based on which the optimizer decides if to use the value or not. To build this network Python 3.6.5 was used. Keras Version 2.1.5 and Tenserflow-GPU version 1.8.0 was used as the backend for Keras library. To run Tenserflow on GPU CUDA version 9.0 and CUDnn version 7.6.4 which was acquired from NVDIA's official website. For training NVDIA GTX 1070 was used. A sequential model with five hidden layers was made. It took around 25-30 seconds to train on the entire dataset.

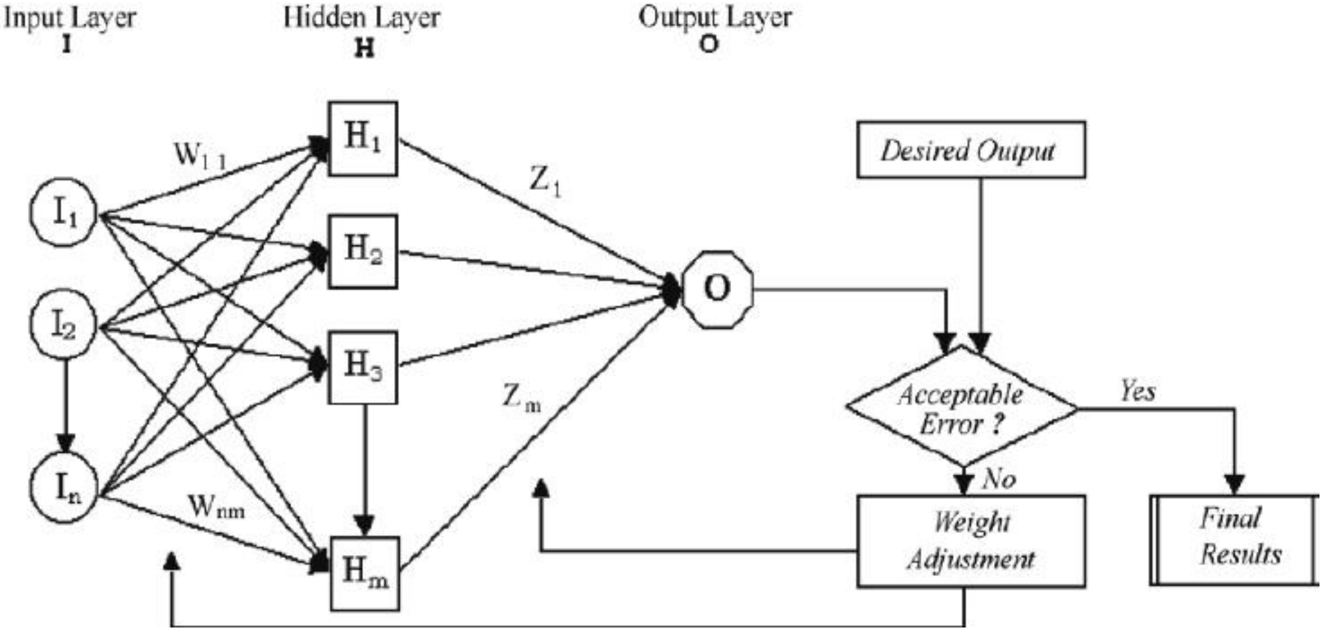


Fig. 4. MLP Network Architecture [9]

A. Input Layer

The main job of this layer is to introduce the input values into the network. In this layer layer, no activation function is present and also processing is not done. The number of MFCC features per audio file is 30. So, the total number of input neurons will also be 30.

B. Hidden Layer

This layer is sandwiched between input and output layer.Also it takes in a set of weighted inputs and produce an output through an activation function. The activation function used is ReLu.

C. Output Layer

It performs similarly to that of hidden layer .In this layer outputs are passed on to the world. The activation function used is Softmax.

During training and feature extraction epochs is set to 50, batch size is set to 5, loss is categorical crossentropy, optimizer used nadam. Also, dropout is kept at 30 percent to avoid overfitting. The learning rate is 0.01 and seed value is 100. For efficient training purpose the original training file was called from a different file to keep the CPU and memory usage at minimum.

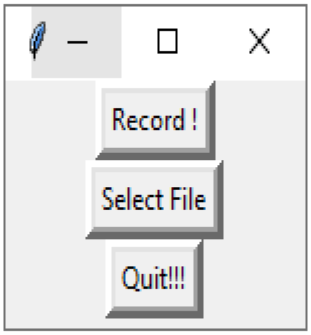
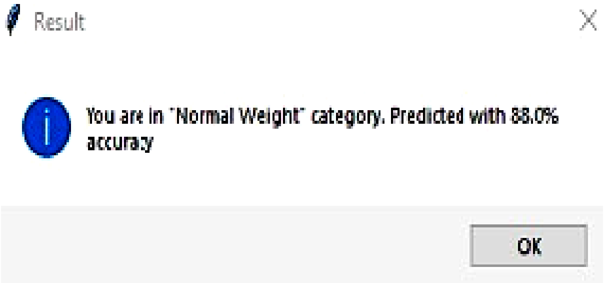


Fig. 5. Using Tensorflow GPU for loading the model

6. RESULT

Recall accuracy of 88 percent was observed with the dataset of 250 audio samples. The accuracy is less due to less number of audios present in the dataset.The algorithm which was used is Multi-layered perceptron because algorithms- K nearest neighbour and probabilistic neural network were already covered.



You are in "Normal Weight" category. Predicted with 88.0% accuracy

Fig. 6. BMI Calculated

7. CONCLUSION

This paper presented a more light weight approach than KNN and PNN with the help of multi layer perceptron. MFCC feature was calculated for every audio file which was provided as an input to the MLP. With 5 hidden layers there are a total of 7 layers in the network architecture with the number of neurons in the input layer being 30. The accuracy of the model was around 88 percent which was obtained by training the model on around 250 audio samples.

Future Scope

Due to the less audio samples used the recall accuracy is less which is 88 percent. Hence, in order to increase the accuracy one can collect more audio samples. Here we have considered only one sound feature "MFCC" but in order to get more accuracy in future one can extract more sound features such as Mel, contrast, Tonnetz, spectrogram.

ACKNOWLEDGMENT

The authors admiringly endorse the contribution of Depart- ment of Information Technology, SIES Graduate school of Technology for their sincere guidance. We would also like to acknowledge the contribution and support of the family members of the authors.

REFERENCES

[1] C. Berkai, M. Hariharan, S. Yaacob, M. I. Omar, " Estimation of BMI status via speech signals using short term cepstral features", IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 195-199, 2015.

[2] C. Berkai, M. Hariharan, S. Yaacob, "Prediction of body mass index using speech signals: A review", 2nd International Conference on Biomedical Engineering (ICoBE), Penang, pp. 1-6, 2015.

[3] C. Tai, D. Lin, "A Framework for Healthcare Everywhere: BMI Prediction Using Kinect and Data Mining Techniques on Mobiles", 16th IEEE International Conference on Mobile Data Management, Pittsburgh, pp. 126-129, doi: 10.1109/MDM.2015.40, 2015.

[4] C. Yadav, N. Razavian, "Using Brain MRI Images to Predict Memory", BMI Age, IEEE International Conference on Humanized Computing and Communication (HCC), Laguna Hills, CA, USA, pp. 126-128, 2019.

[5] H. A. Rahman, S. N. M. Rameli, R. S. S. Abd Kadir, Z. H. Murat and M. N. Taib, "Analysis of correlation between BMI and human physical condition using resonant field imaging system (RFI)", IEEE International RF and Microwave Conference, Kuala Lumpur, pp. 279-282, 2008.

[6] R. M. Simply, E. Dafna and Y. Zigel, "Diagnosis of Obstructive Sleep Apnea Using Speech Signals From Awake Subjects", IEEE Journal of Selected Topics in Signal Processing, vol. 14(2), pp. 251-260, 2020.

[7] G. N. Jayabhavani and N. R. Raajan, "Silent speech generation using Brain Machine Interface, IEEE International Conference on Emerging Trends in Computing", Communication and Nanotechnology (ICECCN), pp. 630-633, 2013.

[8] M. Abdollahian, "The Impact of Body Mass Index on Low Birth Weight", 10th International Conference on Information Technology: New Generations, Las Vegas, NV, pp. 567-572, 2013.

[9] https://www.intechopen.com/books/nuclear-power-system-simulations-and-operation/a-literature-survey-of-neutronics-and-thermal-hydraulics-codes-for-investigating-reactor-core-parame

Wireless Landslide Detection and IOT Based Alarming System

Anjali Jose[1], Lin Maria K Thomas [2], Agna Shaju P [3], Alvin Subash [4], Silpa PA[5]

[1, 2,3,4]Scholar, Sahrdaya College of Engineering and Technology, Kodakara, Kerala, India.
[5]Assistant Professor, Sahrdaya College of Engineering and Technology, Kodakara, Kerala, India.
Email: [1]anjalijosepaul@gmail.com, [2]linmaria98@gmail.com, [3]agnashaju99@gmail.com,
[4]alvincs74@gmail.com, [5]silpapa@sahrdaya.ac.in

Abstract - Extreme rainfall in August 9, 2019 in Kerala triggered major landslides in many districts like Wayanad and Malappuram, it took numerous lives and caused extensive damage to property. This research work focuses to detect landslides and to alarm people through IOT network. The detection system consists of Wireless Sensor Network with sensors like Geophone, Moisture sensor, accelerometers along with a microcontroller. Geophone senses the land movement and the moisture sensor detects the amount of water seepage through the land. This network is buried inside the soil in landslide prone areas. The data collected in the microcontroller is sent to the IOT platform and is analyzed with reference to the threshold values to detect the landslide. On the onset of landslides people are alerted through instant messaging.

Keywords – Landslide, Wireless Sensor Network, Geophone.

1. INTRODUCTION

The world has experienced many natural and man-made disasters which have destroyed large communities and resulted in a great loss of human life. Landslides are one of the most catastrophic disasters that happen around the world. The occurrence of landslides can be related to several causes like morphological, physical and geological effects as well as human activities. Landslide is basically the down-slope movement of soil, rock and other materials due to the influence of gravity. These movements occur suddenly and are a short lived phenomenon that causes remarkable landscape changes. In India, landslides mainly occur due to the intense rainfall. Earthquakes can also cause landslides, however in India these are confined to the Himalayan belt. High rainfall intensity accelerates the slumping and sliding in the hazard zones. The annual loss that occurs due to landslides in India is about $420 million. According to the survey conducted after the floods of 2018 in Kerala, 109 out of the 350 died due to landslide alone. This statistic alone explains how much impact the disaster has on the human lives. The lack of real time monitoring system and improper alarming system in many parts of the disaster hit areas increased the impacts of the landslides in the area. Our research work aims to develop an efficient mechanism to detect and make the people; concerned authorities alert the possibility of occurrence of the disaster. Wireless Sensor Networks (WSNs) is a technology that has the potential for carrying out environmental monitoring. This research work involves interdisciplinary areas such as geology, hydrology, soil mechanics, landslide studies, networking and wireless sensor networks.

2. WIRELESS SENSOR NETWORKS

Wireless Sensor Networks (WSNs) are a recent technological advancement that has immense potential for environmental monitoring as well as monitoring of critical and emergency applications. However, it has limitations such as low memory capacity, power and bandwidth. WSN's has an inherent capability to be deployed in a hostile environment and its low maintenance requirement make it ideally suited for real-time monitoring. These networks are applicable in areas such as environmental monitoring, defense monitoring, Infrastructure monitoring, building monitoring and health care.

2.1 Hardware and Software

The wireless sensor network devices that integrate sensors, processor and wireless networking capabilities are commonly referred to as Motes. The implanted sensors in these nodes will sense the environmental changes and send the details to their neighboring nodes using wireless technology. Different types of wireless nodes with varying capacity of processing, storing, and power have been developed. These special nodes act as the aggregators, gateways, etc. Since all available wireless sensor nodes are constrained by energy and memory, different hardware and software solutions have evolved to deal with these constraints, thus making the technology more useful. Here we use Arduino Uno as the controller board, which is based on Atmega328p for the sensor nodes. It has a total of 14 digital input/output pins, 6 analog input pins, a 16 MHz quartz crystal element, a USB connection for interfacing, a power jack to supply power, an ICSP header and a reset button. It is small, flexible and compatible. It was developed by arduino.cc in Italy. This operates at 5 volts.

3. LITERATURE SURVEY

India has been affected by hydro-geological hazards like landslides since years ago. History depicts that, it has happened mostly in Himalayan regions, this make the landslide monitoring as an important aspects and challenges for the geologists in India. Most of the landslides in India are caused due to heavy rainfall [1]. Rainfall events cause slope failures in areas of certain extent or in large regions. To develop an early warning system for the monitoring of landslides requires the domain expertise, not just to build the instruments but to employ them properly and decode their output for rational purposes. Real time monitoring of landslides is an exigent research area present today in the field of geophysical research. This paper tries to discuss the development of an on field deployment of wireless sensor network based landslide detection system [2]. This system includes a heterogeneous network which is mainly composed of Wi-Fi, wireless sensor nodes, and satellite terminals for the consistent delivery of real time data to the data management center, in order to enable the analysis of data acquired and to give away landslide warnings and risk assessment reports based on the data acquired to the inhabitants of the region. The node that was deployed has an advantage of obtaining optimum results. Moreover, [3] this node has minimum cost and more compatibility for expanding along with other communication devices for more fast responses

4. DESIGN AND DEVELOPMENT

The important parameters for monitoring the rainfall induced landslides are water pressure due to the seepage of water on the soil and the slope movement. Dielectric moisture sensors are employed for this purpose. Accelerometer, geophones and pressure sensors are used for the measurement of the other parameters, based on their relevance in finding the geological factors that cause the landslides under heavy rainfall conditions. Geophones detect ground movements and convert them into voltage. Accelerometers measure acceleration in x, y and z axis, measured values appear as change in voltages. It determines movement of the sensor column which is buried deep inside the earth surface. In order to measure the pressure exerted by the soil movement on the sensor column we use force sensor. Into Combination of these sensors can be used for detecting landslides. The proposed solution is represented using a block diagram in the Figure 1. The diagram shows the three stages of the proposed system.

- Data Collection from different Sensors
- Analysis of data in the controller
- Sending the data to the IOT based server
- Generating output messages and sending instant messages via telegram

The model of the sensor column is depicted here in the Fig.2. Geophone and Moisture sensors are fixed on to a sensor column and the microcontroller board along with the accelerometer is placed on top. The data collected from the sensors are transmitted wirelessly to the Thingspeak server using the node MCU. The Node MCU is connected through the WiFi and data is sent to Thingspeak. The data is stored in different fields allocated to each parameter and is displayed to the user as graphs. The correlation and comparison of the data is done in the Thingspeak.

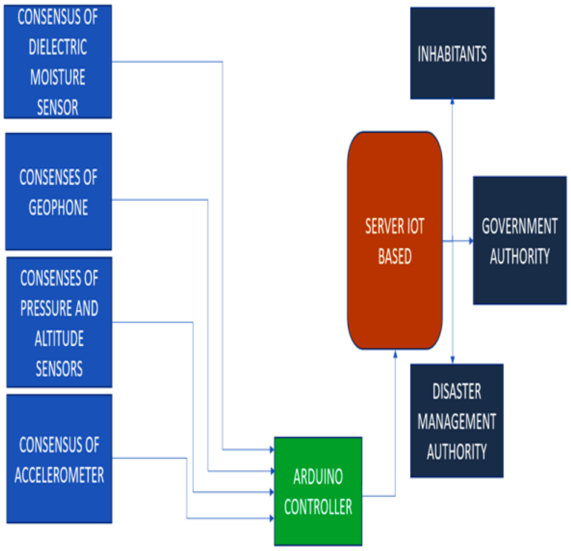

CONSENSUS OF DIELECTRIC MOISTURE SENSOR
CONSENSES OF GEOPHONE
CONSENSES OF PRESSURE AND ALTITUDE SENSORS
CONSENSUS OF ACCELEROMETER
INHABITANTS
SERVER IOT BASED
GOVERNMENT AUTHORITY
DISASTER MANAGEMENT AUTHORITY
ARDUINO CONTROLLER

Fig.1 Block Diagram of the solution

The data is compared with the threshold which was previously set by the researchers analyzing different conditions of the environment. The alert signal is given to the concerned authority using the Matlab-Telegarm channel. Based on the alert received different arrangements can be made in order to help save the lives of many.

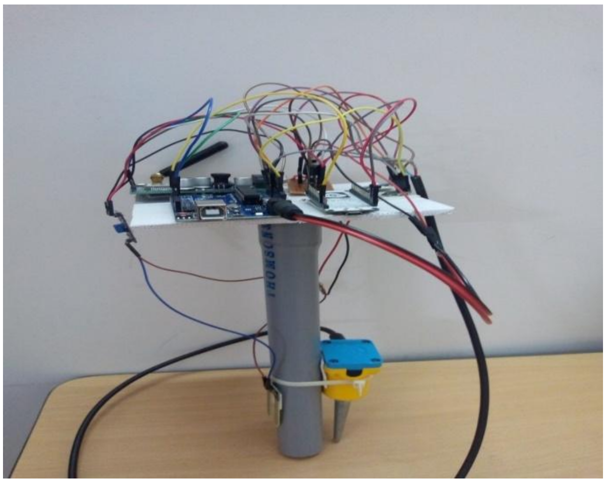


Fig.2 Sensor Column Design

5. EXPERIMENTAL SETUP AND CIRCUIT DIAGRAM

5.1 Experimental Setup

Landslide theory has not yet advanced sufficiently to make the accurate predictions of landslides or to model the phenomena in much depth the way we need them. We have designed and developed an experimental setup to simulate and test the working of the system [4]. The experimental setup provides a test bed for developing, testing, and calibrating the sensors and subsystems of the wireless sensor network [5]. The soil is packed into the setup along with the sensor column for the experiment. Water is then added in the form of seepage, until the slope fails.

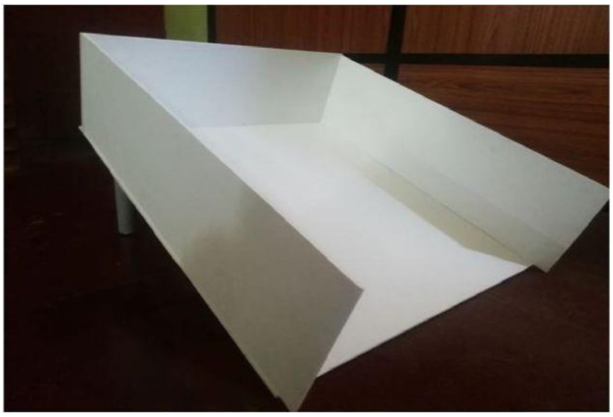


Fig.3 Experimental setup using Foam Sheet

5.2 Circuit Diagram

The Proteus Design is a software tool suite used mainly for the automation of design in electronics field. Arduino Uno is the controller. Sensors are represented as pot-HG. These are connected to analog pins of the Arduino Uno board. Here virtual monitor is used to observe the output. Power supply and ground are separately given to each sensor [6].

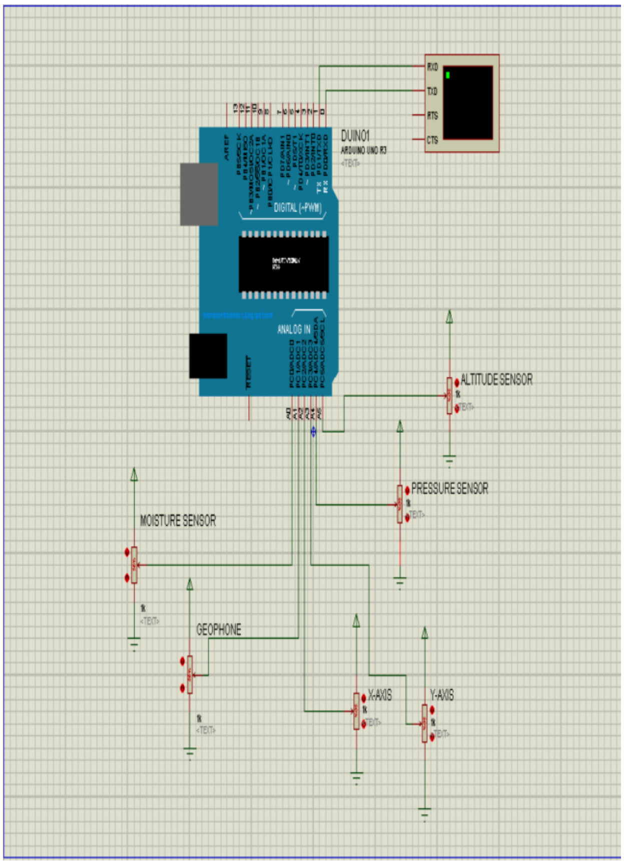


Fig.4 Circuit Diagram using Proteus Design Suite

6. RESULTS AND DISCUSSION

Rainfall induced landslides are among the leading natural disasters experienced in India and all over the world. Though different landslide monitoring techniques have been developed, real-time monitoring of natural disasters like landslides is still one of the current challenges in the field of geophysical research. The developed system aims to provide landslide warnings and risk assessments to the inhabitants of the deployed region. The data was collected by deploying the sensor column with sensors into the experimental setup. The data collected was sent to the Thing speak channel using the Node MCU attached to the Arduino controller. The data sent to the Thingspeak channel can be viewed using the channel.

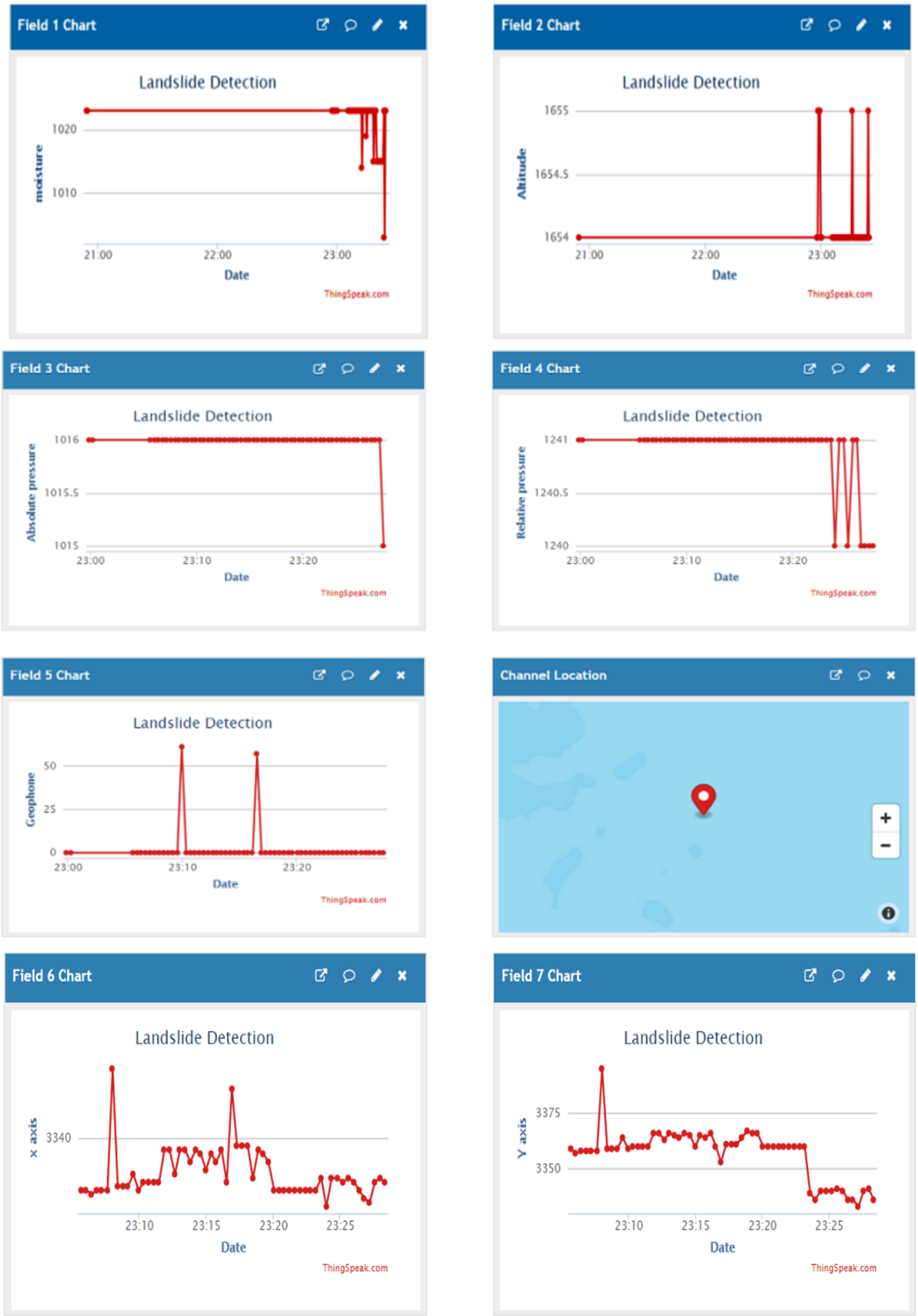
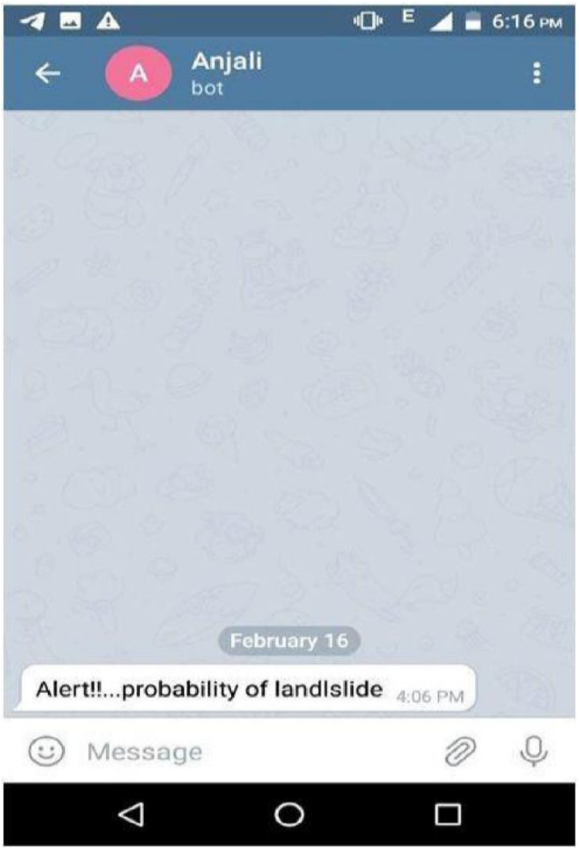


Fig. 5 Data collected from sensors represented as graphs in Thingspeak

The data from different nodes are analyzed using the mat lab and the alert signal is sent via telegram to the users and disaster management authority. The sensors and the column are integrated to the IOT network and the alert is provided instantaneously at a faster rate. Government can evacuate the people. The analysis of different factors was done and the relevant information was collected to execute the solution in landslide prone areas. The software part of the solution has been successfully validated and relevant changes are made to it in order to make the functioning more suitable.



Alert!!...probability of landlslide 4:06 PM

Fig. 6 Alert received by the user through the telegram

7. CONCLUSION AND FUTURE SCOPE

Rainfall induced landslides are among the most catastrophic natural disasters. The Real time monitoring of landslide is still one of the challenges in the field of geophysical research. The real-time data received from the deployed network is analyzed using the Matlab software. The landslide modeling program gives the probability of landslide occurrence. The geological sensor data analysis provides the pattern of variation to the climate change. These tools will help to provide a better understanding of the landslide phenomenon and lead to the development of landslide prediction systems. The entire system can be functional during monsoon seasons. The proposed methodology is setup and made available for the future deployment and waiting for the government sanction to start the project. The analysis of different factors was done and the relevant information was collected to execute the solution in landslide prone areas. Validation of software part is completed and relevant changes are made.

In the future, this network will be used to advance the study of wireless sensor networks and their performance in real-time scenarios, in addition to developing methods for predicting landslides and analyzing the impact of sensor density on such predictions. We also hope that this study will lead to more understanding about the capabilities and usability of wireless sensor networks for other critical and emergency applications. There are varieties of methods that are applicable for landslide detection using various Algorithms such as Distributed algorithm, use of geophone network, geodetic monitoring, etc. which are considered to be essential. Our aim here is to design an effective hardware and software system to predict the landslide real time and monitor them all the time. We would like to extend the project to design the hardware system by integrating with receivers to update the location information. Collection of data from the landslide prone areas for future reference.

ACKNOWLEDGEMENT

We acknowledge with great gratitude all the faculties of Sahrdaya College of Engineering and Technology for giving us constant support and inspirations throughout the completion of the research work.

REFERENCE

[1] Thampi. P. K., Mathai. John., Shankar. G., Sidharthan. S., “Landslides: Causes, Control and Mitigation”, (based on the investigations carried out by the Centre for Earth Science Studies, Trivandrum), 1984.

[2] Ramesh, M. V., “Real-time Wireless Sensor Network for Landslide Detection”, Proceedings of The Third International Conference on Sensor Technologies and Applications, SENSORCOMM 2009, IEEE, Greece, June18 - 23, 2009.

[3] LAN. Hengxing., ZHOU. Chenghu1., C. F. Lee., WANG .Sijing., WU. Faquan., Rainfall- induced landslide stability analysis in response to transient pore pressure – A case study of natural terrain landslide in Hong Kong.

[4] Andreas Gunther, Miet Van-Den Eeckhaut, Jean Philippe Malet, Paola Reichenbacjh and Javier Hervas, Climate physiographically differentiates Pan-European landslide susceptibility assessment using spatial multi-criteria evalution and transnational landslide in-formation Geographology, 224:69-85, November 2015.

[5] D. Petly, Global patterns of loss of life from landslide, Geology, vol.40, no. 10, p.927, 2014.

[6] M. V. Ramesh, Design, development, and deployment of a wireless sensor network for detection of landslides, Ad Hoc Networks, vol. 13, pp. 218, Feb. 2014.

Smart Traffic Management System using Image Processing

Jerome K[1], Manoj S Nagendra[2], Chandrasekar A[3]

[1,2]Scholar, Department of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, India
[3]Professor & HOD, Department of Computer Science and Engineering, St. Joseph's College of Engineering, Chennai, India
Email: [1]ivanjerome99@gmail.com, [2]manojnagendra@gmail.com, [3]drchandrucse@gmail.com

Abstract - The objective of this research work is to develop a traffic management system that automates the current process which is carried out manually. It is divided into 3 modules - simulation to determine the traffic density, detect a crash/accident, detect ambulance using image processing and machine learning techniques. In the first phase, we determined traffic density to minimize the delay caused by traffic congestion and to provide the smooth flow of vehicles. The density of vehicles on each side can be identified by using datasets. If the density is low on a particular side, the period for that side is normal and if the density is high the period will automatically increase compared to normal density. In the second phase, we simulated a crash or accident detection and for the prototype consideration, we have used static accidental image and trained model. In the third phase, analyzed ambulance detection using the dataset, for the prototype consideration for this, used static ambulance image and trained dataset. On detection of an ambulance, the traffic light is automatically changed to green. In each phase, the data updating and monitoring are provided. This scheme is fully automated and identifies the emergency vehicle and controls the traffic lights dynamically. All of these processes are carried out with the help of image processing.

Keywords - Vehicle Density Calculation, Crash Alert, Ambulance Detection, Background Subtraction, Edge Detection, Template Matching.

1. INTRODUCTION

Image processing is a method to convert an image into digital form and perform the necessary operations on it, to get an enhanced image and information. It is a type of signal dispensation in which input is an image, like a video frame or photograph or characteristics associated with that image. The image processing system includes considering images as two-dimensional signals while applying already set signal processing methods to them. It is among rapidly growing technology, with its applications in different aspects of a business. Image Processing forms the core research area within engineering and computer science [1]. Image processing includes Importing the image with an optical scanner or by photography, Analyzing and manipulating the image which incorporates data compression, image enhancement and spotting patterns that are not visible to human eyes like satellite photographs, and the output is obtained in the last stage in which the result can either be an altered image or a report that is based on image analysis.

2. LITERATURE REVIEW

The Association for Safe International Road Travel (ASIRT) reports, annually approx.1.3 million people fatality on the road, 20-50 million of road users are incapacitated. Road crashes cost 1-2% of their annual GDP in different countries. Currently, Road traffic crashes rank as the ninth leading cause of death and account for 2.2% of all fatality globally. If it is not mitigated, road traffic hurts are predicted to become the fifth by 2030 [1]. The challenges imposed on local public servicing outsourcing in saving human lives resulting from vehicle accidents have become a critical concern. An automated and intelligent mobile solution is required for a zero mortality rate since there is a lack of automated on-site medical assistance, late accident reporting, inaccurate geographic location, and lack of injured medical information [2]. The current existing solutions that assist passengers in case of a vehicle accident are concerned with user interaction after the incident. Those mobile solutions require that the injured must launch the app and request help manually and that would not be possible if he/she is under the critical or serious non-vital situation. The situation becomes even worse if passengers go under an unconscious state. Traffic lights which are of current technology use a manual operating system for the time allocation and also require high maintenance during the operation. This causes, time lapsing, and an increase in vehicular traffic [3]. In the existing system, the traffic congestion is predicted manually which is hectic and involves manual efforts. Similarly, accident detection is predicted manually and doesn't ensure quick first-aid. On the whole, in the existing system, the traffic management system is manual and not automated, right from traffic lights, accident detection, emergency vehicle detection, and regulating the traffic which isn't as efficient as the automated system. This proposed system i.e. Smart Traffic Management System dynamically can change the signal lights based on the traffic density [4], detect a crash

[5], and detect emergency vehicles and regulate the traffic accordingly [6].

3. PROPOSED SYSTEM

The objective of the proposed research work is to develop a simulator to determine the traffic density, ambulance, and accident incidents using image processing and machine learning techniques.

In the first phase, we determined traffic density to minimize the delay caused by traffic congestion and to provide the smooth flow of vehicles.

The density of vehicles on each side can be identified by using datasets. If the density is low on a particular side, the time for that side is normal and if the density is high the time will automatically increase compared to normal density. The second phase work simulates a crash or accident detection and for the prototype consideration, used static accidental image and trained model. During the third phase, analyzed the ambulance detection using a dataset, for the prototype consideration used static ambulance image and trained dataset. On detection of an ambulance, the traffic light is automatically changed to green.

3.1 Traffic density identification

A. Video processing to frame

Video processing is a subcategory of Digital Signal Processing techniques where the input and output signals are video streams. In computers, one of the simplest ways to succeed in video analysis goals is using image processing methods in each video frame. In this case, motions are simply realized by comparing sequential frames. Video processing includes pre-filters, which may cause contrast changes and noise elimination alongside video frames pixel size conversions. Highlighting particular areas of videos, deleting unsuitable lighting effects, eliminating camera motions, and removing edge-artifacts are performable using video processing methods. OpenCV library of python is provided with functions that allow us to control videos and pictures.

B. RGB to Grayscale Conversion

In video analysis, converting RGB color image to grayscale mode is completed by image processing methods. The main goal of this conversion is that processing the grayscale images can provide more acceptable leads to comparison to the first RGB images. In video processing techniques the sequence of captured video frames should be transformed from RGB color mode to a 0 to 255 gray level. When converting an RGB image to a grayscale mode, the RGB values for every pixel should be taken, and one value reflecting the brightness percentage of that pixel should be prepared as an output.

C. *Canny Edge Detection*

Object detection is often performed using image matching functions and edge detection. Edges are points in digital images at which image brightness or gray levels change suddenly in amount. The main task of edge detection is locating all pixels of the image that correspond to the sides of the objects seen within the image. Among different edge detection methodologies, the Canny algorithm may be a simple and powerful edge detection method. Since edge detection is vulnerable to noise within the image, the initiative is to get rid of the noise within the image with a 5x5 Gaussian filter.

D. Kalman Filter to detect BLOB

BLOB detection would contain BLOB detection, BLOB analysis, and BLOB tracking. Morphological closure operations including erosion and dilation would be used to increase the accuracy of the system for better detection of vehicles. Further region properties of BLOBs are calculated. Various properties are taken into considerations. These properties help a lot in analyzing the area of the BLOB. Area calculation would give us the density on the roads as low density, medium density, and high density depending on the area covered by the BLOBs.

E. Vehicle Accident Detection

Once the vehicle is detected from the annotated image, with help of the number of pixels that the rear-width of the vehicle occupies in the image and the pre-calibrated distance and number of pixels occupied, the distance between the vehicle at the front and the other cars is determined. With the help of finding the change in distance after a short interval of time, the relative velocity in which the vehicle is moving could be determined. And with knowledge of this and the average speed in which each car moves in each region, the absolute velocity of the vehicle moving also can be determined. When there is a big difference possibility of an accident is detected and each frame is checked for vehicle accidents. Once we found the velocity of the vehicle at the front, we could easily understand the relative speed patterns as we know the speed in which each car moves in a specific region.

With the relative speed that keeps on changes over time, the collision could be identified automatically using machine learning patterns.

Procedure for Traffic Density Calculation

Step1: Traffic conjunction
Step2: Pre processing
Step3: Canny edge detection
Step4: Calculation of region properties
Step5: Calman filter to detect blob
Step6: Classification
Step7: Traffic Density

F. Ambulance Detection

The ripple algorithm is proposed and aims to detect the Ambulance from images obtained by a clear image. It can extract features of the target which are almost invariant when image rotations or target translation and scaling, such that it can detect targets. Moreover, by the templates representing targets, it also can detect the target using machine learning models which we have used to train the system.

Procedure for Accident Detection and Ambulance Detection:

Step1: Traffic conjunction
Step2: Pre processing
Step3: Vehicle distance calculation
Step4: Average velocity, speed predicted for regions
Step5: Classification

Template Matching is a high-level machine vision technique that identifies the parts on a picture that match a predefined template. The Advanced template matching algorithms, Convolutional Neural Network allow us to find occurrences of the template regardless of their orientation and local brightness. The Template Matching techniques are flexible and relatively straightforward to use which makes them one of the most popular methods of object localization. Applicability is restricted mostly by the available computational power as the identification of massive and sophisticated templates is often time-consuming.

4. METHODOLOGY

The research work has been planned to address the objectives of determining traffic density to minimize the delay caused by traffic congestion and to provide the smooth flow of vehicles. Besides, the work simulates a crash or an accident detection and for the prototype consideration, used static accidental image and trained model, and analyzed the ambulance detection using the dataset, for the prototype consideration used static ambulance image and trained dataset. On detection of an ambulance, the traffic light is automatically changed to green. The objectives were screened with literature reviews and identified the research work topic and its significance to the current context of the vehicle movement management to provide a safe drive and safe life.

The methodology is as follows:

Chosen the right images and datasets for training model, captured the density of vehicles in a particular frame, counted the number of vehicles in that frame and based on the density, changed the traffic light colors, used a sample video of an accident and set alert notification upon a crash, and lastly in the third part, used an image of an ambulance which was captured by the webcam and detected it as emergency service accordingly. The input dataset images have been extracted; the background subtraction technique would be applied to each of these frames. This process would lead us to urge a binary image from the frame is processed for BLOB detection. BLOB detection would contain BLOB detection, BLOB analysis, and BLOB tracking. Morphological closure operations including erosion and dilation would be wont to increase the accuracy of the system for better detection of vehicles. Further region properties of BLOBs are calculated.

Various properties are taken into considerations. These properties help a lot in analyzing the area of a BLOB. Area calculation would give us the density on the roads as low density, medium density, and high density depending on the area covered by the BLOBs. These classifications help us in better analysis of the traffic conditions, ambulance vehicle detection, and accidents of vehicle detection.

- **Hardware Interface**

All the physical equipment's i.e. input devices, processor, and output device and interconnecting processor of the computer are called hardware.

- Hard Disk minimum of 40 GB.
- RAM minimum of 2 GB.
- Dual Core and up to 15" Monitor.
- Integrated webcam or external webcam (15 -20 fps).

- **Software Interface**

A set of instructions or programs required to make a hardware platform suitable for the desired task is known as software. The software also can be defined because of the utility programs that are required to drive hardware of the pc.

- Operating system- Microsoft Windows 7 SP 1 or above; Python

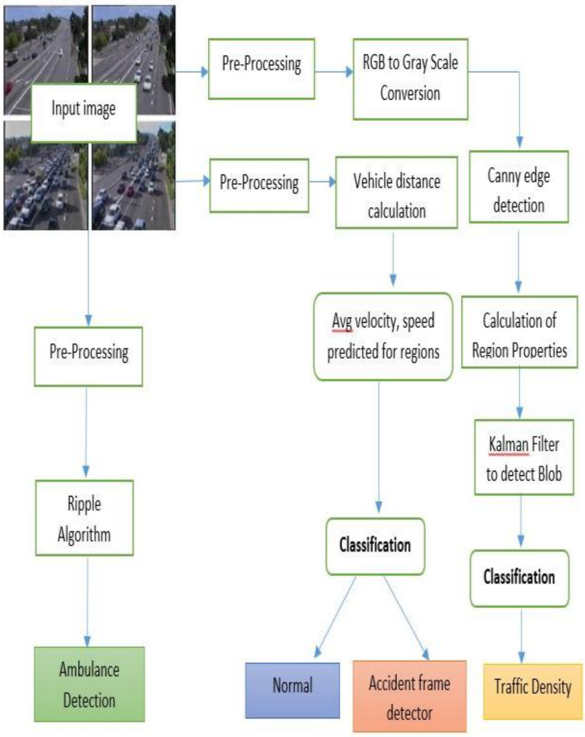


Fig. 4. The architecture of the traffic management system

5. EXPERIMENTAL RESULTS

The following figures are the screenshots of the working module captured in an environment.

Dashboard (Figure 5) is created using Python modules and the Graphical User Interface includes three executable buttons that trigger the respective modules i.e., View Traffic, Find Ambulance, and View Accident in a closed software environment.

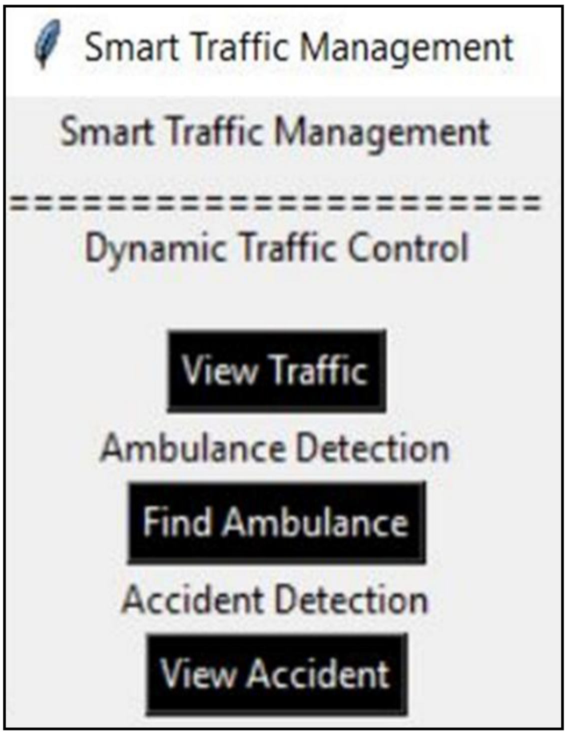


Fig. 5. Dashboard

In the accident detection module, the crash simulation (Figure 6) is done and the video is scanned and divided into frames. The crash alert is displayed at a particular frame (figure 7) to notify upon a crash.

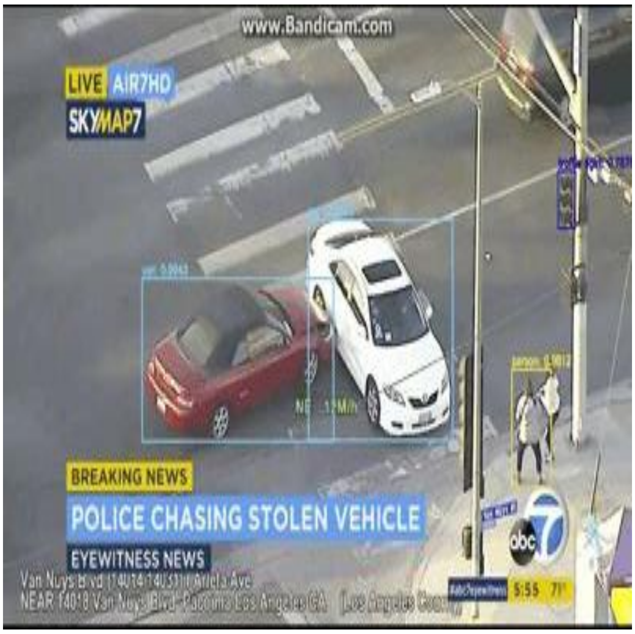
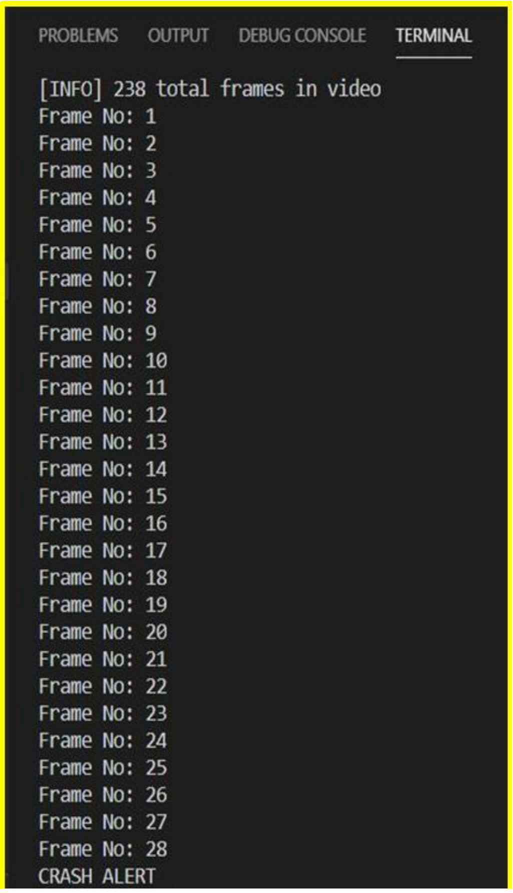


Fig. 6. Vehicle crash simulation



```
[INFO] 238 total frames in video
Frame No: 1
Frame No: 2
Frame No: 3
Frame No: 4
Frame No: 5
Frame No: 6
Frame No: 7
Frame No: 8
Frame No: 9
Frame No: 10
Frame No: 11
Frame No: 12
Frame No: 13
Frame No: 14
Frame No: 15
Frame No: 16
Frame No: 17
Frame No: 18
Frame No: 19
Frame No: 20
Frame No: 21
Frame No: 22
Frame No: 23
Frame No: 24
Frame No: 25
Frame No: 26
Frame No: 27
Frame No: 28
CRASH ALERT
```

Fig. 7. Accident Detection Output

In ambulance detection, the image of an ambulance is captured by the web camera (figure 8) and upon detection, the result is displayed (fig. 9)

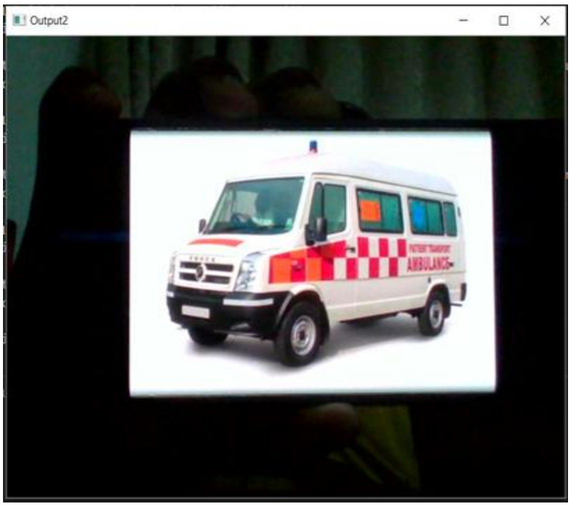
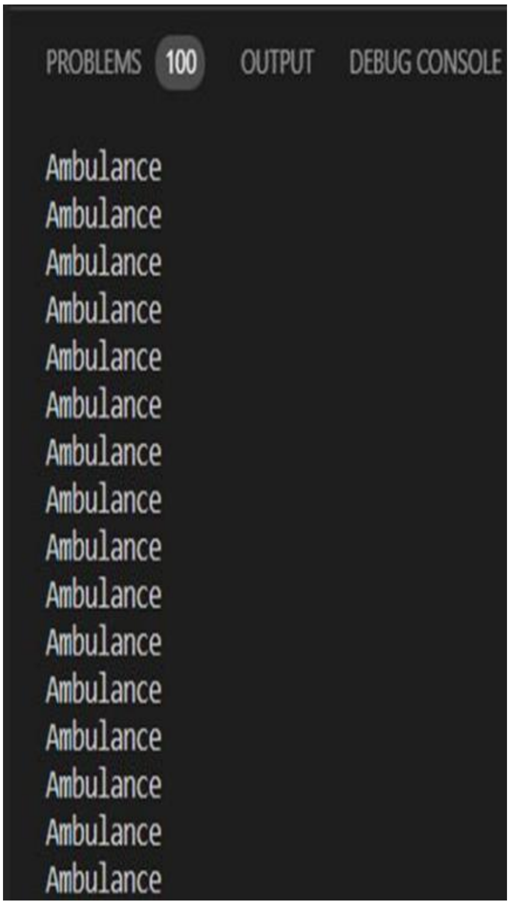


Fig. 8. A static image of an ambulance captured by the camera



```
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
Ambulance
```

Fig.9. Ambulance Detection Output

In traffic density calculation, the number of vehicles are counted (figure 10), divided into frames, and the count is displayed in the result (figure 11).

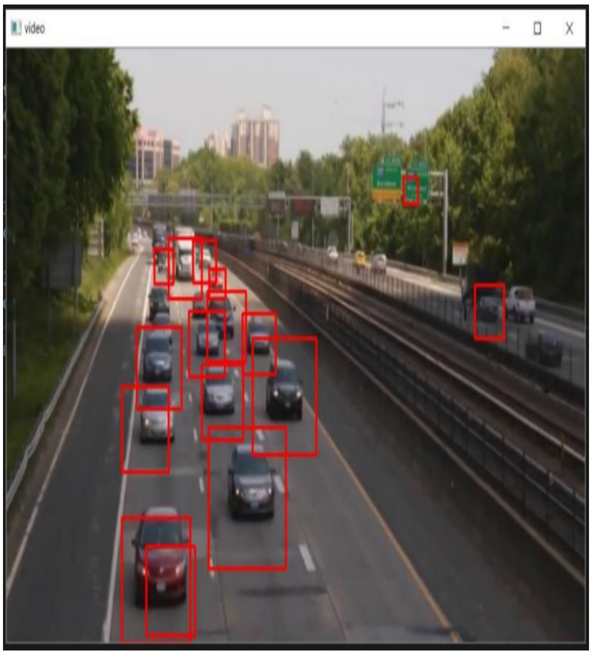
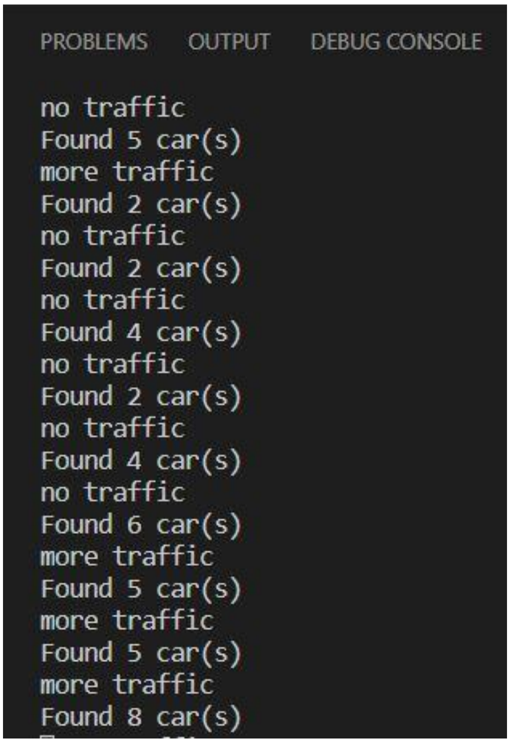


Fig. 10. Count of the number of vehicles



```
no traffic
Found 5 car(s)
more traffic
Found 2 car(s)
no traffic
Found 2 car(s)
no traffic
Found 4 car(s)
no traffic
Found 2 car(s)
no traffic
Found 4 car(s)
no traffic
Found 6 car(s)
more traffic
Found 5 car(s)
more traffic
Found 5 car(s)
more traffic
Found 8 car(s)
```

Fig. 11. Traffic Density Output

6. CONCLUSION

Traffic Density Analysis, Ambulance, and Accident detection System Using Image Processing has been discussed in this proposed system. The primary accomplishment of this system is the automation of traffic management and more specifically, regulation of vehicles in traffic, reducing fatality/mortality by timely detection of accidents, and prioritizing emergency services in a jammed situation. Further, this research work provides a framework that analyses the dataset input images and periodic frames would help to increase the processing speed of the framework. BLOBS increase the efficiency and improve the detection as well as analysis of vehicles. The framework to automatically classify traffic, ambulance vehicle, and accidents in the roads using image processing and machine learning techniques is one of the most successful topic models.

- **Limitations**

The limitations of this simulated model are minimal. Firstly, the detection of vehicles approaching in both directions simultaneously doesn't offer the best result. Detection of vehicles on a single side produces accurate results. Secondly, in accident detection, when two vehicles come together nearby, the crash alert could be triggered sometimes when there is no crash. Thirdly, in ambulance detection, the existence of patients inside the ambulance cannot be identified by this and can be misused for false-emergency by the drivers. There is an enormous scope of image processing in traffic monitoring and analysis for future technologies.

- **Future Enhancement**

In the future, to this model, helmet detection and automated fine system can be added. The camera detects the person not wearing a helmet and captures the vehicle number. Then, the details of the rider are retrieved and a fine is imposed for the violation. Further, as an extension to the accident detection, SOS alert upon detection of a crash can be added. Further advancements in video-based traffic-flow detection can help in developing increasingly robust, real-time, and intelligent traffic management in an optimized system structure.

REFERENCES

[1] Kaviani, Razie, Parvin Ahmadi, and Iman Gholampour, A new method for traffic density estimation based on topic model, Signal Processing and Intelligent Systems Conference (SPIS), IEEE, 2015.

[2] Hasan, Md Munir, et al., Smart traffic control system with application of image processing techniques, Informatics, Electronics & Vision (ICIEV), International Conference on. IEEE, 2014.

[3] Kaviani, Razie, Parvin Ahmadi, and Iman Gholampour, A new method for traffic density estimation based on topic model, Signal Processing and Intelligent Systems Conference (SPIS), IEEE, 2015.

[4] Hasan, Md Munir, et al., Smart traffic control system with application of image processing techniques., Informatics, Electronics & Vision (ICIEV), International Conference on. IEEE, 2014.

[5] Surendra Gupte, Osama Masoud, Robert F. K. Martin, and Nikolaos P. Papanikolopoulos, Detection and Classification of Vehicles, IEEE Transactions on Intelligent Transportation Systems, Vol. 3(1), pp.37- 47, March 2002.

[6] Suárez, P.D., Conci, A., de Oliveira Nunes, E., VideoBased Distance Traffic Analysis: Application to Vehicle Tracking and Counting, IEEE CS Journals and Magazines, Vol 13(3), pp. 38- 45, 2011.

Role of Convolutional Neural Network Implication Attacks on Image and Attribute in Social Networks

Amita Mishra[1], A Punitha[2], Venkateshwar Rao Pasam[3]

[1,2]Assistant Professor, Department of Computer Science & Engineering, CMR Engineering College, Hyderabad, India.
[3]Assistant Professor, Department of Computer Science & Engineering, Malla Reddy Engineering College, Hyderabad, India.
Email: [1]amita19092010@gmail.com, [2]punitha26@gmail.com, [3]venkatcse534@gmail.com

Abstract - In fashionable culture, social networks play a fundamental responsibility for on-line users. Conversely, one un-ignorable drawback after the thriving of services in privacy problems. At an equivalent point in time, neural networks are fleetly developed in current years, and square gauge established to be appallingly effective in abstract thought attacks. This paper proposes a verity novel structure for abstract thought attacks in social networks that efficiently integrates and modifies the present progressive Convolution Neural Network (CNN) models. This framework will employment wider pertinent eventualities for abstract thought attacks regardless of whether or not a user includes a legit outline image or not. Likewise, the structure is ready to spice up the recent supercilious accuracy CNN for susceptible data prediction. Additionally this paper additionally analyzes and elaborates arrangement of totally Connected Neural Networks to deal with abstract attacks. Furthermore ancient machine learning algorithms square measure enforced to check the outcome from FCNN. Additional privacy concerns are discussed in this paper. Compared to existing approaches the discussed one gives improved results.

Keywords - FCNN- Convolutional Neural Networks completely Connected Neural Networks, Attack, CNN- Convolutional Neural Networks

1. INTRODUCTION

These days, on-line informal communities zone unit a crucial half for everyone. In the US, the amount of your time that people pay on-line interpersonal organizations is consistently expanding; half-hour ever spent online is at present distributed to informal organization cooperation. Social organize stages constantly advance their devices and decisions to attract and have association new audiences.

Their territory unit a great many messages be-ing sent a day in on-line interpersonal organizations. In any case, one un-ignore capable downside behind the blasting of the administrations is security issues. When an individual registers a sound record for an informal community, a prfile ought to be made, that verifies that his family, companions and partners zone unit ready to decide himself. The profile incorporates numerous things of knowledge. Various they zone unit compulsory, and a couple of territory unit no mandatory. De-pending on a client's inclination, he should understand a harmony between communicating good information of himself and covering touchy individual information. In any case, by conveying the assaults started by Convolutional Neural Networks completely Connected Neural Networks (FCNNs), or elective AI calculations, onto an outsized amount of available data, the concealed touchy individual information are frequently deduced and clients' security are regularly undermined.

The inspiration of this paper lies in this delicate information assaults have significant effects for each on-line clients and publicists. Understanding the assaults will profoundly encourage creating cautious measures to stop protection run for on-line clients. Their territory unit numerous outcomes once clients' security is uncovered. In the first place, touchy information could encourage foes to re-spread clients' insider facts since current mystery recuperation mechanisms in some cases raise clients' delicate information before causing secret key recuperation joins. Second, on-line client security run may affect clients' disconnected activities. to Illustrate, realizing clients' expounded in-arrangement like name, birthday, and address may help to fashion charge cards or maybe distinguishing proof documents. Third, delicate information will encourage promoters to convey advertisements for focused clients. to Illustrate, knowing a client's age, sexual orientation, and code may uncover the client's modus vivendi, that significantly will expand the air conditioner curacy of promotion conveyance.

The inspiration of this paper lies in this delicate information assaults have significant effects for each on-line clients and publicists. Understanding the assaults will profoundly encourage creating cautious measures to stop protection run for on-line clients. Their territory unit numerous outcomes once clients' security is uncovered. In the first place, touchy information could encourage foes to re-spread clients' insider facts since current mystery recuperation mechanisms in some cases raise clients' delicate information before causing secret key recuperation joins. Second, online client security run may affect clients' disconnected activities. To Illustrate, realizing clients' expounded inarrangement like name, birthday, and address may help to fashion charge cards or maybe distinguishing proof documents. Third, delicate information will encourage promoters to convey advertisements for focused clients. To illustrate, knowing a client's age, sexual orientation, and code may uncover the client's modus vivendi that significantly will expand the air conditioner curacy of promotion conveyance.

2. RELATED WORK

Numerous investigations are done that represent considerable authority in surmising concealed client data sup-ported old AI calculations [1]. These examinations generally fall under 2 classes, with the essential one pack clients into totally various classes by conveying unattended [2], [3]. AI calculations and furthermore the second one utilizing antiquated AI calculations joined with tongue process. In any case, there zone unit 2 disadvantages from these examinations [4]. To begin with, the exhibitions of old AI algorithms on informal communities region unit typically poor, because of the nature of informal communities, there region unit regularly very ten open qualities, and it's commonly hard to seek after a direct connection between the overall population characteristics and furthermore the Target concealed trait [5]. Second, a few examinations use tongue procedure to flavor up the presentation. On the other hand, the intercalary procedure is long and is moreover incapable to broaden the presentation basically on account of the limitation of old AI calculations.

As of late, various investigations [8], [9] are acknowledged to use FCNN based for the most part legitimate intuition assaults in interpersonal organizations. FCNN is picked because of its reasonable at look for in propelled connections between input traits and yield characteristics [10]. This preferred position turns out to be especially important once it's relentless to speak to a relationship by a liner articulation. In any case, existing investigations have 2 major inadequacies [11]. In the first place, these exactions neither unmistakably show the designs of FCNNs nor the detailed characteristics of the contemplated datasets. It's essential to exhibit anyway a FCNN is worked since totally various setups significantly affect the presentation of intelligent deduction at-tacks in informal organizations [12]. Second, none of the examinations look at the exhibitions among FCNNs and antiquated AI calculations. Since FCNN is generally one type of AI calculations, it's necessary, from the world point of view, to convey such reasonably examinations

.

Since the arranged system is that the reconciliation and alteration of the overall models, the way anyway the models square measure coordinated and along these lines the inspirations driving the modification square measure outlined from area three.2 to 3.4. Enormous ground-truth information got the opportunity to be slithered to mentor the neural systems; accordingly area three.5 shows the best approach to gather information from true interpersonal organizations [13]. Segment 3.6 plans to chase the relationships between the client open traits and there-front the objective shrouded characteristic. During this progression, a mama trix is worried to know the attributes of the offered information. Since existing examinations don't show the cautious structure of FCNNs for informal community attacks, segment 3.7 settles this extraordinary issue by elaborating a potential and compelling FCNN for derives once assaults bolstered open client data. Segment 3.8 shows the execution of some ordinary antiquated mama chine learning calculations. [14] Those calculations square measure applied to coordinate the presentation of neural systems. In particular, call tree, Na¨ıve Thomas Bayes, and k-NN square measure utilized. Cross-approval is one among the impartial ways that to explore the appropriately grouped p.c for AI calculations. In conclusion, Section 3.10 professional vides resistance instruments with exile.

3. METHODOLOGY

A. Integration-of-Faster R-CNN Face-Detector and CNN-Age Classifier

Quicker R-CNN face finder and CNN age classifier upheld pictures square measure develop models. Snappier R-CNN face indicator can do perfect results to separate countenances in an image precisely and expeditiously. CNN age classifier upheld pictures can even succeed pleasant outcomes if pictures containing single face square measure the information. One in everything about most commitment of this paper is that each models square measure coordinated to perform thinking assaults.

There square measure 2 endowments to send this combination. To start with, R-CNN face locator will isolate the profile images that don't contain human appearances. Thus, time is spared to avoid the progressed CNN classifier on pictures and rather, execute light-weight FCNN classifier. Second, the blending will support the exactness of CNN age classifier on pictures. Since the principal classifier is prepared on face pictures exclusively, it per-shapes best once the information is apparent face pictures. Speedier RCNN face indicator ensures that each image that takes care of the CNN classifier contains a transpalease single face. The method of reasoning is that R-CNN face detector has the adaptability to extricate all the face locales in an image and at a proportionate time, dispose of streaky

Foundation and elective articles. At that point, if there's just 1 genuine face, the removed face area in-stead of the main picture becomes one contribution of the CNN classifier.

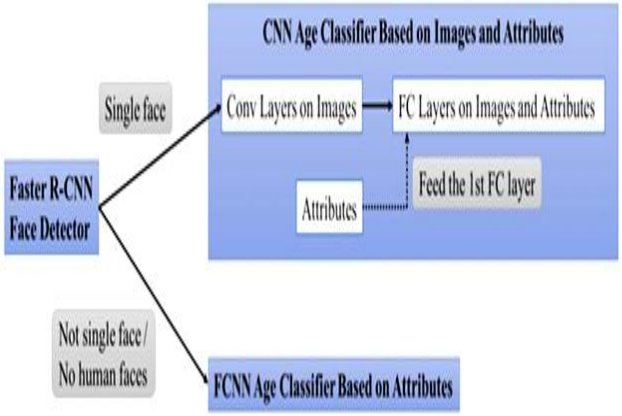

CNN Age Classifier Based on Images and Attributes
Single face
Conv Layers on Images
FC Layers on Images and Attributes
Faster R-CNN Face Detector
Feed the 1st FC layer
Attributes
Not single face / No human faces
FCNN Age Classifier Based on Attributes

Fig.1. System overview of inference attacks based on images

B. Correlation Matrix

Connection network will speak to the connection ships among the picked traits. It shows the dependency between any 2 properties. Condition (1) shows the framework between 2 variable x and y. x is that the variable. y is that the variable. n is that the assortment of in-development focuses inside the example. x and y square measure the mean of x and y, severally. Sx and sy square measure test fluctuation of x and y, respectively.

$$r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \qquad (1)$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \qquad (2)$$

From equation 1 & equation 2 finding RNN, the relationship cost r(x; y) might be an assortment between balanced. In a really immediate connection, as x will expand, y will increment. In a really correlation, as x will build, y diminishes. On the off chance that the value is near the precarious edge of zero, it infers that x and y square measure approximately associated. In the event that 2 properties have higher connection, the foundation shade of that cell is darker. All the overall population characteristics have powerless relationships with each other, and each one the overall population qualities have frail associations with the individual age trait. Since all the traits aren't independent with each other, it's an ideal setting to implement neural systems.

C. Fully Connected Neural Network Construction

R-CNN face locator and CNN age classifier supported pictures are unmistakably incontestable in [13], However; existing papers don't plainly show the configuration of FCNN age classifier on properties. Totally various designs will drastically have an effect on the exhibition of FCNN; consequently it's squeezing to call attention to the important part of FCNN development.

There are 2 motivations to pick FCNNs. To start with, by the character of the issue, the data sorts for the traits are clear, and there are just many recognized characteristics and a few others focused on categories. FCNN is sweet at taking care of this sort of settings. Second, FCNN is practical in calculation and may be just implemented. Subsequently, it's flexible to manage the boundaries of the system to chase the easiest feasible arrangement for the issue. For FCNNs, the mystery is to search out the easiest scope of layers and scope of units in each layer to suit the issue properly. Since there are no fundamental standards to follow to plot those qualities, the least difficult technique is through experimentation, and furthermore the characterize spasms of the issue. When enduring makes an endeavor, the least complex structure of the neural system is that there are a couple of concealed layers and each layer has thirteen units. To minimize (J), backpropagation algorithmic standard is implemented execution; angle checking and irregular initialization are withdrawn.

4. RESULTS AND DISCUSSION

Figure1 shows the examination and correlation among the different pieces of the anticipated system. It will be seen obviously that FCNN bolstered characteristics contains a decent capacity to anticipate the age change for a client inside the informal community. The outcome's in regards to multiple times of the outcome from the arbitrary conjecture. Along these lines, it demonstrates that the overall population ascribes can possibly predict clients' non-open traits by altogether using FCNNs. Be that as it may, since the information contained inside the open credits keeps on being limited, the prediction rate keeps on being underneath five hundredth.

In Figer 2 conjointly shows a correlation between the outcomes before R-CNN balanced and once R-CNN advertisement jested. The introduction of R-CNN is to broaden the malleability of the genius presented structure. In this manner, disregarding the p.c of clients United Nations organization have substantial profile pictures in an incredibly dataset, the casing work will adjust the dataset and expand forecast exactness. To pass judgment on adaptability, completely the qualification before R-CNN balanced and once R-CNN balanced underneath each situation is estimated. On the off chance that the varieties territory unit small, it infers that not exclusively the system is flexible to address totally unique datasets anyway conjointly high generally speaking expectation rate will be justified. In figure, that shows that the expectation capacity of the system is promising and solid. It can likewise be seen that the system yields high by and large expectation rate at 78:00%.

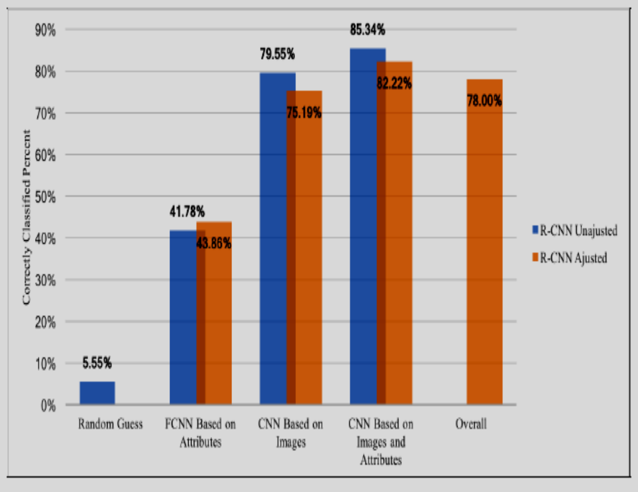


Fig.2. R-CNN balanced

5. CONCLUSION

This paper proposes a substitution system for theoretical idea assaults in informal organizations. The arranged system well integrands and changes the overall dynamic CNN models. Accordingly, it will work more extensive circumstances for unique idea assaults independent of whether a client incorporates a genuine profile picture or not. There square measure 2 significant restrictions during this paper. To begin with, just 1 social net-work is prepared and tried. Second, just 1 focused on non-open trait, the age change, is taken into air conditioning tally. Inside the future work, a ton of informal communities and a lot of focused attributes are investigated to more authorize the show of the arranged structure and to demonstrate a ton of a scope of unique idea at-attaches interpersonal organizations.

REFERENCES

[1] D. Liben-Nowell and J. Kleinberg, “The link-prediction problem for social networks”, Journal of the Association for Information Science and Technology, vol. 58(7), pp. 1019–1031, 2007.

[2] M. Pennacchiotti and A.-M. Popescu, “A machine learning approach to twitter user classification”, ICWSM, vol. 11(1), pp. 281–288, 2011.

[3] N. Benchettara, R. Kanawati, and C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks", Advances in Social Networks Analysis and Mining (ASONAM), International Conference on. IEEE, pp. 326– 330, 2010.

[4] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Data Mining: Practical machine learning tools and techniques", Morgan Kaufmann, 2016.

[5] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases", Computational linguistics, vol. 27(4), pp. 521–544, 2001.

[6] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," Computer Science Department Faculty Publication Series, p. 3, 2005.

[7] Y. Matsuo, J. Mori, M. Hamasaki, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka, "Polyphonet: an advanced social network extraction system from the web", Web Semantics: Science, pp. 262-278, 2007.

[8] I. Habernal, T. Ptacek, and J. Steinberger, "Sentiment analysis in czech social media using supervised machine learning", Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 65–74, 2013.

[9] R. Michalski, P. Kazienko, and D. Krol, "Predicting social network measures using machine learning approach", Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, IEEE Computer Society, pp. 1056–1059, 2012.

[10] A. Luna, M. N. del Prado, A. Talavera, and E. S. Holguın, "Power demand forecasting through social network activity and artificial neural networks", IEEE ANDESCON, Oct 2016, pp. 1–4, 2014.

[11] Z. Li, D. y. Sun, J. Li, and Z. f. Li, "Social network change detection using a genetic algorithm based back propagation neural network model", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 1386–1387, 2016.

[12] A. Khadangi and M. H. F. Zarandi, "From type-2 fuzzy rate-based neural networks to social networks' behaviors", IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1970–1975, 2016.

[13] G. Levi and T. Hassner, "Age and gender classification using con-volutional neural networks", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 34– 42, 2015.

[14] Incorporation of DCT and MSVQ to Enhance Image Compression Ratio of an image, International Research Journal of Engineering and Technology, Vol. 03(03), 2016.

Monitoring Product Fake Review for Genuine Rating

Chandra Shekar K[1], Ambatapudi Shyamala [2], Chanda Soumya [3], Allenki Neha [4]

[1] Associate Professor, Department of Computer Science and Engineering,
Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India
[2,3,4]Scholar , Department of Computer Science and Engineering,
Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India
Email: [1]chandhra2k7@gmail.com, [2]shyamalaambatapudi@gmail.com,
[3]soumyachanda123@gmail.com, [4]allenkineha5@gmail.com

Abstract - Review plays a vital role for the sales of products in today's world. Whenever a product is being bought, people check reviews about the product. Users go through many reviews in the website, but they are unable to get whether the review is genuine or not. In some websites, reviews are added by the company itself to produce false positive product reviews for increasing the publicity. Since most of the good reviews are given by the company, the users suspect the soundness of the reviews. In this paper, fake product review monitoring and removal is done using opinion mining. In order to find out fake reviews in a website, the system will trace the fake reviews with the help of review posting patterns and by identifying IP address. The IP address is identified / recorded, and if the same IP address is repeated multiple times, the system eliminates the duplication of that review. This helps the user to find genuine review about the product.

Keywords--Opinion Mining, Fake Review, Genuine Rating, Order Tracking.

1. INTRODUCTION

In today's generation, people generally tend to use online shopping rather than going out to buy any product. But to purchase the products, users tend to analyze the ratings/reviews of products given by various persons to know whether the product is worth enough to purchase or not. Users come across many reviews in the website about the product, but the user cannot identify whether the reviews are genuine or fake. In some of the websites, most of the good reviews are added by the company people itself to make product famous inclining the people to buy the product. The company people make their own reviews for many products manufactured by their own firm to make the product sale increasing. To identify and remove these fake reviews, this monitoring system is developed. This process is done by identifying the IP address along with MAC address of the user from which multiple instances of the reviews are generated with different names or details, making them fake reviews. These reviews will be identified and sent to the administrator, who can analyze the reviews and remove them if they are considered as fake. This system also allows administrator to rate a product based on the user reviews.

2. RELATED WORK

Generally, e-commerce websites facilitate their customers to post reviews and feedback of the product in the form of ratings. These type of reviews cannot be verified as the genuinity of the user cannot be checked. These can also be spammed in order to make product famous and also to get more profit, or may contain fake reviews or malicious opinions, which will mislead the endusers. Shashank kumar Chauhan et. al. have proposed a system that detect spam and fake reviews, and filter out reviews through vulgar and curse words, by incorporating sentiment analysis.

Hamzah Al Najada et. al. have proposed a system to distinguish between spam and non-spam reviews by using supervised classification methods. In order to tackle this issue, they employed a bagging based approach and have build a number of balanced datasets, through which a set of spam classifiers are trained and use their ensembles to detect review spams.

Based on fuzzy analytic hierarchy process, Xinkai Yang used a method of risk assessment standards of industrial control system, introduced a fuzzy consistent matrix and entropy method which is used to overcome the lack of fuzziness in the evaluation result of traditional analytic hierarchy process.

Ruxi Yin et. al. targeted on opinion spam detection methods in which spam refers to fake reviews, in which well-designed fake comments are targeted at damaging a specific product by an individual or an organization.

Sonu Liza Christopher et. al. used Big Pi model which will be useful in tracking such reviews through their associated social media accounts.

SP.Rajamohana et. al. have done a detailed survey using various machine learning techniques for detecting spam and genuine reviews.

K. Chandra Shekar et. al. proposed a hybrid technique of ensemble classifier, to provide a better solution for the classification problem.

3. PROPOSED METHODOLOGY

In this paper, a proposed system, termed as " Monitoring and removing system"is developed which helps to find out whether a review of the product is fake or not. The reason for introducing this system is to purchase the products with genuine rating, as fake reviews are also added by anybody, who may be an owner / manufacter or seller of that particular product. If the reviews are fake, the users or customers of the product can't able to find these types of reviews, by identifying them with the help of IP address of users from which the posts are made.

Proposed methodology provides an advantage over the existing system, as people get genuine reviews about the product, and customers can post their own review on a product, so that people can spend money on worthy products instead of the hyped products.

Fig. 1 shows Architecture diagram, which includes two modules. They are namely admin, user. Here the user will register and login to the website. For placing an order the user will be able to see the reviews. The admin can check and add genuine reviews about the product and remove any fake reviews present in the website.

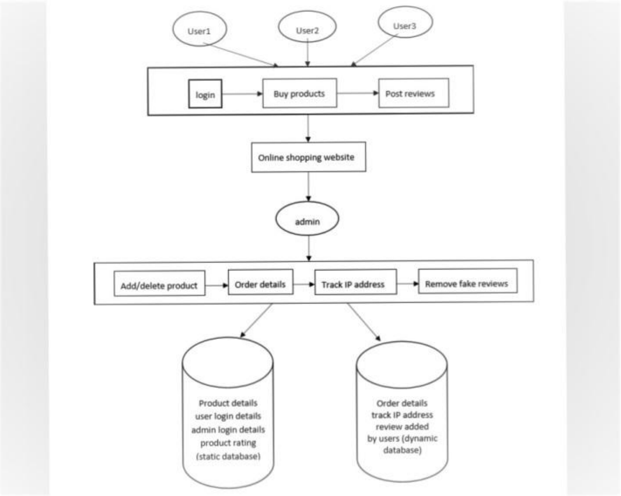
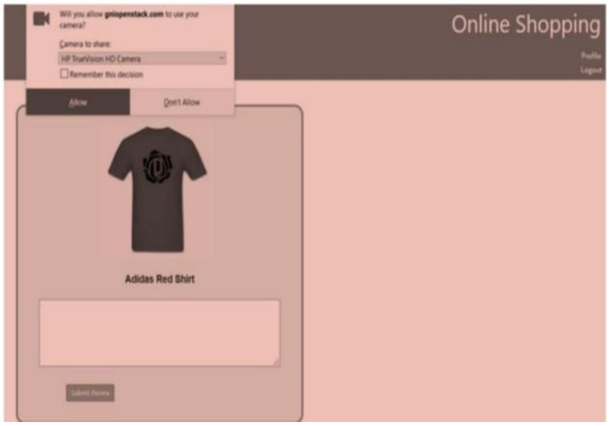


Fig. 1 : System Architecture of the Proposed Methodology

4. RESULTS AND DISCUSSIONS

The results obtained after implementation of the system are provided in this section. The home pageis displayed when the user opens the website. The login page where user has to login with their credentials. The main objective of our work is to create a system which will detect spam and redundant reviews and to filter them so that user obtains correct knowledge about the product. The aim of our project is to enhance customer satisfaction as well as to make online shopping reliable. The project will detect the fake reviews by fetching IPs. Results are shown in Fig. 2 and Fig. 3.

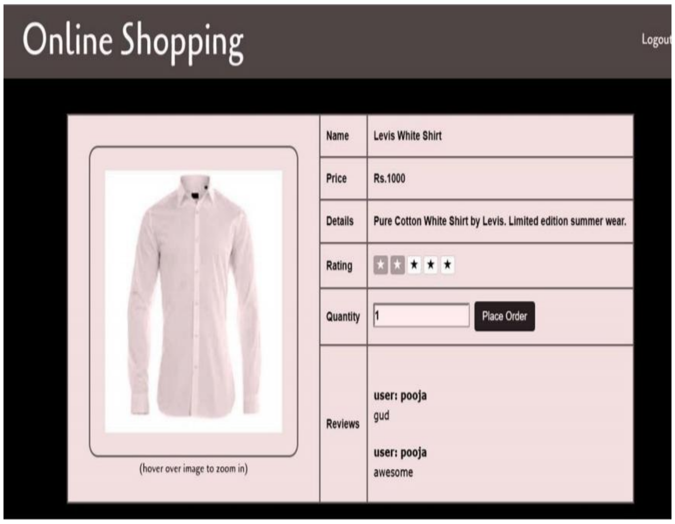


Will you allow gniopenstack.com to use your camera?

Camera to share: HP TrueVision HD Camera

□ Remember this decision

Fig. 2 : Page to insert reviews



Name	Levis White Shirt
Price	Rs.1000
Details	Pure Cotton White Shirt by Levis. Limited edition summer wear.
Rating	★ ★ ★ ★ ★
Quantity	1
Reviews	user: pooja gud user: pooja awesome

Fig. 3 : Order Confirmation page

5. CONCLUSION AND FUTURE DIRECTIONS

Finally we conclude from our work that this application helps the user to spend money for valuable product rather than getting into the nest of hyped products. This application will analyze and post only the genuine reviews of the products. The users can acquire the correct details about the products availability as well as the genuine reviews on the application or website.

As a part of our future work, we would try to develop the method by including the concept of calculating the sentiment score of the reviews. We would also take the necessary precautions in updating the dictionary comprising of sentiment words. We would also ensure to add more words in the dictionary and update the weights given to those words to get more accuracy through calculated score of the reviews. A direction for future research is to implement the system and check performance by applying proposed approach to various benchmark data sets.

REFERENCES

[1] Shashank Kumar Chauhan, Prafull Goel, Anupam Goel, Avishkar Chauhan and Mahendra K Gurve, Research on product review analysis and spam review detection, 4th International Conference on Signal Processing and Integrated Networks (SPIN), ISBN (e):978-1-5090-2797-2, pp.1104-1109, 2017.

[2] Sonu Liza Christopher and H.A Rahulnath, Review authenticity verification using supervised learning and reviewer personality traits, International Conference on Emerging Technological Trends (ICETT), ISBN (e): 978-1-5090-3752-0, pp.1-7, 2016.

[3] Hanshi Wang, Lizhen Liu and Ruxi Yin, Research of integrated algorithm establishment of spam detection system, 4th International Conference on Computer Science and Network Technology (ICCSNT), ISBN (e): 978-1-4673-8173-4, pp. 390-393, 2015.

[4] Xinkai Yang, One methodology for spam review detection based on review coherence metrics, International Conference on Intelligent Computing and Internet of Things, ISBN (e): 978-1-47997534-1, pp.99-102, 2015.

[5] Hamzah Al Najada; Xingquan Zhu, iSRD: Spam review detection with imbalanced data distributions, Proceedings of the IEEE 15th International Conference on Information Reuse and Integration, ISBN (e):978-1-4799-5880-1, 2014.

[6] SP. Rajamohana, K. Umamaheshwari, M. Dharani, R. Vedackshya, A survey on online review SPAM detection techniques, International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT) 2017, ISBN(e): 978-1-5090-5778-8., International Conference on Innovations in Green Energy and Healthcare Technologies (IGEHT), ISBN(e):978-1-5090-5778-8, 2017.

[7] K. Chandra Shekar, Priti Chandra, K. Venugopala Rao, An Ensemble Classifier Characterized by Genetic Algorithm with Decision Tree for the Prophecy of Heart Disease, International Conference on Innovations in Computer Science and Engineering, 2019.

[8] K. Chandra Shekar, K. Ravi Kanth, K. Sreenath, Improved Algorithm for Prediction of Heart Disease using case based reasoning technique on non-binary datasets, International Journal of Research in Computer and Communication technology, ISSN 2278-5841,Vol 1(7), 2012.