ENRICHING HEALTH CARE FRAUD DETECTION SYSTEM USING ANN

N.I.Ujoomwale[1], Aditi Kamath[2], Darshana Akadkar[3], Pooja Divase[4], Shraddha Hundalekar[5]

[1] *Assistant Professor , Department of Computer Engineering, MES College Of Engineering, Maharashtra, India*
[2] *BE Student, Department Of Computer Engineering, MES College Of Engineering, Maharashtra, India*
[3] *BE Student, Department Of Computer Engineering, MES College Of Engineering, Maharashtra, India*
[4] *BE Student, Department Of Computer Engineering, MES College Of Engineering, Maharashtra, India*
[5] *BE Student, Department Of Computer Engineering, MES College Of Engineering, Maharashtra, India*

ABSTRACT

Unavailability of a medical fraud detector encourages the rate of fraud in the healthcare insurance sector to increase by a great extent such that the medical practicians delude the insurance holders very well. Consequently, in order to do away with these fraudulent insurance claims, a system that discovers the fake claims and facilitates in scaling down the rate of insurance claim frauds is essential. Several techniques like Social Network Analysis (SNA), Duplicate and Gap testing, Spike Analysis, Social Customer Relationship Management (SCRM) and Predictive Modelling are utilized in order to attain a system owning to a successful fraud detection process. But, on the contrary, these systems face troubles of a large time and space complexities. To enhance the process of fraud claims detection of the doctors at the insurance company's end the proposed method puts forwards an idea of identifying fraud claims by clustering the claims based on the protocols by using the K-means clustering technique which is then powered with ANN to extract the fraud list and this process is supported by fuzzy logic classification hypothesis.

Keyword: *-Data labels, Data vector, Preprocessing, Protocol collection, K-means Clustering, ANN, Fuzzy Classification, probability, Fraud Estimation, Fraud claim identification, Insurance, Algorithm design and analysis, Drugs (Medicines)*

1. INTRODUCTION

The Healthcare insurance industry has gone through a enormous alteration over the few years. But India faces a huge economic loss in this sector because of the increasing amount of fraud claims. This sudden uprise in the fake claims is due to the fraudulent practices of the medical practicians. A claim generated to cover or deform information which is actually designed and modernized to produce better healthcare benefits, is termed as a fraudulent health insurance claim.

Therefore, to bring the level of frauds to a very low rate, a system must be developed which can accurately identify the different fraudulent claims. The proposed system puts forth the idea, where the doctor's claims are collected and processed such that the data is then labelled numerically.This labelled data is further put into different groups and then depending on this grouping of data, the claims can be easily identified as a fraud or not. Major Data mining and Data analytic techniques like K –means, Artificial Neural Networks and Fuzzy Logic Classification are used such that the data is efficiently and accurately identified for any sort of fraud.

1.1 K-MEANS CLUSTERING

A heuristic partitioning method used in order to partition the large data sets and solve the clustering problem efficiently is named as K-means Clustering. In the K-means approach the data objects are classified based on their attributes or features into “k” number of clusters. The k centroids have to be determined as the primary step for every cluster. The distance between the different clusters must be maximized while the distance between the data points within the cluster must be minimized. Assign points of the given data set to the nearest centroid and then calculate the new centroid for each cluster depending on the data points present in that cluster using suitable distance formulas. Repeating the same method again, group the data points based on the new centroid. With every iteration the centroid changes and the data points are shuffled from one cluster to another. This process has to be done until no data point is moved from one cluster to another and the clustered data are obtained.This k-means clustering algorithm can be used in the study of earthquakes based on the regions which are affected by the earthquakes.Also it can be useful in identifying the groups of houses based on the geographical locations.

1.2 ARTIFICIAL NEURAL NETWORKS

Systems similar to the Neurological functioning of the Human Body are defined by experts as Artificial Neural Networks (ANN). These systems keep learning from the different examples, without having any previous knowledge about those examples. The system learns from its previous experiences. ANN is based mainly on the aggregation of the different nodes called as the Artificial Neurons. The signals that are transmitted, are in the form of real numbers, while the output obtained from each of the artificial neuron is measured by a nonlinear function. The artificial neurons and the connections between them have a certain definite weight, which adjusts itself as the system proceeds with the process of learning. ANN is useful in recognizing patterns and speech, machine translation and video game application.

1.3FUZZY LOGIC CLASSIFICATION

Fuzzy logic classification is said to contain a set of different elements whose membership function is defined by a truth value. Multivalued logic in which real numerates are taken as truth values that range between 0 and 1 is termed as Fuzzy logic. Fuzzy logic uses IF-THEN rules which help map input considered truth values. Fuzzy logic is used in different fields like pocket computers, flight aid for helicopters, controlling subway system to improve the driving comfort.

2. LITERATURE SURVEY

Paper [1] describes K-means as an outlier detection for finding an outlier in network analysis, to find a community overlapped user in the network. Also it finds more k-clique that describes the strong coupling of data. k-means clustering is applied to make clusters with the help of a Euclidean distance formula.

Paper [2] introduces a design of k-means that shows the map reduce design for handling large datasets so that it will be effective in reduction of execution time. It starts with an small number of k-centres, selection and update iteratively for minimization of an error function. The benefits of this paper are simpler implementation with high efficiency and scalability and linear complexity of time.

Paper [3] explains the recently proposed hybrid binarization methods regarding k-means clustering algorithm. It consists of handling effective scanning of documents which consists of text based data on a simple background, in which application is carried out on each block independently. It gathers pixel values and numbers related to respective clusters of every block for the performance of corrective loop.

Paper [4] puts forth the idea of the application of k-means algorithm for overcoming cluster genes and redundant genes. In this every sample represents a large set of genes so this makes the algorithm more time consuming hence, using k-means pre filtering method by dividing genes into clusters and the accurate and effective output is obtained.

The proposed system in paper [5] gives Multilayer perceptron (MLP) artificial neural network (ANN) is the ideal model for allocation and detection. It allocates liver to recipients who will most possibly die, but have chances of survival after Liver Transplantation (LT). The Proposed model estimates the successful long-term probability of survival for periods including six months to 13 years. The 27 input attributes in the dataset are given to the MLP model. The model trained the clinical attributes of a patient related to LT using back propagation algorithm. The trial and error method is used to select 13 nodes for hidden

The system in paper [6] gives ECG based authentication for health care data security using Artificial Neural Network (ANN). The system uses two stage identification using 2 NN models which are "generally" NN model and "personal" NN model. The "General" NN model is used for preliminary screening and "personal" NN model is used for specific recognition. The ECG signals QRS/QRST are pre-processed and features are extracted from it. Later, this feature is matched with templates to generate a matching matrix and fed it to both NN models. The input data is a matrix which consists of the distance between features dataset and template features given to both NN models. The result of the system depends on the results of both models.

The paper [7] describes ANN for analysing health and active power of wind turbines using SCADA data. SCADA is a software used to gather data in real time from remote locations for controlling equipment and conditions. The wind turbine dataset in pre-processed and passed to the back propagation ANN (BP-NN). There are three different transfer functions such as tangent sigmoid, logarithmic sigmoid and hard limit is used to construct BPNN. A network has one hidden layer and 10 neurons. The Neural Network output power and actual power are compared to get health value.

The proposed system in paper [8] gives the diagnostic model of obstructive sleep apnea hypopnea syndrome (OSAHS) using artificial neural networks (ANN). The medical records of 43 patients in which 31 persons suffered from OSAHS in records and others suffered from other respiratory diseases. The logistic regression model is used for analysing medical records. The ANN classification model consists of 9 units in the input layer, hidden layer has 3 and the output layer has 1 unit. The input layer parameters are past medical history (hypertension, coronary heart disease), oxygen desaturation index, breathing type , pulse, sleep position, Epworth sleepiness scale assessment, smoking history, Mallampati grade and STOP-BANG questionnaire.

Paper [9] explains the study related to the daily-care of β-thalassaemia patients with the help of Clinical Decision Support Systems which is a 3-tier architecture model. The data of the thalassaemia patients are collected which is in different forms like digital images, clinical test information, Magnetic Resonance Imaging (RMI) etc.. The Fuzzy Inference Machine uses the a linguistic variable. The variable is characterized by a quintuple and the assignment equation L has the form x=: UF (L). The Fuzzy Inference Machine has been implemented to evaluate the Iron overload and status of liver and heart of the patients. LIC and T2 [23] are the two parameters used for assessing the Iron overload of heart and liver, respectively and are the inputs to the fuzzy logic. The main aim of the machine is to synthesize on the liver and heart status. The fuzzy relational model known as Mamdani scheme is implemented into the system. Each rule in this is represented by an "if antecedent, then consequent" relationship. Depending on the given input the membership αj of the rule Rj is calculated. In the next step the result of antecedent is applied to consequent. The input of the implication process is a single number given by the "antecedent" and the output is a fuzzy set. Finally the defuzzification process is being performed starting from the output fuzzy set resulting from the aggregation process. For the evaluation of iron overload a colour scale and is compared with the patient's clinical status. Red label indicates high risk and green level indicates no risk. An interface of patient's clinical details is being displayed which enables the patients to know about their health status and also shows whether the treatment is ineffective or if the patient has followed an ideal treatment sequence.

Paper [10] aims at providing the Healthcare-as-a-service which is presented by using a fuzzy-rule based classifier.The proposed system enbles the doctors to input the symptoms of new diseases and can correctly find out only those patients who have been affected.The initial clusters formed are been collected,retrieval and processing of big data is taken place in the cloud environment.Fuzzificztion and defuzzification processes include the membership functions.These functions are designed so as to get the inferences from the collected data.Various evaluation metrics put forward are average response time,accuracy,computationcost,classification time and false positive ratio.A fuzzy-classifier is used inorder to retrieve the fuzzy rules specified by the experts and also stores the fuzzy data values into clusters. The doctors queries the inference system and the results of the respective patients is sent back to them. Effective results are obtained with respect to evaluation of various performance metrics in a cloud computing environment.

The proposed paper [11] elaborates the practical and effective use of regenerative braking energy which is used in controlling the speed of electric vehicle. Both the output of motor controller of electric vehicle along with fuzzy logic controller act as the feedback to the motor and thus the improvement in speed is observed. This strategy uses regenerative and frictional braking force for the stability and braking safety. This results in the conversion of kinetic energy into electrical energy and this can be stored in the battery. MATLAB software is used for modelling the FLC. The rules of fuzzy logic are built by having any mutual relation between battery, speed of vehicle and braking force. The output is the ratio between the regenerative braking force and total braking force which is again given as a feedback to BLDC thus improving the speed as it uses regenerative braking energy. Thus, by using the FLC the speed of the vehicle gradually increases in positive and reduces the torque to a very acceptable a value.

The paper [12] explains an application of Fuzzy Logic Controller, which is built based on Automatic Voltage Regulator system. AVR device controls the reactive power and the terminal voltage of the synchronous generator is being controlled at a safer level.Triangular membership functions are used by the AVR systems for analysing. The FLC takes two inputs as error voltage and change in error voltage. The falsification changes the input data values to linguistic values. Fuzzy control rules are being provided by the data from the databases. The core part of the FLC is the interference engine as it provides a decision making capabilities based on the fuzzy concept and interference rules in fuzzy logic. In the defuzzification process, by scale mapping output variables are decided and these linguistic output variables are converted into the numerical values.The simulation consists of the performance check of the AVR system with six different PID controllers. As a result, it is found that the FLC based AVR system gives an efficient result and performs better the PID controller.

Paper [13] brings about the idea of New approaches in the Health Care Fraud Detection Methods. This paper proposes a survey of the fraud detection methodologies which is based on the Big Data approach. Reduction in the health care cost, Increase in the efficiency of the fraud detection and providing high quality data to the health care system are the significant features of the system developed. Also the paper shows major advantages like Automatic learning of Fraud Patterns from the Data, Identifying newer and different types of frauds which was not possible previously. The Big data method helps to analysis a huge amount of data which is highly complex and impossible to gain control of,

Paper [14] identifies Anomalous Insurance Claims using the Machine Learning model which helps to detect when physicians exhibit anomalous behaviour in the processing of the medical insurance claims. The model developed made use of the multinomial Naïve Bayes algorithm and it is assessed by calculating the precision, recall and F-score. This is done with the assistance of the 5-fold cross-validation. With the development of this system, a fraud free insurance claim can be obtained which provides a relief to the commoners who can now avail the benefits of the medical facilities. A Supervised Learning approach of Naïve Based classification helps in the predicting and detecting the fraudulent claims.

3. PROPOSED METHODOLOGY

The proposed methodology of fraud claim detection through aritificial neural network can be broadly narrated using the following steps. And the whole process is depicted in figure 1.

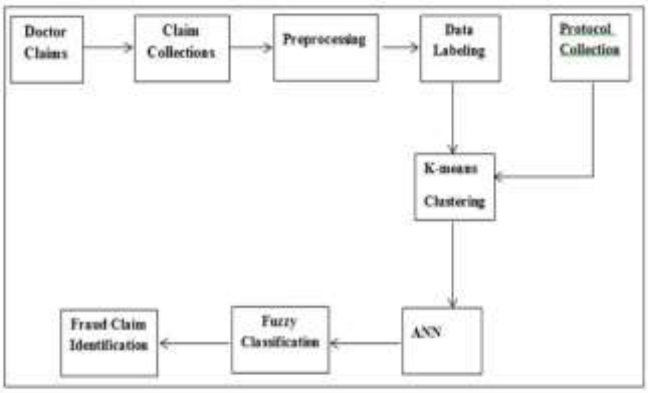

Figure 1: Proposed System overview

Step 1: Doctor Claim List- This is the initial step of the proposed model where the doctors who are willing to claim his/ her service is mentioned all the attributes in the work book and uploaded in the system. The Insurance personnel will feed these insurance claims of the services belong to the different doctors to the sytem.This workbook is having some attributes like Claim ID, Doctor name, Patient name, Disease, Suggesting medicines, Number of recalls , Period between the recall , Referring Doctor , Service cost, Suggesting procedures and sggesting dietaries.

Once the system accepts these claims as the input, then these claims are read from the workbook file into the double dimension list using JXL third party API.

Step 2: Protocol Collection - Proposed model sets some standard protocols for each and every type of the diseases for all the service related queries that can arise at the Doctor's end on behalf of the medical counsil. These protocols are stored in the database and this data is collected from the database in the form of the double dimension list.

Step 3: *Preprocessing and Labeling* - Once the doctor's claims are received, it is statically stored in the instance fraud detection process. And the double dimension list of data are subjected to preprocessing, where predefined attribute column data is selected for the further process of fraud detection. These preprocessed column's data is used to label them with the unique integer numerical value. This is done by analyzing the all protocols with respect to the doctors claims, then a ratio is estimated to get the numerical values. These numerical values evetually indicate the strength of the claims. After this all the rows of the preprocessed list are assigned with the integer numerical value which is subjected to cluster in the next step of the model.

Step 4: K means Clustering - Obtained a numerical list of the past step is subjected to find the sum of all the numerical values to append at the end of the each row. Then these rows are sorted in the ascending order using bubble sort which yields the minimum and maximum values of the summation list. Then by using these minimum and maximum values a distance is being evaluated that is divided into the two halves and it is considered as the data points of the Kmeans clusters. Then each row is added into the estimated datapoint ranges to form clusters of the claims.

Step 4:Artificial Neural network (ANN) -First Layer - Here in the First layer of ANN claim clusters are subjected to find the mean and standard deviation of the sumamtion values of the rows. Then these clusters are tend find the convolution inner range and the outer range based on the mean and standard deviation ranges as mentioned in the below equation 1 to 4.

$$\mu = \frac{(\sum_{i=1}^{n} Edi)}{n} \qquad \_\_\_\_\_\_(1)$$

$$\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Edi-\mu)^2|} \qquad \_\_\_\_\_\_(2)$$

$$f(I_p) = \mu - \delta \rightarrow \mu + \delta \ \_\_\_\_\_\_(3)$$

$$f(Op) = < \mu \rightarrow > \mu \ \_\_\_\_\_\_\_(4)$$

Where

δ - Standard Deviation

μ- Mean

Edi - Euclidean distance of instance row

N- Number of Rows in cluster

$f(I_p)$ - Inner range Function

$f(O_p)$ - Outer range Function

Based on these inner and outer range size a probability cluster is identified for the claim clusters.

Secondary layer - Here in this step all the respective values of the columns belong to each neuron cluster are subjected to summation function to get the mean value. Then these neuron rows are merged to get a single qualitative neuron than holds the traces of the fraud claims.

Deep Layer - Here the merged neuron is considered to find the factorized values in between the neuron rows to yield a two dimension list with two columns . where the second column indicates the ANN Score of the rows.

Step 5: Fuzzy Classification - This is the final step of the fraud detection process of the proposed model. Here ANN Score list is subjected to Fuzzification process where each of the Fraud claims are considered as the neuron representing with a score. These Claims are converted into fuzzy crisp value ranges like VERY LOW, LOW, MEDIUM, HIGH AND VERY HIGH which eventually indicates the fraud levels. Based on these crisp values IF- THEN rules are being applied to get the classified fraud according to the given criteria. The process of Fuzzy ANN can be shown with the below mentioned alogorithm1.

ALGORITHM 1: FUZZY ANN

```
//Input : ANN Score Vector $E_v$
//Output: Classified list $C_L$
Start
Set small=0, big=0
    For i=0 to size of $E_v$
        $T_{Set} = E_{vi}$   [ $T_{set}$ = Temporary Set]
        Sc=$T_{set[1]}$
        IF ( Sc <small) [ sc= Score]
        small=Sc
        IF(Sc>big)
        big=Sc
End for
    d=( big-small)/5   [ d= Distance ]
        For i=1 to 5
            IF(i==0)
        Fc(min=small, max=d) [ Fc = Fuzzy Crisp Set ]
            else
            Fc(min=Fci-1(max),max= Fci-1(max)+d
        End For
            For i=0 to Size of Fc
                For j=0 to size of Ev
                $T_{Set}$ = Evi   [ Tset = Temporary Set]
                Sc=Tset[1]
                    IF Sc ∈ Fci
                    add $T_{Set}$ to $C_L$
                    END IF
                End For
            End For
return $C_L$
```

4. RESULT AND DISCUSSION

The proposed methodology of Fraud detection model is deployed in windows based java machine using netbeans as IDE and mysql as the database server. To measure the performance of the system for the accuracy of the fraud detection, model considers the precision and recall as the measuring parameter.

Precision and recall are considered as the one of the best parameter to measure the performance of our system. Precision can be defined as the positive classfified values that indicates the amount of relevant fraud claitms are identitied by the system.

Precision can be described as the ratio of number of relevant fraud are detected for the input number of claims to the sum of number of relevant and irrelevant fraud claims are classified for the input number of claims. Relative effectiveness of the system can be evaluated thoroughly by using precision parameters.

Recall indicates the relevant results extracted over the extracted relevant results. Recall can be described as the ratio of number relevant fraud claims are detected to the sum of relevant fraud claims are not detected. Absolute accuracy of the system can be properly denoted by this system.

Precision and recalls can be more effectively explained as below

- ✓ A = The number of relevant fraud claims are detected for the given number of claims
- ✓ B= The number of irrelevant fraud claims are detected for the given number of claims
- ✓ C = The number of relevant fraud claims are not detected for the given number of claims

So precision can be given as

Precision = (A / (A+ B)) *100

Recall = (A / (A+ C)) *100

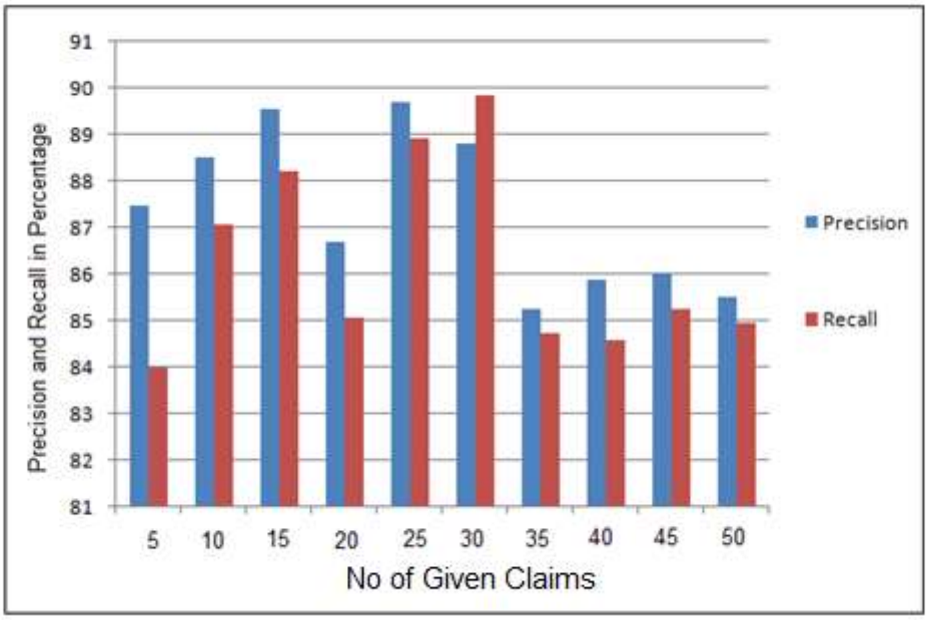

Figure 2: Performance Evaluation through precision and recall

The graph in above figure 2 clearly indicates that proposed model for fraud claims detection sytem achieves 87.34 % of average precision and 86.27 % of average recall for the given number of claims. Which is high and indicates successful deployment of the proposed idea.

5. CONCLUSION AND FUTURESCOPE

It is being said that doctors are like god, So patient oftenly believe them as just trusting like a god. This trust oftenly takes doctors in a wrong way to take advantage of the patient's innocency. Generally, this kind of things happened when doctors claim the fraud claims regarding the services provided to the patients. So the proposed model putforwards an idea of finding the fraud claims of the doctors by using data mining as the tool. This is achieved by the K means clustering and the Artificial neural network with a blend of fuzzy logic. And proposed model achieves good precision and recall rates as discussed in the past section .

The proposed model can be deployed in real time scenario of insurance companies for the big data mode using distributes computing paradigm.

6. REFERENCES

[1]. Parmeet Kaur, “Outlier Detection using Kmeans and Fuzzy Min Max Neural Network in Network Data”, 1-5090-1144-5/16, IEEE 8th International Conference on Computational Intelligence and Communication Networks, 2016
[2]. Amira Boukhdhir, OussamaLachiheb, Mohamed Salah Gouider, “An improved MapReduce Design of Kmeans for clustering very large datasets”, 978-1-5090-0478-2/15, IEEE, 2015
[3]. Mahmoud Soua, RostomKachouri, Mohamed Akil, “Improved Hybrid Binarization based on Kmeans for Heterogeneous document processing”, 978-1-4673-8032-4/15, IEEE 9th International Symposium on Image and Signal Processing and Analysis (ISPA) 2015
[4]. Jianqiang Li, Fei Wang, “Towards Unsupervised Gene Selection: A Matrix Factorization Framework”, 1545-5963, IEEE ACM, 2016
[5].C. G. Raji and S. S. Vinod Chandra, “Long-Term Forecasting the Survival in Liver Transplantation Using Multilayer Perceptron Networks”, 2168-2216, IEEETransactions On Systems, Man, And Cybernetics: Systems, 2017
[6]. Ying Chen and Wenxi Chen, “Finger ECG-based Authentication for Healthcare Data Security Using Artificial Neural Network”, 978-1-5090-6704-6, IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), 2017
[7]. Hui Li, Jiarong Yang, Menghang Zhang, ShuangquanGuo, Wei Lv and Zongchang Liu, “A method based on artificial neural network to estimate the health of wind turbine”, 978-1-4799-7016-2, IEEE, 2015

[8]. Jing Bin; MENG Hai-bin; YANG Song-chun, ZHAO Dong-sheng and SHANG Xue-yi , ”The diagnostic model of obstructive sleep apnea hypopnea syndrome based on artificial neural networks” , 978-1-5090-3906-7, IEEE 8th International Conference on Information Technology in Medicine and Education, 2016
[9].S. Santini, A. Pescape, A. S. Valente, V. Abate, G. Improta, Maria Triassi and P. Ricchi, A. Filosa,“Using fuzzy logic for improving clinical daily-care of β-thalassemia patients” 978-1-5090-6034-4/17, IEEE, 2017
[10].Anish Jindal, Amit Dua, NeerajKumar,Ashok Kumar Das, A. V. Vasilakos and Joel J.P.C. Rodrigues, “Providing Healthcare-as- a-Service Using Fuzzy Rule-Based Big Data Analytics in Cloud Computing”, 2168-2194 IEEE Journal of Biomedical and Health Informatics, 2018
[11].Surabhi Agrawal and Dr.VivekShrivastava ,” Speed Control of Electric Vehicle Using Fuzzy Logic Controller” ,IEEE International Conference on Information,Communication,Instrumentation and Control(ICICIC), 2017
[12].Tripti Gupta, D. K. Sambariya,” Optimal design of Fuzzy Logic Controller for Automatic Voltage Regulator”, 385, IEEE International Conference on Information,Communication,Instrumentation and Control(ICICIC), 2017
[13].EbruAydo˘ganDuman, SerefSa˘gıro˘glu, “ Health Care Fraud Methods and New Approaches”,978-1-5386-0930-9/17, IEEE 2nd International Conference on Computer science and Engineering, 2017
[14].Richard A. Bauder, Taghi M. Khoshgoftaar, Aaron Richter, Matthew Herland, “Predicting Medical Provider Specialties to Detect Anomalous Insurance Claims”, 2375-0197/16, IEEE 28th International Conference on Tools with Artificial Intelligence, 2016

******