Improved Fraud Detection in e-Commerce Transactions

Jisha Shaji
PG Scholar
Department of Computer Engineering
St. Francis Institute of Technology, Mumbai, India
jishatshaji@hotmail.com

Dakshata Panchal
Assistant Professor
Department of Computer Engineering
St. Francis Institute of Technology, Mumbai, India
dakshatapanchal@sfitengg.org

***Abstract*—Online transactions have gained popularity in the recent years with an impact of increasing fraud cases associated with it. Fraud increases as new technologies and weaknesses are found, resulting in tremendous losses each year. Since the transactions associated with e-commerce are large in number, the dataset associated with them is also large; therefore, it requires fast and efficient algorithms to identify fraudulent transactions. Most of the methods used for fraud detection are rule-based or are systems that require re-training when newer patterns of fraud occurs. Detecting fraud as it is happening or within a short time span is not easy and requires advanced techniques. As the demand has arisen for self-learning predictive systems, the main objective is to detect the fraudulent transactions by using Adaptive Neuro-Fuzzy Inference System, which is a hybrid of neural networks along with fuzzy inference, wherein the system can adapt to newer instances of fraud.**

Keywords*—Fraud Detection; hybrid approach; e-commerce; Machine Learning; e-payment.

I. INTRODUCTION

There has been a tremendous increase in electronic transactions during the last decades, due to the popularization of the World Wide Web and e-commerce [1]. The number of card issuers, card users and the online merchants has increased [2]. This is mainly due to the various online retailers like eBay, Flipkart, Amazon, Walmart to name a few. Individuals have changed their mode of payment significantly with the growth of modern technology. Most of them make use of online payment modes while shopping online or at the market. The fraudulent activity on a card affects the cardholder, the merchant, the acquiring bank and the issuer. With regard to the cost of fraud, the most affected participant is the merchant, because the cost of fraud is greater than the cost of goods sold.

Cyber-crime is a crime committed over internet. Fraud is generally defined as a criminal activity committed by the criminal in order to obtain financial/personal gain [2, 3]. Fraud can be mainly divided into two types: Offline Fraud and Online fraud.

- **Offline fraud** is the one which involves some physical activity such as stealing purse/wallet which contains valuables like credit card, ID proofs etc. and using the crucial information within them.
- **Online Fraud** occurs when the fraudster uses an electronic medium or creates a website and presents their website as genuine to obtain crucial personal information and perform illegal transactions on such customer accounts. Other ways through which the fraudsters collect/steal personal information are Hacking, Phishing, Spoofing, Spyware, Shoulder Surfing, Dumpster Diving etc. [3].

The online transactions take place within a fraction of seconds. Since a large number of people are associated with such e-commerce transactions, the dataset associated is also large. There exists a need for developing fast and efficient algorithms to process such large datasets and to search for fraudulent and deceitful transactions.

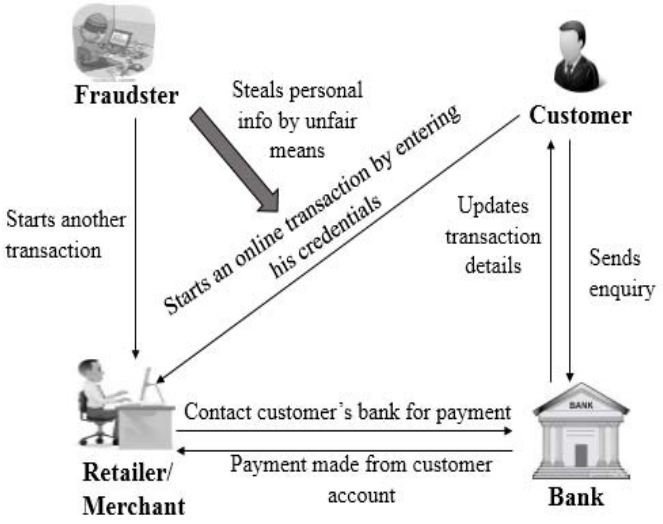

Fig. 1. Example of Transaction fraud.

In the Fig. 1, it can be seen that, the fraudster gains the personal credential information of the customer via some unfair means and uses the same for online shopping. The cardholder realizes that fraud has occurred and starts an enquiry for the same at a later stage after the transaction is complete and the money is lost. Although there are several fraud detection techniques based on Data Mining, Knowledge Discovery and

Expert System etc., they are not capable enough to detect and prevent the ongoing fraud.

The major challenge of fraud detection is the very limited time span in which acceptance or rejection is to be done. Also another peculiarity is the large number of transactions that has to be processed at a given time. Detecting fraud at the earliest when the transaction is being processed is a major concern in e-commerce systems. Another major concern of the research is to avoid rejecting the genuine customers. The fraudulent activities impose considerable financial losses to merchants and therefore fraud detection becomes a necessity.

Also the most important thing to remember is that the fraudsters are constantly updated, so there arises a need for systems which keeps on adapting themselves as and when newer instances of fraud occurs. In this paper, a system to detect fraudulent transactions with increased accuracy and adaptation to newer instances of fraud has been proposed.

The paper is organized as follows: Section II describes the literature review in the area of fraud detection in electronic transactions. Section III describes the proposed system. Section IV describes the detailed experimental methodology and results and finally Section V presents conclusions and future work.

II. BACKGROUND AND RELATED WORK

Electronic or credit card fraud detection has drawn a lot of attention in the last few decades. Some of the works that are related to fraud detection in electronic transactions or credit card operations are described in this section.

In [1, 4] the authors have used a Neural Network based approach which uses MLP. In Neural Network [2], the interconnection weights between different nodes are learned during the process of training and the processing ability of the network is computed by the learnt weights. The features of the transaction were given as inputs and the inputs are weighted. This weight shows the intensity of how a particular input influences the output value. The network computes the difference between the actual and the desired output and propagates it backward to adjust the weights assigned to the inputs. Thus, it can be inferred that, fraud detection using neural network is based on Pattern Recognition, i.e. when a fraudulent transaction is detected; the weights of the inputs related to that transaction pattern are updated. The disadvantage of this approach is that every time a new pattern of fraud occurs the entire network has to be re-trained.

In [1], the authors have also used a Bayesian learning approach. Bayesian Networks were also used in different comparative studies for detecting fraud in electronic transactions especially in credit card transactions. Bayesian Networks represents dependencies between variables of a probabilistic model, where each node represents a random variable and the arcs represent the relationship and dependencies between variables. In the fraud detection problem, the variables are the features or attributes that influence the transaction. These features were given as inputs. In the fraud detection problem, initially the network is unknown. To construct the Bayesian Network, the data has to be learned. Later from the graph that is constructed, the set of dependent variables to happen fraud is calculated. Bayesian Networks are more prominently used for classification problems. The network provides easy and fast training but is impacted when applied to newer instances.

In [5], in order to detect fraudulent transactions, a model based on Hidden Markov Model (HMM) has been proposed by the authors. Initially, the model is trained with the normal behavior or spending pattern of the cardholder. To identify the spending behavior of the cardholder or customer, K-means clustering algorithm is used. The spending profile is characterized as low, medium and high. The HMM is used to find out any deviation or variance in the spending patterns. HMM works just by remembering the customer spending behavior. Therefore, if the HMM rejects an incoming transaction with sufficiently high probability, then the transaction is considered fake or fraudulent. The model will generate an alarm to stop the transaction if any deviation or variance is observed from the spending patterns of the cardholder.

In [6], the authors have described a fusion approach based on Dempster-Shafer theory and Bayesian learning. The fraud detection system has four components: Rule-based filter, Dempster-Shafer adder, transaction history database and Bayesian Learner. Here, the rule-based filter calculates the suspicion level of each of the transactions with respect to its deviation from the genuine or good pattern. The Dempster-Shafer adder combines several such evidences and computes or calculates an initial belief. The initial beliefs are then combined to obtain an overall belief by applying the Dempster-Shafer theory. The transaction is judged as fraud or genuine based on the initial belief. Once the transaction is found to be suspicious, the belief is further enhanced or weakened by checking its similarity with fraud or genuine using Bayesian learning. The system was efficient but was highly expensive and processing power was low.

Many of the researches make use of a fuzzy-based clustering approach. Fuzzy clustering is basically used to find the deviation in the spending patterns of users. Since the clustering boundaries in such problems have an overlapping nature, the use of hard clustering is avoided. To reduce such limitations, fuzzy clustering has been used, so that individual data points may belong to more than one cluster. This helps to include all probabilities and find out the deviation from the actual pattern [7].

There are also some expert systems, where rules are generated from the data by a human expert and those are stored into the rule-based systems as IF-THEN rules. Fuzzy logic also addresses the problem of uncertainty of inputs by allowing the input definitions in the form of linguistics variables (e.g. small, medium and high) [8].

III. PROPOSED SYSTEM: FRAUD DETECTION USING ADAPTIVE NEURO-FUZZY INFERENCE SYSTEM

In this paper, a fraud detection system based on Adaptive Neuro Fuzzy Approach (ANFIS) is proposed. This technique allows to utilize the advantage of both neural networks and fuzzy inference system. The advantage of self-learning from neural networks and the advantage of specifying or generating

fuzzy rules and inferences based on the newer instances of fraud are combined together.

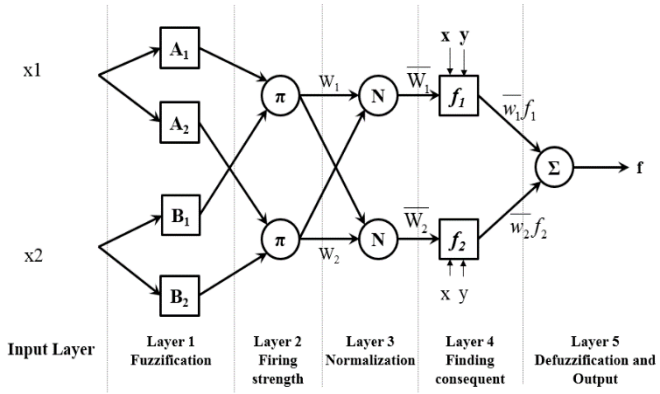

Fig. 2. ANFIS Architecture [9].

In [9], the Adaptive Neuro-Fuzzy approach was initially presented by Jang. ANFIS uses fuzzy if-else rules to improve the performance of complex systems. In these systems, prior knowledge can be given as rules to help train the system faster. It can easily incorporate both numeric and linguistic rules for problem solving. This technique provides fast and accurate learning.

Fig. 2 describes the basic architecture of ANFIS system, the circle indicates a fixed node, whereas a square indicates an adaptive node. The architecture is five layered [9, 10]. The ANFIS system uses a hybrid learning algorithm, which combines least square estimator and gradient descent method. Each epoch here consists of a forward pass and a backward pass. The forward pass adjusts the consequent parameters while the antecedent parameters remain fixed. In the backward pass, the antecedent parameters are tuned and the consequent parameters remain fixed. The ANFIS output is calculated by employing the consequent parameters found in the forward pass.

The output error is used to adapt the antecedent or premise parameters by means of a standard back propagation algorithm [9]. This hybrid approach proves to be highly efficient in training ANFIS systems [11].

A. Working Principle

The system will be initially trained with the prior knowledge of the expert or based on the training data. In this phase, the rule sets will be designed based on the previous instances of fraud, cases registered regarding loss of cards or credentials etc. The inputs to the system will be the inputs that characterize each transaction.

When an individual user comes into the e-commerce system and enters his credentials, his spending patterns, time when the purchase is made, his purchase history, credit card number and status/history of earlier purchases, location etc. are some of the attributes to be considered. Fig. 3 shows the proposed model for fraud detection using Adaptive Neuro-Fuzzy Inference System. Based on the expert knowledge, the Rule Base will be designed which will be the combination of different attributes that influence fraud in the previous cases.

The inputs will be preprocessed and it will be fed to the Inference Engine which will use Adaptive Neuro Fuzzy Inference System. The rules in the Rule Base are fed to the inference engine. The rules can be automatically generated where the system takes the permutation and combination of all the inputs present or can be manually fed. The Inference Engine checks the inputs that come into the system based on these rules. The network can learn or adapt itself, it learns by adjusting the weights assigned to inputs. The weights assigned to the inputs will be on the basis of how much it influences the fraudulent transaction. Depending on the fraudulence in transaction the rulebase will be updated. Similarly, for every successful transaction the rulebase will be updated. The fraudulent customers can be blocked by some software systems which can be later on integrated with the proposed system.

B. Fraud Detection

When a customer comes in to the website, his credentials are taken as the input for that transaction. The inputs are initially normalized and fed into the inference engine. The inference engine compares the input parameters to the already existing rules sets and creates an inference based on the range of influence of the inputs and the Euclidean distance whether the class is good or bad. The class whose distance is minimum is detected.

Detecting fraud does not always mean that the transaction under consideration is fraudulent. Therefore, in this study, an adaptive Neuro-fuzzy approach is used, where the network predicts fraud based on the earlier history as well as the newer instances of fraud. It takes the permutation and combination of all the attributes present. To ensure that the network is working correctly, the network was trained with some values from the dataset and then during the testing phase, it was provided values which were not used during the training phase. In such cases also, the network accurately predicted fraud, thus preventing the genuine customers from getting rejected.

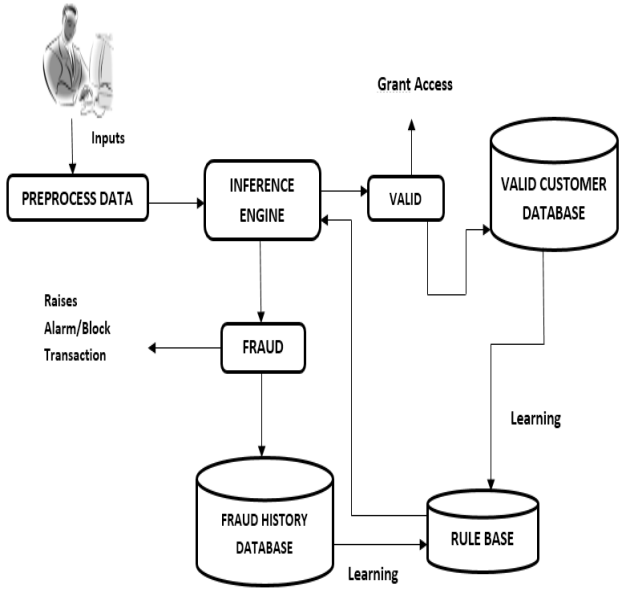

Fig. 3. Proposed model for fraud detection using Adaptive Neuro-Fuzzy Inference System.

IV. CASE STUDY

This section describes a case study where three different computational intelligence techniques has been applied on the same dataset to detect fraud.

A. Dataset Overview

Since the dataset related to online banking transactions are highly confidential, it is difficult to get them for academic research. Similarly, getting a dataset which is correctly classified is difficult. So, a similar dataset which is popularly known as German Credit Data [12] is used for this study. It is a dataset which classifies people described by a set of attributes, as good or bad credit risks [13]. This dataset is available on the UCI Machine Learning Repository. The dataset contains 20 attributes and 1000 instances and also a classification attribute for each instance. This dataset has been earlier used to study fraud in credit card transactions and that is the reason why the same was chosen for this study.

The purpose of this work is to analyze the classification done in the dataset using the features or attributes that characterize each of the instances or transactions as good or bad, by applying computational intelligence techniques like Bayesian Networks, Neural Networks and Adaptive Neuro-Fuzzy Inference System over the same data and to compare the performance of each of them.

In total, 21 attributes were used for the study, of which the first 20 attributes are the features or attributes that characterize the transaction and the last attribute is the classification or class: Good or Bad.

Once the system is trained entries can be added manually to check if it is belonging to class Good or Bad. And these entries will again help the system to adapt to newer instances of fraud.

B. Methodology

The process starts with the characterization of the dataset by removing items with lower significance and categorizing some numeric values. The same methodology is used for all the three techniques. The relative gain of each of the attributes was checked using InfoGain [14].

After this process the training and the testing sets have to be defined. But to increase the accuracy of ANFIS and Neural Networks, during the preprocessing phase itself, a method is used, wherein the entries are shuffled and given to the network, such that, every time the network is trained, the network gets a different set of training samples thus increasing its efficiency.

To evaluate the fraud detection techniques different environments were used.

- The ANFIS model was built using subtractive clustering [15, 16], which generates Fuzzy Inference rule structure from data, with radii or the clusters range of influence 0.3. Since subtractive clustering is used there is need to give both input and output data as input arguments. So the first 20 attributes (inputs) and the last one attribute i.e. the 21st attribute (desired output) is given in a matrix form as input arguments. The optimization method is selected as 0 i.e. back propagation.
- The second model built was based on Neural Network. The network is a multi-layer perceptron with one hidden layer and the output layer consisting of one neuron. The activation function used is hyperbolic tangent sigmoidal transfer function between input and the hidden layer and between hidden and the output layer. For the training stage the algorithm used was Scaled conjugate gradient back propagation algorithm for pattern recognition and classification.
- For the model based on Bayesian Network, K-fold cross validation was used for dividing the dataset. The network analyses the relationship between attributes and calculates the probability of fraud from a set of dependent variables. Hill climbing algorithm [1, 17] was used to search for the network. Also, the maximum number of father nodes is set to 1 as the number of our response variable is also 1.

The performance of the proposed system is evaluated by comparing the results of the proposed system against other two techniques using the parameters Root Mean Square Error (RMSE), Mean Squared Error (MSE) and Mean Absolute Error (MAE).

C. Results

Table I summarizes the results of the techniques previously mentioned. The proposed system based on ANFIS outperforms the other two techniques and indeed proves to be efficient and best among the three for analyzing fraud.

One other thing to be noticed is the training time required is comparatively less and it eliminates the process of retraining, since the network itself can adapt to newer instances of fraud. Even though a new instance of fraud comes into the system, it classifies it correctly based on the learning from the previous instances and the combination of rules from the inference network.

From Table I it can be clearly seen that ANFIS is a better technique for detecting fraud as compared to the other two machine learning techniques under consideration. The proposed system takes a lot more combinations of input features and gives correct prediction thus avoiding false alarms and rejecting genuine customers.

TABLE I. RESULTS FOR DIFFERENT MACHINE LEARNING TECHNIQUES ON THE GERMAN CREDIT DATA

	Machine Learning Techniques		
	ANFIS	***Neural Network***	***Bayesian Network***
RMSE	0.0036255	0.46904	0.4299
MSE	1.3144e-05	0.22	0.1848
MAE	0.0026	0.2000	0.3263

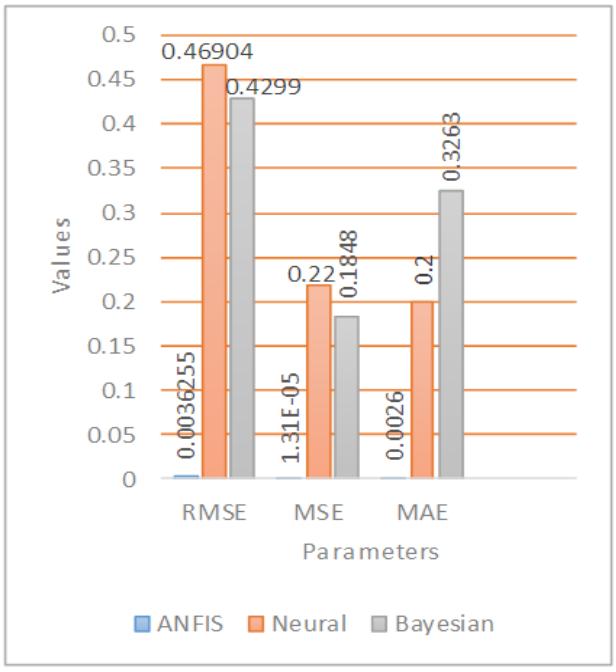

Fig. 4. Machine Learning techniques by parameters and values.

Fig. 4 shows the comparison of the three machine learning techniques over the German Credit Data. It can be clearly seen that ANFIS has lower values for all the three parameters. The lower values indicate that the ANFIS model predicts fraud with better accuracy and a smaller error occurs between actual and predicted values.

Similarly, the error mean and error standard deviation for proposed model is -0.0006274 and 0.003588 whereas error mean and error standard deviation for neural is 0.2 and 0.4264 respectively.

For ANFIS, the RMSE and MSE when 100 samples were considered was around 0.0036255 and 1.3144e-05 respectively and RMSE and MSE when 400 samples were considered was around 0.0020 and 4.0237e-06 respectively. This can be clearly visualized in Fig. 5 and Fig. 6 respectively. For Neural Networks, the RMSE and MSE when 100 samples were considered was around 0.46904 and 0.22 respectively and RMSE and MSE when 400 samples were considered was around 0.8544 and 0.73 respectively. Thus, it can be inferred that, in ANFIS systems, as the number of training samples increases the error decreases and the results are more refined. Whereas for neural network the error increases with the increase in training samples.

The training time required for the proposed system increases as the number of samples increases. But the time taken for testing after the initial training is comparatively very less. This is because, it uses the rule sets which were already designed during the training phase and hence the system does not have to re-frame and re-compute it. For e.g. for 100 training samples the network takes about 6 to 7 seconds for training whereas the time taken for testing is only 1 to 2 seconds. The training and testing errors and the time taken decreases over time, as every time the inputs to the network are shuffled and the network gets trained efficiently.

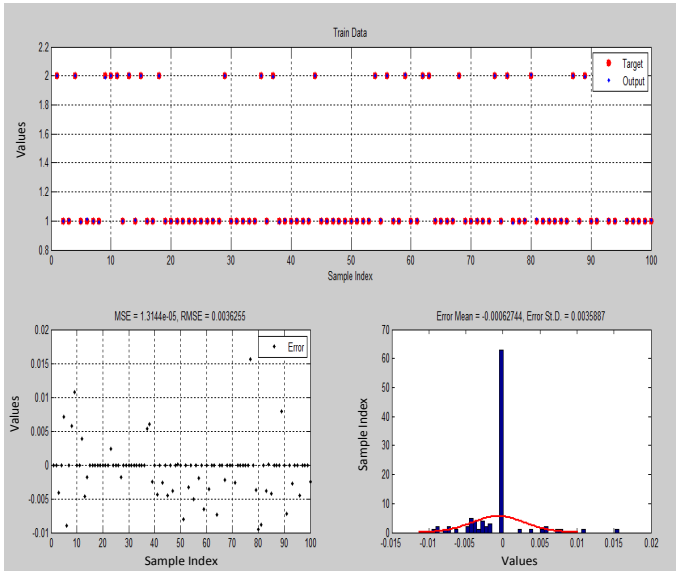

Fig. 5. ANFIS Output for 100 training samples.

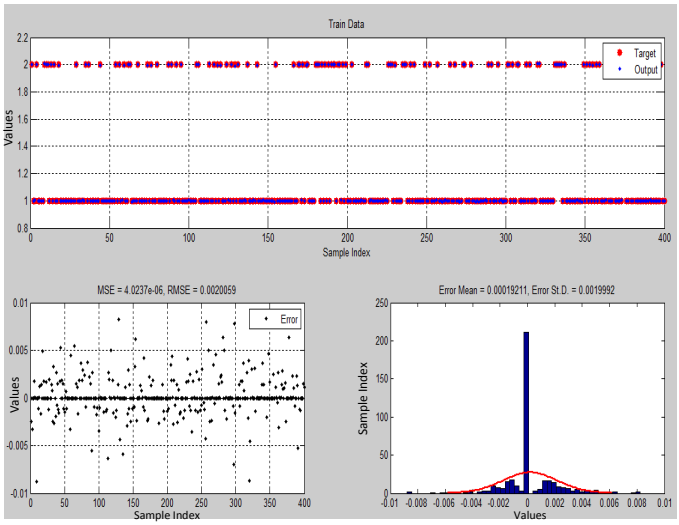

Fig. 6. ANFIS Output for 400 training samples.

During the testing phase, newer instances which were not present in the training data set were given to the trained network and that was also predicted accurately by the network with an RMSE of 0.00447. This clearly indicates that, this approach will prove excellent in fraud detection systems, wherein the major aim is to avoid rejecting genuine customers and to predict fraudulent customers as quickly as possible.

Also, the ease of implementation of the three techniques can be considered as a comparison factor. Bayesian networks are simple to implement but it faces difficulty when newer instances come in. Similarly, in neural network every time the network is retrained, the weights assigned to the nodes have to be changed according to the error that is back propagated. So, the implementation and maintenance of neural network based fraud detection system is difficult and takes time. ANFIS systems on the other hand takes time during initial implementation, but once trained it can adapt itself and there is no need for back tracking or retraining.

V. CONCLUSION

This work is presented as an analysis of computational intelligence techniques used for fraud detection in electronic transactions and also a new technique based on Adaptive Neuro-Fuzzy approach is proposed. The proposed technique was evaluated against other approaches based on Neural Network and Bayesian Networks by implementing them for comparison purpose over the same dataset. The results proved that, the proposed technique based on ANFIS would be a better choice for fraud detection, to improve the efficiency of fraud detection process.

One challenge of this research was the availability of an actual dataset. However, since the network can adapt to the changes it can be implemented over any dataset. The only constraint is that, the data has to be provided to the network in the way it requires it, i.e. in a matrix form. Once this is done, the same technique can be used to predict different types of frauds like insurance frauds, credit frauds to name a few.

Although, the proposed system gives good results with large number of inputs, future work will concentrate on reducing the number of inputs required to predict fraud, i.e. input reduction. The future work will also concentrate on studying and analyzing the different input reduction techniques and to check if there is a significant difference in the results.

REFERENCES

[1] E. Caldeira, G. Brandao and A. C. M. Pereira, "Fraud Analysis and Prevention in e-Commerce Transactions", 9th Latin American Web Congress, Ouro Preto, pp. 42-49, 2014.

[2] E. Caldeira, G. Brandão, H. Campos and A. Pereira, "Characterizing and Evaluating Fraud in Electronic Transactions", Eighth Latin American Web Congress, Cartagena de Indias, pp. 115-122, 2012.

[3] S. Parvinder and M. Singh, "Fraud Detection by Monitoring Customer Behavior and Activities", International Journal of Computer Applications, vol. 111, no. 11, pp. 23-32, 2015.

[4] J. J.-S. Roger, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence", 1997.

[5] A. Srivastava, A. Kundu, S. Sural and A. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model", in IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37-48, 2008.

[6] K.K. Tripathi and R. Lata, "Hybrid Approach for Credit Card Fraud Detection", International Journal of Soft Computing and Engineering, vol. 3, no. 4, pp. 8-11, 2013.

[7] T. K. Behera and S. Panigrahi, "Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network", Second International Conference on Advances in Computing and Communication Engineering, Dehradun, pp. 494-499, 2015.

[8] S. Sorournejad, Z. Zahra, R. E. Atani, and A. H. Monadjemi, "A Survey of Credit Card Fraud Detection Techniques: Data and Technique Oriented Perspective", https://arxiv.org/abs/1611.06439, 2016.

[9] J. S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system", IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, 1993.

[10] J. S. R. Jang and C.-T. Sun, "Neuro-fuzzy modeling and control", in Proceedings of the IEEE, vol. 83, no. 3, pp. 378-406, 1995.

[11] A. Al-Hmouz, J. Shen, R. Al-Hmouz and J. Yan, "Modeling and Simulation of an Adaptive Neuro-Fuzzy Inference System (ANFIS) for Mobile Learning", in IEEE Transactions on Learning Technologies, vol. 5, no. 3, pp. 226-237, 2012.

[12] Lichman, Moshe, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2013.

[13] https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data).

[14] Weka Machine Learning Project, http://www.cs.waikato.ac.nz/~ml/weka.

[15] https://in.mathworks.com/help/fuzzy/genfis2.html

[16] L. Edwin and E. Klement, "Online adaptation of Takagi-Sugeno fuzzy inference systems", Proceedings of CESA, 2003.

[17] https://en.wikibooks.org/wiki/Artificial_Intelligence/Search/Iterative_Improvement/Hill_Climbing.