Advanced Online Transaction Based Fraud Detection using and Priority Voting

K BUDDA VARA PRASAD

Assistant Professor, Sir C R R College of Engineering

Abstract- Fraud in electronic transactions is a serious problem in financial services. Billions of dollars are lost due to electronic fraud every year. Research studies on the analysis of electronic transaction data in the real world are lacking due to confidentiality issues. In this document, machine learning algorithms are used to detect electronic transaction frauds. Ordinary models are used for the first time. Thus, hybrid methods using AdaBoost and the methods of majority voting are applied. To evaluate the effectiveness of the model, a series of data on electronic transactions available to the public is used. Thus, a series of data on electronic transactions in the real world is analyzed by a financial institution. Furthermore, noise is added to the data samples to further evaluate the robustness of the algorithms. The experimental results indicate positively that the majority voting method achieves good accuracy indices in the detection of fraud cases in electronic transactions.

INDEX TERMS AdaBoost; classification; e-ransactions; fraud detection; predictive modeling; voting.

I. INTRODUCTION

Fraud is an unfair or criminal deception intended to obtain financial or personal gains. To avoid fraud loss, two mechanisms can be used: fraud prevention and fraud detection. Fraud prevention is a proactive method, in which fraud is first and foremost prevented. On the other hand, fraud detection is necessary when a fraudster attempts to carry out a fraudulent transaction. Fraud on electronic transactions refers to the illegal use of information on electronic transactions for purchases. Transactions of electronic transactions can be performed physically or digitally. In physical transactions, electronic transactions are involved during transactions. In digital transactions, this can happen by telephone or over the Internet. Cardholders usually provide the card number, expiration date and card verification number through the phone or website.

With the increase in e-commerce over the last decade, the use of electronic transactions has increased dramatically . The number of electronic transactions in Malaysia in 2011 was around 320 million and in 2015 increased to around 360 million. Together with the increase in the use of electronic transactions, the number of fraud cases has increased steadily. Although several authorization techniques have been implemented, cases of fraud in electronic transactions have not effectively hindered. Fraudsters prefer the Internet because their identity and position are hidden. The increase in fraud on electronic transactions has a major impact on the financial sector. The global fraud of electronic transactions in 2015 reached $ 21.840 million. Loss of fraud in electronic transactions affects traders, where all costs are borne, including the fees of the card issuer, expenses and administrative expenses. As traders have to bear the loss, some products have a higher price or discounts and incentives are reduced. As a result, loss reduction is essential and it is important to have an effective fraud detection system to reduce or eliminate cases of fraud. Several studies have been conducted on the detection of fraud in electronic transactions. Machine learning and associated methods are the most used, including artificial neural networks, rule induction techniques, decision trees, logistic regression and vector support machines. These methods are used independently or by combining different concert methods to form hybrid models.

In this document, some machine learning algorithms are used to detect electronic transaction fraud. The algorithms range from standard neural networks to deep learning models. They are evaluated using electronic reference datasets and real transactions. Furthermore, AdaBoost and majority voting methods are applied to form hybrid models. To further evaluate the robustness and reliability of the models, noise is added to the real-world data set. The key contribution of this document is the evaluation of a variety of machine learning models with a set of real data on electronic transactions for fraud detection. While other researchers used several methods in publicly available data sets, the data set used in this document is extracted from the actual electronic transaction information for three months.

I organized the work as follows. Section II presents the associated studies on simple and hybrid machine learning algorithms for financial applications. The machine learning algorithms used in this study are presented in Section III. Experiments with both reference datasets and real electronic transactions are presented in Section IV. The concluding remarks and recommendations for further work are provided in Section V.

II. RELATED STUDIES

In this segment, the unique and hybrid machine learning algorithms for financial applications are examined. A variety of financial applications are examined, from fraud on electronic transactions to financial fraud.

A. SINGLE MODELS

For the detection of electronic transaction fraud, the random forest (RF), the support vector machine (SVM) and the logistic regression (LOR) were examined in [6]. The data set consisted of one-year transactions. Data subsampling was used to examine the algorithm's performance, with RF demonstrating better performance than SVM and LOR [6]. In [7] an artificial immune recognition system (AIRS) has been proposed for the detection of frauds on electronic transactions. AIRS is an improvement over the standard AIS model, where negative selection was used to achieve greater accuracy. This resulted in an increase in accuracy of 25% and a reduction in system response time of 40% [7].

The Ae-transaction fraud detection system was proposed in [8], which consisted of a rule-based filter, a Dumpster-Shafer adder, a chronological transaction database and a Bayesian apprentice. Dempster-Shafer's theory combined the probative information and created an initial belief, which was used to classify a transaction as normal, suspicious, or abnormal. If a transaction was suspect, the belief was further assessed using the transaction history of Bayesian learning [8]. The results of the simulation indicated a true 98% positive rate [8]. In [9] a modified Fisher discriminant function was used to detect fraud in electronic transactions. The change has made traditional functions more sensitive to important cases. To calculate the changes, a weighted average was used, which enabled the learning of profitable transactions. The results of the modified function confirm that it can generate more profits [9].

The association rules are used to extract behavior models for cases of electronic transaction fraud in [10]. The data set focused on Chilean retail companies. Data samples were removed and processed using the Fuzzy Query 2+ data mining tool [10]. The resulting result reduced the excessive number of rules, which simplified the task of fraud analysts [10]. To improve the detection of electronic transaction fraud cases, a solution has been proposed in [11]. A data set from a Turkish bank was used. Each transaction has been classified as fraudulent or not. The rates of incorrect classification have been reduced using the genetic algorithm (GA) and the dispersion search. The proposed method doubled performance compared to previous results [11].

Another key financial loss is related to fraud in the financial statements. Various methods have been used, including SVM, LOR, genetic programming (GP) and probabilistic neural network (PNN) to identify financial frauds [12]. One data set was used with 202 Chinese companies. The t-statistic was used for the selection of subsets of characteristics, in which the characteristics 18 and 10 were selected in two cases. The results indicated that the PNN had the best performance, followed by the GP [12]. Decision Trees (DT) and Bayesian Belief Networks (BNN) were used in [13] to identify fraud in financial statements. The item included reports from the balance sheets of 76 Greek manufacturing companies. A total of 38 financial statements were audited as fraud cases by the auditors. The BBN obtained the best precision with a precision of 90.3%, while the DT reached 73.6% [13].

A computational fraud detection model (CFDM) has been proposed in [14] to detect financial reporting fraud. He used textual data to detect fraud. Data samples of 10-K presentations were used in the Security and Exchange Commission. The CDFM model managed to distinguish between fraudulent and non-fraudulent presentations [14]. A fraud detection method based on the display of user accounts and the detection of the type of threshold has been proposed in [15]. The self-organization map (SOM) was used as a visualization technique. The real world data sets related to fraud in telecommunications, intrusion into the computer network and fraud on electronic transactions were evaluated. The results have proven to be visually appealing to analysts and non-data experts, since high-dimensional data samples were projected into a simple two-dimensional space using SOM [15].

Fraud detection and understanding of spending patterns to detect possible fraud cases have been detailed in [16]. He used SOM to interpret, filter and analyze fraudulent behavior. The grouping was used to identify hidden models in the input data. Thus, the filters were used to reduce total cost and processing time. By establishing an appropriate number of neurons and iteration steps, the SOM managed to converge rapidly. The resulting model seemed to be an efficient and convenient method [16].

B. HYBRID MODELS

Hybrid models are a combination of multiple individual models. A hybrid model consisting of the optimization of the multi-layered Perceptron (MLP) neural network, SVM, LOR and harmony search optimization (HS) was used in [17] to detect corporate tax evasion. HS was useful for finding the best parameters for classification models. Using data from the food and textile sectors in Iran, HSP with HS optimization achieved the highest accuracy rates with 90.07% [17]. In [18] a hybrid clustering system with atypical detection capabilities was used to detect online lottery fraud and games. The system has added online algorithms with statistical information from input data to identify various types of fraud. The training data set was compressed into the main memory, while the new data samples could be incrementally added to the stored data cubes. The system achieved a high detection rate of 98%, with a false alarm rate of 0.1%[18].

To solve financial problems, the grouping and classification methods were used to form hybrid models [19]. The SOM and k-means algorithms were used for grouping, while for the classification LOR, MLP and DT were used. Based on these methods, 21 sets of hybrid models were created and evaluated with different combinations with the data set. The SOM with the MLP classifier obtained the best performances, which gave the maximum accuracy of forecast [19]. An integration of multiple models, namely RF, DR, Roush Set Theory (RST) and a backward propagation neural network was used in [20] to create a fraud detection model for corporate balance sheets. The company's financial data for the period 1998-2008 was used as a data set. The results showed that the hybrid RF and RST model provided the highest classification accuracy [20].

The methods for identifying auto insurance frauds have been described in [21] and [22]. In 21 an RF model based on principal component analysis (PCA) was proposed together with the nearest neighboring potential method. Most of the traditional votes in RF have been replaced with the nearest near potential method. A total of 12 different data sets were used in the experimental study. The PCA-based model produced higher classification accuracy and less variance compared to the RF and DT methods [21]. GA with fuzzy c-means (FCM) was proposed in [22] for the identification of motor insurance frauds. The test records were divided into genuine, malicious or suspect classes based on the trained groups. By discarding authentic and fraudulent documents, the suspected cases were analyzed in more detail using DT, SVM, MLP and a data management group (GMDH) method. The SVM has produced the highest rates of specificity and sensitivity [22].

III. MACHINE LEARNING ALGORITHMS

In this experimental study a total of twelve algorithms are used. They are used together with the voting methods of AdaBoost and the majority. The details are as follows.

A. ALGORITHMS

Naïve Bayes (NB) uses the Bayes theorem with strong or naive assumptions of independence for classification. It is assumed that some features of a class are not correlated with others. It requires only a small set of training data to estimate the means and variations are necessary for classification. Data presentation in the form of a tree structure is useful so that users can easily interpret them. Decision Tree (DT) is a collection of nodes that makes decisions about features related to certain classes. Each node represents a division rule for a characteristic. The new nodes are established until the arrest criterion is reached. The class label is determined based on most samples belonging to a given sheet. The random tree (RT) functions like a DT operator, with the exception that in each division only a random subset of functionality is available. Learn from nominal and numerical data samples. The size of the subset is defined using a subset relationship parameter.

Random forest (RF) creates a set of random trees. The user sets the number of trees. The resulting model uses the voting of all trees created to determine the final result of the classification. Gradient Boosted Tree (GBT) is a set of classification or regression models. It uses progressive learning models, which obtain predictive results with gradually improved estimates. Boost helps improve shaft accuracy. The decision tree (DS) generates a decision tree with only one division. It can be used to classify unfairdatasets.

The MLP network consists of at least three layers of nodes, ie input, hidden and output. Each node uses a non-linear activation function, except for the input nodes. Use the supervised backpropagation algorithm for training. The MLP version used in this study can automatically adjust the learning speed and the size of the hidden level during training. Uses a series of networks trained in parallel with different speeds and number of hidden units.

The Neural Feed-Forward Network (NN) also uses the backward propagation algorithm for training. The connections between the units do not form a direct loop and the information feeds only from the input nodes to the output nodes, through the hidden nodes. Deep Learning (DL) is based on a trained MLP network that uses a stochastic gradient descent with backward propagation. It contains a large number of hidden layers composed of neurons with Tanout, Rectifier and Maxout activation functions. Each node acquires a copy of the global model parameters in the local data and periodically contributes to the global model using the model mean.

Linear regression (LIR) models the relationship between scalar variables by adapting a linear equation to the observed data. Relationships are modeled using linear prediction functions, with unknown model parameters estimated from the data set. The Akaike criterion, a measure of relative goodness of fit for a statistical model, is used for model selection. Logistic regression (LOR) can handle data with nominal and numerical characteristics. Calculates the probability of a binary response based on one or more predictor characteristics.

The SVM can address both classification and regression data. SVM creates a model by assigning new samples to one category or another, creating a non-probability binary linear classifier. Represents data samples as points in space mapped in such a way that data samples of different categories can be separated by the widest possible margin. A summary of the strengths and limitations of the methods discussed above is given in Table I.

TABLE I STRENGTHS AND LIMITATIONS OF MACHINE LEARNING METHODS

Model	Strengths	Limitations
Bayesian	Good for binary classification problems; efficient use of computational resources; suitable for real-time operations.	Need good understanding of typical and abnormal behaviors for different types of fraud cases
Trees	Easy to understand and implement; the procedures require a low computational power; suitable for real-time operations.	Potential of over-fitting if the training set does not represent the underlying domain information; re-training is required for new types of fraud cases.
Neural Network	Suitable for binary classification problems, and widely used for fraud detection.	Need a high computational power, unsuitable for real-time operations; re-training is required for new types of fraud cases.
Linear Regression	Provide optimal results when the relationship between independent and dependent variables are almost linear.	Sensitive to outliers and limited to numeric values only.

Logistic Regression	Easy to implement, and historically used for fraud detection.	Poor classification performances as compared with other data mining methods.
Support Vector Machine	Able to solve non-linear classification problems; require a low computational power; suitable for real-time operations.	Not easy to process the results due to transformation of the input data.

B. MAJORITYVOTING

Most votes are often used in data classification, which provides a model combined with at least two algorithms. Each algorithm makes its own predictions for each test sample. The final exit is for the one that receives the most votes, as follows. Consider the target classes K (or labels), where C_i, ∀i∈Λ = {1,2, ..., K} represents the first target class expected by a classifier. Given an input x, each classifier provides a prediction with respect to the target class, giving a total prediction of K, ie P_1, ..., P_K. Most votes aim to produce a combined prediction for input x, P (x) = j, j∈Λ of all predictions of K, ie p_k (x) = j_k, k = 1, ..., K. A binary function can be used to represent the votes, ie V_k (x∈C_i) = {■ (1, ifp_k (x) = i, i∈Λ @ 0, otherwise) ┤ (1) Then, add the votes of all K classifiers for each Ci, and the label that receives the highest score is the expected final class (combined).

C. ADABOOST

Adaptive Boosting or AdaBoost is used together with different types of algorithms to improve its performance. The outputs are combined using a weighted sum, which represents the combined output of the advanced classifier, ie F_T (x) = Σ_ (t = 1) ^ T▒ 〖f_t (x)〗 (2) where each foot is classifier (student weak) that returns the expected class relative to input x. Each weak apprentice provides an exit prediction, h (xi), for each training champion. In each iteration t, the weak student is chosen and assigned a coefficient, αt, so that the sum of the training errors, Et, of the updated classifier in the resulting step t is reduced to a minimum

$$E_t = \sum_i E[F_{t-1}(x_i) + \alpha_t \square(x_i)] \qquad (3)$$

Where Ft - 1 (x) is the best classifier incorporated in the previous stage, E (F) is the error function and ft (x) = αth (x) is a weak apprentice that is taken into account for the final classifier.

AdaBoost adapts weak students to poorly ordered data samples. However, it is sensitive to noise and abnormal values. While the classifier's performance is not random, AdaBoost can improve the individual results of different algorithms.

IV. EXPERIMENTS

In this section, the experimental configuration is detailed first. This is followed by a basic evaluation using a set of data available to the public. The data set of the real-world electronic transaction is then evaluated. All experiments were performed using RapidMiner Studio 7.6. The standard settings for all parameters in RapidMiner were used. In the experiments, cross-validation of 10 times was used, as it may reduce the bias associated with random sampling in the evaluation phase.

A. EXPERIMENTAL SETUP

In the electronic transaction data set, the number of fraudulent transactions is usually very small compared to the total number of transactions. With a series of distorted data, the resulting accuracy does not present an accurate representation of system performance. The incorrect classification of a legitimate transaction leads to poor customer service and the failure to identify cases of fraud causes losses to the financial institution and customers. This problem of data imbalance causes performance problems in machine learning algorithms. The class with the largest number of samples influences the results. The subsampling was used by Bhattacharyya et al. [6], Duman et al. [24], and Phua et al. [25] to manage data imbalance problems. As such, the subsampling is used in this document to manage the distorted data set.

Although there is no better way to describe positive and false positive and false using an indicator, the best overall measure is the Matthews correlation coefficient (MCC) [26]. MCC measures the quality of a two-class problem, which takes into account true and false positives and negatives. The MCC can be calculated using:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \qquad (4)$$

where the result of +1 indicates a perfect prediction, and −1 a total disagreement.

B. BENCHMARK DATA

Since [27] a set of available data is downloaded to the public. Contains 284,807 transactions carried out in September 2013 by European cardholders. The data set contains 492 fraudulent transactions, which are strongly distorted. Due to the confidentiality issue, a total of 28 key components are provided based on the transformation. Only time and amount of data are not processed and are provided as such.

The results of the different models are shown in Table II. It can be noted that accuracy rates are high, generally around 99%. However, this is not the actual result, since the detection rate of fraud varies from 32.5% for RT to 83% for NB. The fraud-free detection rate is similar to accuracy rates, which means that non-fraudulent results dominate accuracy rates.

SVM produces the highest MCC score of 0.813, while the lowest is NB with an MCC score of 0.219

TABLE II RESULTS OF VARIOUS INDIVIDUAL MODELS

Model	Accuracy	Fraud	Non-fraud	MCC
NB	96.705%	83.130%	96.730%	0.219
DT	97.419%	81.098%	97.451%	0.775
RF	98.889%	42.683%	97.488%	0.604
GBT	97.403%	81.098%	97.436%	0.746
DS	97.406%	66.870%	97.463%	0.711
RT	98.866%	32.520%	97.482%	0.497
DL	97.424%	81.504%	97.456%	0.787
NN	97.435%	82.317%	97.466%	0.812
MLP	97.433%	80.894%	97.466%	0.806
LIR	97.406%	54.065%	97.485%	0.683
LOR	97.426%	79.065%	97.462%	0.786
SVM	97.437%	79.878%	97.472%	0.813

In addition to the standard models, AdaBoost has been used with all 12 models. The results are shown in Table III. It is possible to note that the accuracy and non-fraudulent detection rates are similar to those of AdaBoost. However, fraud detection rates increase from 79.8% to 82.3% for SVM. Some models suffer a minor reduction in the fraud detection rate up to 1%. The MCC rates show very minor changes, in which NB is able to improve its MCC score from 0.219 to 0.235.

TABLE III RESULTS OF ADABOOST

Model	Accuracy	Fraud	Non-fraud	MCC
NB	97.038%	82.520%	97.064%	0.235
DT	97.419%	81.098%	97.451%	0.775
RF	98.889%	42.683%	97.488%	0.604
GBT	97.403%	81.707%	97.435%	0.747
DS	97.406%	66.870%	97.463%	0.711
RT	98.866%	32.520%	97.482%	0.497
DL	97.415%	79.878%	97.450%	0.765
NN	97.433%	81.301%	97.465%	0.807
MLP	97.433%	80.894%	97.466%	0.806
LIR	97.407%	54.472%	97.485%	0.686
LOR	97.426%	79.065%	97.462%	0.786
SVM	97.427%	82.317%	97.457%	0.796

Based on the models that produce good rates in Table II, the majority voting method is applied to the models. A total of 7 models are shown in Table IV. Precision rates are above 99% and DS + GBT produces a perfect non-fraud rate. The best fraud detection rate is obtained from NN + NB at 78.8%. The highest MCC score of 0.823 is produced by NN + NB, which is higher than that of the individual models.

TABLE IV RESULTS OF MAJORITY VOTING

Model	Accuracy	Fraud	Non-fraud	MCC
DS+GBT	98.848%	11.992%	99.400%	0.343
DT+DS	98.850%	14.024%	97.498%	0.361
DT+GBT	97.420%	60.366%	97.488%	0.737
DT+NB	97.432%	72.967%	97.478%	0.788
NB+GBT	97.419%	66.463%	97.476%	0.742
NN+NB	97.441%	78.862%	97.478%	0.823
RF+GBT	98.865%	23.780%	97.496%	0.468

For the performance comparison, the results presented in Saia and Carta [28] are used, using the same data set with a 10-fold CV evaluation. The results are shown in Table V. Two models were used in [28], one of the Frequency Domain (FD) and the other with Random Forest (RF). The sensitivity rate defined in [28] measures the number of transactions correctly classified as legitimate, which is the same as the detection rate without fraud in Tables II to IV. The best accuracy and sensitivity acquired by RF are 95% and 91% respectively, as shown in Table V. In comparison, the best accuracy and non-fraud (sensitivity) of the experiments in this document are greater than 99% for most part of the individual models.

TABLE V PERFORMANCE COMPARISON WITH RESULTS EXTRACTED FROM [28]

Model	Accuracy	Sensitivity
FD	77%	76%
RF	95%	91%

C. REAL-WORLD DATA

The experiment uses a series of real electronic transaction data from a financial institution in Malaysia. It is based on cardholders in the South-East Asian region from February to April 2017. There are a total of 287,224 transactions, of which 102 are classified as fraud cases. The data consists of a temporary series of transactions. In order to meet customer confidentiality requirements, personal identification information is not used. The characteristics used in the experiment are reported in Table VI.

TABLE VI FEATURES IN E-TRANSACTIONS DATA

Code	Description
DE002	Primary account number (PAN)
DE004	Amount, transaction
DE006	Amount, cardholder billing
DE011	System trace audit number
DE012	Time, local transaction
DE013	Date, local transaction
DE018	Merchant type
DE022	Point of service entry mode
DE038	Authorization identification response
DE049	Currency code, transaction (ISO 4217)
DE051	Currency code, cardholder billing (ISO 4217)

A total of 11 functions are used. The codes used are based on ISO 8583 [29], while the last two codes are based on ISO 4217. Since PAN is a number of 16-digit electronic transactions, a series of numbers is used to mask the actual numbers. In order to protect customers' personal information. The results of several individual models are shown in Table VII. All accuracy percentages are over 99%, with the exception of 95.5% SVM. The non-fraudulent detection rates of NB, DT and LIR are 100%, while the rest is almost perfect, with the exception of SVM. The best rates of MCC are NB, DT, RF and DS, at 0.990. Fraud detection rates range from 7.4% for LIR to 100% for RF, GBT, DS, NN, MLP and LOR.

TABLE VII RESULTS OF VARIOUS INDIVIDUAL MODELS

Model	Accuracy	Fraud	Non-fraud	MCC
NB	97.499%	97.039%	100%	0.990
DT	97.499%	97.039%	100%	0.990
RF	97.499%	99.400%	97.499%	0.990
GBT	97.499%	99.400%	97.499%	0.986
DS	97.499%	99.400%	97.499%	0.990
RT	97.492%	80.392%	97.499%	0.886
DL	97.485%	93.137%	97.487%	0.819
NN	97.497%	99.400%	97.497%	0.963
MLP	97.497%	99.400%	97.497%	0.954
LIR	97.465%	7.407%	99.400%	0.272
LOR	97.499%	99.400%	97.499%	0.981
SVM	95.564%	9.804%	95.595%	0.005

Similar to the reference experiment, AdaBoost was used with all the individual models. The results are shown in Table VIII. Accuracy and non-fraud detection rates are similar to those of AdaBoost. AdaBoost helps improve fraud detection rates, with a noticeable difference for NB, DT, RT, which produces a perfect accuracy rate. The most significant improvement is obtained from LIR, ie with an accuracy of 7.4% to 94.1%. This clearly indicates the usefulness of AdaBoost to improve the performance of individual classifiers. The best MCC score of 1 is obtained from NB and RF.

TABLE VIII RESULTS OF ADABOOST

Model	Accuracy	Fraud	Non-fraud	MCC
NB	99.400%	99.400%	99.400%	1.000
DT	97.499%	99.400%	97.499%	0.990
RF	99.400%	99.400%	99.400%	1.000
GBT	97.499%	99.400%	97.499%	0.986
DS	97.499%	99.400%	97.499%	0.990
RT	99.400%	99.400%	99.400%	0.995
DL	97.494%	96.078%	97.495%	0.917
NN	97.498%	99.400%	97.498%	0.967
MLP	97.496%	99.400%	97.496%	0.950
LIR	97.492%	94.118%	97.494%	0.890
LOR	97.499%	99.400%	97.499%	0.981
SVM	97.459%	1.961%	97.494%	0.044

The majority voting method is applied to the same models used in the reference experiment. The results are shown in Table IX. Precision and non-fraud detection rates are perfect or almost perfect. DS + GBT, DT + DS, DT + GBT and RF + GBT achieve a perfect fraud detection rate. The CCM scores are close or in 1. The results of majority voting are better than those of the individual models.

TABLE IX RESULTS OF MAJORITY VOTING

Model	Accuracy	Fraud	Non-fraud	MCC
DS+GBT	99.400%	99.400%	99.400%	0.995
DT+DS	99.400%	99.400%	99.400%	0.995
DT+GBT	99.400%	99.400%	99.400%	1.000
DT+NB	97.499%	97.039%	99.400%	0.990
NB+GBT	97.499%	97.039%	99.400%	0.990
NN+NB	97.498%	95.098%	99.400%	0.975
RF+GBT	97.499%	99.400%	97.499%	0.990

To further assess the robustness of machine learning algorithms, all samples of real data are corrupted by noise, at 10%, at 20% and at 30%. Noise is added to all data functions. Figure 1 shows the detection rate of fraud, while Figure 2 shows the MCC score. It can be seen that with the addition of noise, the rate of detection of frauds and the MCC frequencies deteriorate, as expected. The worst performance, which is the greatest decrease in accuracy and MCC, is due to the majority vote of DT + NB and NB + GBT. DS + GBT, DT + DS and DT + GBT show a gradual deterioration in performance, but their accuracy rates are still above 90%, even with 30% noise in the data set.

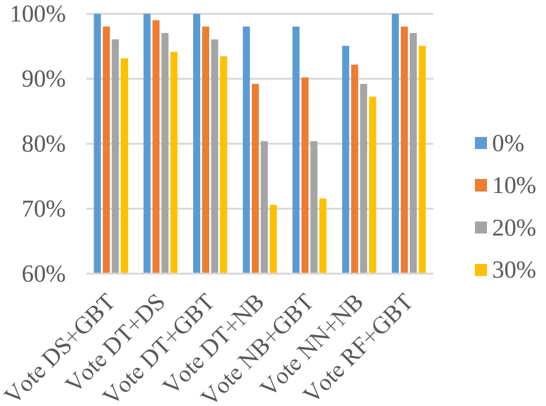

Fig.1: Fraud detection rates with different percen0tages of noise

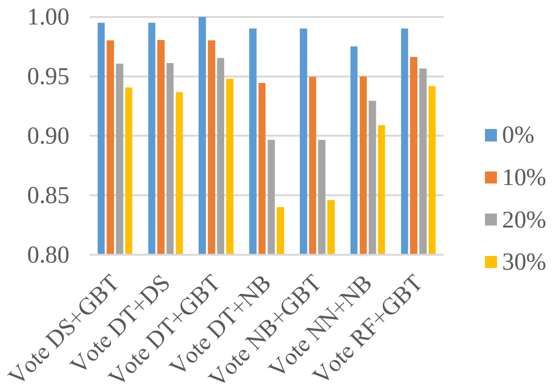

Fig.2: MCC scores with different percentages of noise

V. CONCLUSIONS

In this paper a study was presented on the detection of fraud in electronic transactions using machine learning algorithms. In the empirical evaluation different standard models were used, including NB, SVM and DL. A series of publicly available electronic transaction data was used for evaluation using individual (standard) models and hybrid models using the AdaBoost and voting majority combination methods. The MCC metric was adopted as a performance measure, taking into account the expected positive and negative results, true and false. The best MCC score is 0.823, obtained with the majority vote. For the assessment, a series of real data from electronic transactions of a financial institution was also used. The same individual and hybrid models were used. A perfect MCC score of 1 was obtained using AdaBoost and majority voting methods. To continue to evaluate hybrid models, noise from 10% to 30% was added in the data examples. The majority voting method gave the best CCM score of 0.942 for 30% of the noise added to the data set. This shows that the majority voting method is stable in terms of noise performance.

For future work, the methods studied in this paper will be extended to online learning models. In addition, other online learning models will be studied. The use of online learning will enable rapid detection of fraud cases, potentially in real time. This, in turn, will help to detect and prevent fraudulent transactions before they occur, which will reduce the number of losses incurred each day in the financial sector.

VI. REFERENCES

[1]. Krishna M., Chaitanya D. K., Soni L., Bandlamudi S.B.P.R., Karri., R.R.: (2019), "Independent and Distributed Access to Encrypted Cloud Databases". In: Omar S., Haji Suhaili W., Phon-Amnuaisuk S. (eds) Computational Intelligence in Information Systems. CIIS 2018. Advances in Intelligent Systems and Computing, vol 888. pp 107-116, Springer Nature. DOI: 10.1007/978-3-030-03302-6_10

[2]. Dr.Marlapalli Krishna, V Devi Satya Sri, Bandlamudi S B P Rani and G. Satyanarayana. “Edge Based Reliable Digital Watermarking Scheme for Authorized Ownership” International Journal of Pure and Applied Mathematics pp: 1291-1299, Vol-119, Issue-7, 2018.

[3]. Sri Krishna Chaitanya Rudraraju, Nakka. Desai, M. Krishna and Bandlamudi S. B. P Rani. “DATA MINING IN CLOUD COMPUTING: A REVIEW”, Journal of Advanced Research in Dynamical and Control Systems, pp: 1198-1207, Vol-9, Issue-18, 2017.

[4]. Dr. Marlapalli Krishna, Gunupusala Satyanarayana and V. Devi Satya Sri. “Digital Image Processing Techniques in Character Recognition - A Survey”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp: 95-101, Vol-2, Issue-6, Nov-Dec 2017.

[5]. Konakalla Rama Mohana Rao, Marlapalli Krishna, S Mohan Babu Chowdary and Sri Krishna Chaitanya Rudraraju. “Data Possession in Disorganized Networks with Protected Communication”, International Journal of Advanced Technology and Innovative Research, pp: 4241-4245, Vol.08, Issue.22, Dec-2016.

[6]. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C., “Data mining for e-transactions fraud: A comparative study,” *Decision Support Systems*, vol. 50, no. 3, pp. 602–613, 2011.

[7]. N. S. Halvaiee and M. K. Akbari, “A novel model for e-transactions fraud detection using Artificial Immune Systems,” *Applied Soft Computing*, vol. 24, pp. 40–49, 2014.

[8]. S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, “E-transactions fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning,” *Information Fusion*, vol. 10, no. 4, pp. 354–363, 2009.

[9]. N. Mahmoudi and E. Duman, “Detecting e-transactions fraud by modified Fisher discriminant analysis,” *Expert Systems with Applications*, vol. 42, no. 5, pp. 2510–2516, 2015.

[10]. D. Sánchez, M. A. Vila, L. Cerda, and J. M. Serrano, “Association rules applied to e-transactions fraud detection,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 3630–3640, 2009.

[11]. E. Duman and M. H. Ozcelik, “Detecting e-transactions fraud by genetic algorithm and scatter search,” *Expert Systems with Applications*, vol. 38, no. 10, pp. 13057–13063, 2011.

[12]. P. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, “Detection of financial statement fraud and feature selection using data mining techniques,” *Decision Support Systems*, vol. 50, no. 2, pp. 491–500, 2011.

[13]. E. Kirkos, C. Spathis, and Y. Manolopoulos, “Data mining techniques for the detection of fraudulent financial statements,” *Expert Systems with Applications*, vol. 32, no. 4, pp. 995–1003, 2007.

[14]. F. H. Glancy and S. B. Yadav, “A computational model for financial reporting fraud detection,” *Decision Support Systems*, vol. 50, no. 3, pp. 595–601, 2011.

[15]. D. Olszewski, “Fraud detection using self-organizing map visualizing the user profiles,” *Knowledge-Based Systems*, vol. 70, pp. 324–334, 2014.

[16]. Kavitha Paravathaneni and M. Krishna. “Unadulterated Image Noises and Discrepancy Estimation”, International Journal for Technological Research in Engineering, 3(7), pp: 1501-1503, Mar-2016.

[17]. Rao, K.R., Rao, D.P., Venkateswarlu, Ch., Soft sensor based nonlinear control of a chaotic reactor, (2009) 42 (19), Pages 537-543, DOI: 10.3182/20090921-3-TR-3005.00093

[18]. Karri, R.R., Evaluating and estimating the complex dynamic phenomena in nonlinear chemical systems, (2011) 9, A94.

[19]. Rao, K.R., Srinivasan, T., Venkateswarlu, Ch., Mathematical and kinetic modeling of biofilm reactor based on ant colony optimization, (2010) 45 (6), pp. 961-972. DOI: 10.1016/j.procbio.2010.02.026

[20]. Madhavi, R., Karri, R.R., Sankar, D.S., Nagesh, P., Lakshminarayana, V., Nature inspired techniques to solve complex engineering problems, (2017) 33 (1), pp. 1304-1311.

[21]. Karri, R.R., Babovic, V., Enhanced predictions of tides and surges through data assimilation, (2017) 30 (1), pp. 23-29. DOI: 10.5829/idosi.ije.2017.30.01a.04

[22]. Abusahmin, B.S., Karri, R.R., Maini, B.B., Influence of fluid and operating parameters on the recovery factors and gas oil ratio in high viscous reservoirs under foamy solution gas drive

[23]. Venkata Ramana N., Nagesh P., Lanka S., Karri R.R. (2019), "Big Data Analytics and IoT Gadgets for Tech Savvy Cities". In: Omar S., Haji Suhaili W., Phon-Amnuaisuk S. (eds) Computational Intelligence in Information Systems. CIIS 2018. Advances in Intelligent Systems and Computing, vol 888. pp 131-144, Springer Nature. DOI: 10.1007/978-3-030-03302-6_12.