



An Efficient Pre and Post filter-based anomaly detection technique for credit card fraud detection

Venkata Ratnam Ganji

Research Scholar, Department of Computer Science & Engineering
Acharya Nagarjuna University, Guntur, Andhra Pradesh, India

gvr.jntuk@gmail.com

Aparna Chaparala

Professor, Department of Computer Science & Engineering

R.V.R & J.C.College of Engineering

Guntur, Andhra Pradesh, India

Abstract

Machine learning plays a major role in the homogeneous and heterogeneous outlier detection process. Traditional outlier detection models are difficult to find the outliers in the heterogeneous uncertain datasets due to sparsity issue. In this work, hybrid pre-filter based outlier detection is implemented on the anomaly database in order to find the initial outliers for the post outlier detection process. In the post filtering process, a hybrid clustering based outlier detection model is proposed to predict the anomaly and non-anomaly classes. In this paper, a hybrid outlier detection based classification framework is proposed in order to eliminate the noise in the data for the class prediction process. In the classification problem, a cluster base classification model is implemented on the filtered data in order to optimize the false negative and false positive rate. Experimental results show that the present model has better false negative rate and error rate than the conventional approaches.

Keywords: Anomaly databases, machine-learning, support vector machine.

DOI Number: 10.14704/nq.2022.20.8.NQ44219

NeuroQuantology 2022 ;20(8):1987-2001

1987

1. Introduction

Credit scoring is still very limitedly known by the world. Although credit scoring started in 1941, until the turn of the century, the use of credit scoring was limited to client evaluation. The uses of credit scoring have expanded to different areas in the short period of life. The following section deals with the review of studies in various fields concerning the application of credit scoring. The first credit scoring application, who recognized the need for a numerical value that could quantify the applicant's creditworthiness and assist the lender in the process of subscription. The tool's

prominence has also increased with increased pressure on the banking sector to generate more credit facilities. Wider product ranges have forced creditors to customize their models. The use of credit scoring has historically been restricted to the assessment of home loans, mortgages, and consumer credit. Lending to small businesses however has been viewed by most lenders as a risky game. Banks preferred to lend with large valued collateral, accounted financials, and proven credit history to large enterprises that processed systematic future plans. Many small businesses find it very difficult to get loans from lenders. A major hindrance has



been experienced by lenders because they have had no place to grow. Lenders are looking for opportunities for their loan customers' repayment history to improve. Banks and other formal lenders are now targeting small businesses whose financial needs have been met primarily by local money lenders. Therefore, one of the lender's important priorities is to look out for creditworthy borrowers. Credit scoring has proven to be an effective tool for assessing the associated credit risk and reducing lending costs to small business borrowers. Many studies have contrasted neural networks with other methods of scoring. In a study that compared discriminant analysis, logistic regression and neural network models, the neural network model had an overall efficiency rating that was higher than logistic regression and support vector machine.

The traditional class imbalance approach is used to solve the problem. A number of advanced methods are developed to achieve increased precision in standard classification approaches to solve the class-imbalanced problem. The cost-sensitive method of learning generally includes a cost matrix for all error or case categories. It aims mainly to facilitate learning from imbalanced data sets. In the process of sampling the minority class, this mechanism has an equivalent effect. It can end by overfitting training with specific rules or regulations. ANN is a characterized statistical learning algorithm, based on the Neural Network. The neural network is a neuronal network to detect cases when activated. Approximation and feature assessment can be performed by looking for network sizes that take inputs[1]. The interconnected neurons of ANNs may also, besides pattern recognition, be defined as input/output measurement in line with the machinery learnings (pattern recognition). K-Nearest neighbour (KNN) is an instance-based classification system to support kernel learning in anomaly repositories. The kernel design classification function KNN optimizes the anomaly pattern removal function [2].

Before classification classes are applied, the dataset which is totally skewed towards the majority class is required in the pre-process. In some cases, along with the cases of anomaly, minority classes can be detected. The minority class anomaly is much higher than the majority group. In the case of majority classes, most classification models therefore provide a high level of precision, whereas the minority type is less accurate [3].

The extraction of hidden information from heterogeneous database is one of the major issues in real-time applications. In data mining research, the data mining algorithm for both high-dimensionality and cardinal data was considered one of the challenges. The large number of data generated with different heterogeneous features are used in the conventional models for pattern discovery [4]. Based on the test with new techniques such as Gaussian, a mixture of Gaussian, naive and support vector data description by biomedical data sets, an approach is designed for One-Class Classification (OCC) [5]. There was no need for continuous data on the capacity of the process through data sets. When there is a benchmark data set for a one-class classification, the estimate includes multiple classes for all data sets. Each class is taken as a target class and newly classified units are taken from the units in other classes. They conceived problems of class imbalance in medical data sets. For the prediction model, a balanced data set is essential [6]. Class labels fail to balance medical data sets. The classification methods, while the data set is imbalanced, run on minority class examples. It is also used to increase the accuracy of all classes without a relative distribution [7]. The methods of over-sampling and under-sampling are used. To identify right hyperplane, they explained the linear combinational model. It identifies a suitable hyperplane for the classification of the target classes. It is necessary to select the hyperplane in such a way that it effectively classifies the target class. Hyperplane selection differs from one scenario to another. If three hyperplanes exist, the target



classes are separated accurately by all three. The concept of distance margin is used in such a scenario [8]. It allows for easier identification of a suitable hyper plane [9]. Similarly, to get better results, the margin distance measures are adjusted according to scenarios. Some of the SVM classification techniques that operate from margin distance measures are the maximum margin classifier and the soft margin classifier. Through which it conducts an effective process of classification.

2. Related works

Generic models are referred to as models that are developed by bureaus or institutions based on industry data applicable to all participants or subscribers. Custom models are models built for a particular group of borrowers, who have specific requirements or need a more in-depth subscription compared to other borrowers". It is possible to classify FICO (Credit scores like Fair Isaac Corporation), Vantage and CIBIL etc as generic models. Most of the infamous models specific to certain institutions or groups of customers, on the other hand, are classified as custom models. Although conceptually customized scores are better off in a customer's assessment, certain significant issues such as feasibility, development and implementation can hinder the implementation of customized models. The credit scoring story begins around the beginning of the 19th century, despite the history of credit dating back to 2000BC, and the real universal growth in credit scoring research activities was not until the end of the 20th century [10]. The literature available therefore is limited. Although credit scoring has gained importance in recent years, it is not reflected in the amount of research, particularly in developing countries. The following section deals with the review of available research related to the classification of credit scoring models, the comparison of credit scoring tools and the credit scoring assessment criteria. Credit scoring models were categorized as three types in one classification, namely

eISSN1303-5150

heuristic models, statistical models and causal models. Heuristic models or expert models are based on the experience of experts, who are subjective in their understanding and decisions. For evaluations, these models do not undertake complex mathematical or statistical techniques. There is a more comprehensive and organised analysis of credit card fraud detection techniques that was done by [11] that also provides an overview of detection techniques in telecommunications fraud intrusion and money laundering, as well as in medical and academic fraud. According to a common view of credit card fraud detection, the goal is to accurately classify transactions as legitimate or fraudulent. In this field, there hasn't been a lot of study done yet. Despite the fact that fraud detection has been around for a long time. Because banks are unwilling to hide their sensitive customer transaction data for reasons of privacy and lack of access to real-world data, researchers are unable to conduct experiments. In addition, researchers are unable to learn about actual fields because the field names have been changed over time. Simple rule matching techniques are the most commonly used in automated fraud detection in the financial sector. This includes the detection of credit card fraud. Experts in the field of fraud detection usually develop the rules and code the detection system using some form of descriptive language. When a transaction or set of transactions passes through the system and the preconditions of the rule are met, the detection rules "fires." This type of expert system falls under the category of misuse detection. For the detection of credit card fraud, the expert system approach is popular because it uses the rules defined for the standard events that indicate a credit card fraud (e.g. if the transaction is done above the threshold value declared in the system, or if the purchase of any item or service is done simultaneously in two different places which are very far away). Although expert systems are good at detecting well-known types of fraud, they aren't very adaptable when it comes to



handling new types of fraud that deviate from the standard pattern. Academics have studied the differences between standard expert systems; for example, some have been applied to the detection of fraud in the insurance industry. Combining methods improves the effectiveness of fraud detection models and may also provide better coverage. They combined the three different methods of fraud detection into a single system and recommend that rules be mined from a labelled dataset and changed over time to address new patterns of mobile phone cloning fraud as they emerge. Fuzzy logic, a soft computing concept, can be used in expert systems, which can be seen as the evolution of standard expert systems. There are sets of fuzzy rules that can allow a given set of parameters to have either full or partial membership in preconditions of a fuzzy rule. An attractive and useful feature of fuzzy models is their ability to be both grammatically correct and mathematically sound at the same time. An electrochemical process is commonly used to communicate between neurons. Neurons communicate with each other via a network of dendrites (input connections) and axons (output connections), which are linked by synapses (output connection). Using a customer's past spending data, the system builds a neural network that can be used to spot unusual patterns in their current spending. Large datasets can be handled by this system, and the parameters of an analysis can be easily changed through a graphical user interface. Customers' purchasing patterns are represented by three transaction features: the time since the last purchase in the same category, the transaction amount, and the category of the purchase. The system was tested by preparing synthetically generated data. As a result of this design flaw, each customer will require their own custom neural network. As a result, the network as a whole becomes extremely large, necessitating higher maintenance budgets. Fisher's discriminant analysis was used by them to distinguish between legitimate and fraudulent

eISSN1303-5150

operations. In addition, a neural network-based fraud detection system known as Minerva was created. Credit card fraud detection relies heavily on locating itself deep within the network's transaction servers. Because the system relies solely on previous transactions, it doesn't necessitate a large database and can classify a transaction in as little as 60 milliseconds. Because of the difficulty in obtaining useful datasets for training and determining a meaningful set of detection variables, this system has a number of drawbacks. By combining a rule-based classification approach with a neural network algorithm, they identified fraud cases. It was the neural network that verified the transaction classification after the rule-based classifier had checked for fraud. This technique increases the likelihood of correctly identifying fraud, thereby reducing the number of false alarms while simultaneously boosting confidence in the system.. Neuronal networks for fraud detection can be trained either supervised or unsupervised. This last step does not necessitate any prior labelling of classes. "Self Organizing Maps" (SOM) [12] is the most well-known unsupervised segmentation model approach. On the basis of customer behaviour, self-organizing map neural networks are used to detect credit card fraud. Many studies employ this clustering technique. Unsupervised methods are known as "anomaly detection" in the context of intrusion detection, whereas supervisory methods are commonly called "misuse detection" in the context of fraud detection. Fraud detection is concerned with classifying transactions into valid and fraudulent categories. Commercial intrusion detection systems are available, but most of these are based on the misuse-detection approach with static detection rules (i.e. rule-based expert systems or very simple pattern matching techniques), which means they use very simple techniques. Using multiple local fraud classifiers with different learning algorithms (ID3, CART, RIPPER, and BAYES), started a research project to combine these into meta-classifiers. Since the base classifiers are run

www.neuroquantology.com

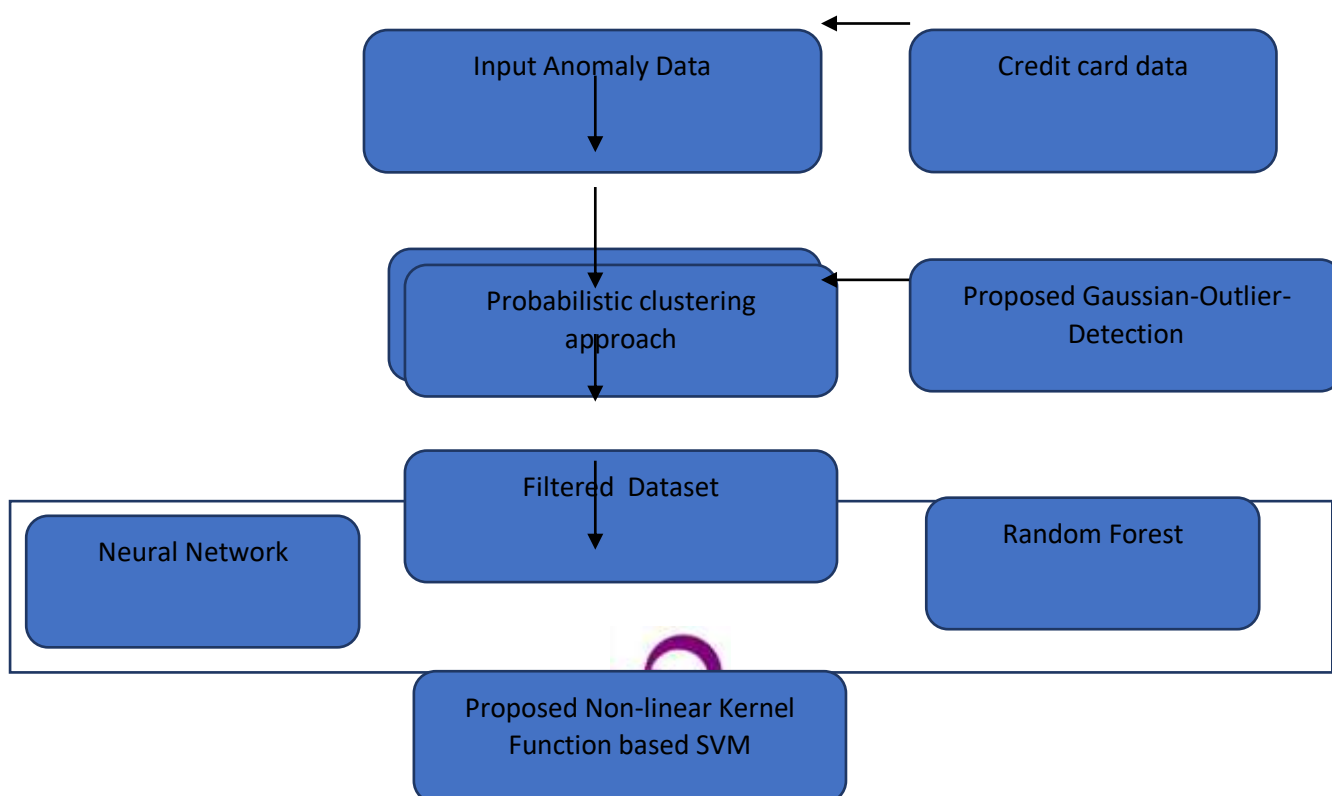


locally and fed into the meta-classifier for final classification, organisations don't have to share their private data. Support Vector Machine (SVM) is used for collaborative work. [13] Proposed an unsupervised credit card detection technique that uses unusual spending patterns and transaction frequency to identify suspicious activity. Peer group analysis and break point analysis are two unsupervised outlier detection techniques that Bolton and Hand proposed, respectively. According to the paper, PGA techniques can be used to detect anomalies in local data, and the BPA technique can be used to define fraudulent behaviour by equating transactions at the start and end of a given time window. [14] Proposed a fraud density map (FDM) technique to improve a neural network's learning efficiency. The fraud density map attempts to address the issue of the incompatible distributions of fraudulent and genuine transactions among the real data and the training data because there is a lack of knowledge of fraudulent transactions in training data sets. In addition, data mining is used to enhance their work. The fraud pattern mining algorithm and association rules, which provide information on the features present in fraudulent transactions, have been developed to mine fraud. In order to thwart bank heists, banks are implementing new

fraud patterns into their fraud detection systems.

3. Contextual Gaussian Cluster based Classification model for Anomaly Detection

In the proposed framework, a hybrid Gaussian filter based clustering and classification model is proposed on the credit card and KDD databases. In this model, novel pre-outlier instances are detected and filtered using the Gaussian elimination process. These filtered data is clustered to find the anomaly and non-anomaly classes for the anomaly classification process as shown in the figure 1. In this framework, different types of anomaly datasets such as credit card and KDD are taken to find the outliers as pre-outlier analysis. Here, a hybrid probabilistic clustering measure is used to detect the anomaly class and non-anomaly class for the classification problem. In the proposed framework, a hybrid Gaussian outlier model is proposed in order to improve the overall classification rate and error rate. In this Gaussian approach, an advanced Gaussian probability measure is used to find the anomaly object in the dataset. These Gaussian values are used to find the outlier instances and non-outlier instances in the given database as pre-outlier analysis.



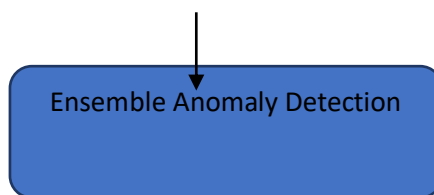


Figure 1: Proposed Model

Pre-outlier detection using Gaussian Estimation measure

Input: Training Samples T, Features-set FS.

Output: Anomaly samples O, Non-anomaly samples N.

1. Read the training data D.
2. To each attribute A in the feature set FS.
3. do
4. Perform the proposed Gaussian probability measure to find the anomaly object in the given training data.

$$\text{Gaussian_outlier_measure} := \text{GOM}(T) = \max \left\{ \frac{1}{\sqrt{2\pi}} \left\{ \frac{(A_i - \text{mean } i) \cdot \exp(A - \text{mean}(i) / \text{sd}(i))}{|N|} \right\} \right\}$$

Where GOM(T) defines the gaussian probability measure of each sample I.

A_i defines the sample of its attribute.

Done

5. Sort and find the instances with highest gaussian probability values.
6. For $i=1$ to GOM(T)
7. Do
8. If $\{\text{GOM}(T) < \lambda\}$
9. Assign sample I as Outlier.
10. $O(i) = \text{GOM}(T(i));$
11. Else
12. Assign sample I as Not Outlier.
13. $N(i) = \text{GOM}(T(i));$
14. Construct non-outlier training samples using N.
15. Done
16. Perform hybrid clustering approach for post outlier analysis.

The nonlinear transformation function is used to implement a hybrid Gaussian outlier detection algorithm on the non-outlier objects. In order to find the training data outliers, a non-linear transformation function is applied to each feature.

Hybrid Post outlier detection using the Clustering Measure:

Step 1: Load the pre-filtered non-anomaly data.

Step 2: Select k random objects as initial cluster centres for outlier detection.

Step 3: Compute non-linear central distance as



$$dc = \rho_i = \sum_{j \neq i} F \left(\sum_{t=1}^m (x_i^t - x_j^t)^2 - T \right)$$

where T represents the distance threshold, $F(x) = 1$ if $x < 0$ and otherwise $F(x) = 0$.

$$\eta_i = \sum_{j \in i, j \neq i} e^{-\left(\frac{d_{i,j}}{dc}\right)^2}$$

$$d_{ij} = \sqrt{\sum_{t=1}^m (x_i^t - x_j^t)^2}$$

Step 4: Update initial clusters using the proposed cluster probability measure as

$$CU = \frac{1}{n} \sum_k P(C_k) \sum_{ij} \left[P(A_i | C_k)^2 - P(A_i)^2 \right] * \sqrt{\eta}$$

Where, $A_i = V_{ij}$ represents Attribute value and C_k represents Classes.

Step 5: Repeat till k clusters and predefined number of iterations.

Hybrid Ensemble classification model



BOOSTINGALG

```

Let the input dataset is represented as ID = {In1, In2, In3, ... Inn} be the given dataset
Set of proposed classifiers and base classifiers are represented as ensemble classifiers as
EC = {Neuralnetwork, Randomforest, ProposedSVM},
Tr = Training data taken from ID, Tr ∈ ID
Te = Test data taken from the ID, Te ∈ ID
l = length(ID)
for i=1 : R(#number of partitions)
do
Let h = i
S(h) ⊂ Tr
MC = {}; // Model classifier
CO={}; Classifier output
for j = I to l
do
if (j > 1)
s(j) = Set of wrongly classified records of jth model MC(i) on S(h)
CO = CO ∪ CC(j); // CC(j): Correctly classified instances
end if
done
for(k = 1 : |EC|
R(k) = Te classified by CO(k)
Output = max (R(k) : k = 1, 2, ..., | EC |)
done
    
```

1994

Non-linear kernel SVM is used to predict anomalies from filtered data in this algorithm. The high-dimensional anomaly class prediction is computed using a hybrid kernel function on the filtered anomaly data. Classifiers C1 and C2 have been proposed for use in the boosting approach to improve classification. In order to test the majority voting of each sample, a boosting classifier is applied to the hybrid base classifiers c1 and c2.

Hybrid Multi-class SVM for outlier detection

The objective function of the proposed multi-class SVM for outlier classification process is given as

$$\min_{w, b, \chi, \rho} \left\{ \frac{1}{2} w^T w - v \cdot \eta + \chi \sum_{i=1}^l \lambda_i \right\}$$

s.t.

$$y_i (w^T \cdot \phi(x_i, y_i) + b) \geq \eta - \lambda_i$$

$$\lambda_i \geq 0, i = 1 \dots l, \eta \geq 0$$

$$\phi(x_i, y_i) = e^{\chi \cdot \log(\sum |y_i|^2)}; \text{if } (x_i > y_i)$$

$$= e^{\chi \cdot \log(\sum |x_i|^2)}; \text{if } (x_i < y_i)$$

$$= e^{\chi \cdot \log(\sum |x_i - y_i|^2)}; \text{if } (x_i = y_i)$$



The decision boundary is given by

$$\text{sgn}\left(\sum_{i=1}^1 y_i \cdot \phi(x_i, y_i) + b\right)$$

4. Experimental Results

Third-party libraries are used to simulate the experimental results in the Java and NetBeans environments. The proposed study uses the kdd cancer and credit card datasets, both of which include a significant number of features. The proposed model is tested on cloud-based training anomaly datasets.

Outlier detection and data transformation algorithms are used to filter these datasets at first. Following that, the filtered data is fed into a proposed classification algorithm for anomaly detection and decision-making. The credit card dataset for the data classification task is shown in Table 1.

Table 1: Sample Credit card Data

checking_status	credit_limit	purpose	credit_age	savings_account	employment_status	installment_plan	personal_status	other_parts	residence	property_status	age	other_pay	housing	existing_credit	job	num_dep	own	tele	foreign	so	class
<=0	5	'critical/ot'	radio/ty	1189	'no known'	>=7	4	'male singl'	none	4	'real estate'	67	none	own	2	'skilled'	1	yes	yes	good	
0<=X<200	48	'existing pt'	radio/ty	5951	<500	1<=0<=4	2	'female di'	none	2	'real estate'	22	none	own	1	'skilled'	1	none	yes	bad	
'no checki	32	'critical/ot'	education	2098	<100	4<=0<=7	2	'male singl'	none	3	'real estate'	40	none	own	1	'unskilled r'	2	none	yes	good	
<=0	41	'existing pt'	furniture/e	7882	<100	4<=0<=7	2	'male singl'	guarantor	4	'life insura	45	none	'for free'	1	'skilled'	2	none	yes	good	
<=0	24	'delayed p'	'new car'	4875	<100	1<=0<=4	3	'male singl'	none	4	'no known'	53	none	'for free'	2	'skilled'	2	none	yes	bad	
'no checki	36	'existing pt'	education	9055	'no known'	1<=0<=4	2	'male singl'	none	4	'no known'	35	none	'for free'	1	'unskilled r'	2	yes	yes	good	
'no checki	24	'existing pt'	furniture/v	2835	500<=0<=11<=7		3	'male singl'	none	4	'life insura	53	none	own	1	'skilled'	1	none	yes	good	
0<=X<200	36	'existing pt'	'used car'	6948	<100	1<=0<=4	2	'male singl'	none	2	car	35	none	rent	1	'high quali'	1	yes	yes	good	
'no checki	12	'existing pt'	radio/ty	3059	>=1000	6<=0<=7	2	'male di/v'	none	4	'real estate'	61	none	own	1	'unskilled r'	1	none	yes	good	
0<=X<200	30	'critical/ot'	'new car'	5234	<100	unemploye	4	'male mar,	none	2	car	28	none	own	2	'high quali'	1	none	yes	bad	
0<=X<200	12	'existing pt'	'new car'	1295	<100	<1	3	'female di'	none	1	car	25	none	rent	1	'skilled'	1	none	yes	bad	
<=0	48	'existing pt'	business	4308	<100	<1	3	'female di'	none	4	'life insura	24	none	rent	1	'skilled'	1	none	yes	bad	
0<=X<200	12	'existing pt'	radio/ty	1567	<100	1<=0<=4	1	'female di'	none	1	car	22	none	own	1	'skilled'	1	yes	yes	good	
<=0	24	'critical/ot'	radio/ty	1195	<100	>=7	4	'male singl'	none	4	car	60	none	own	2	'unskilled r'	1	none	yes	bad	
<=0	35	'existing pt'	'new car'	1403	<100	1<=0<=4	2	'female di'	none	4	car	28	none	rent	1	'skilled'	1	none	yes	good	
<=0	24	'existing pt'	radio/ty	1282	100<=0<=51<=0<=4		4	'male singl'	none	2	car	32	none	own	1	'unskilled r'	1	none	yes	bad	
'no checki	24	'critical/ot'	radio/ty	2424	'no known'	>=7	4	'male singl'	none	4	'life insura	53	none	own	2	'skilled'	1	none	yes	good	
<=0	30	'no credit'	business	8072	'no known'	<1	2	'male singl'	none	3	car	25	bank	own	3	'skilled'	1	none	yes	good	
0<=X<200	24	'existing pt'	'used car'	12579	<100	>=7	4	'female di'	none	2	'no known'	44	none	'for free'	1	'high quali'	1	yes	yes	bad	

Table 2: Comparative analysis of number of outliers detected in the credit card dataset

Test Data	ITree	Gaussian Outlier	Pre outlier
Test#1	96	87	61
Test#2	83	85	63
Test#3	94	83	63
Test#4	82	82	58
Test#5	81	97	65
Test#6	93	90	66
Test#7	89	92	63
Test#8	81	78	68
Test#9	86	80	57
Test#10	85	98	69
Test#11	82	91	65
Test#12	93	77	60
Test#13	80	83	63
Test#14	99	96	58
Test#15	85	96	59
Test#16	95	77	67
Test#17	85	81	61
Test#18	82	96	68
Test#19	79	99	62
Test#20	83	79	66



Table 2, describes the performance results of proposed outlier detection model to the conventional models on credit-card dataset. As shown in table2, traditional models have large number of candidate sets than the proposed model on credit-card dataset.

Table 3: Comparative analysis of runtime (ms) detected in the credit card dataset

Test Data	ITree	Gaussian Outlier	Pre outlier
Test#1	6230	7223	3045
Test#2	6066	7622	2796
Test#3	6073	6416	3157
Test#4	7637	7641	2814
Test#5	7119	6935	2798
Test#6	7899	7169	3143
Test#7	7859	7153	3102
Test#8	6253	6127	2831
Test#9	6874	6270	2717
Test#10	6164	6363	2724
Test#11	7611	7460	3232
Test#12	6188	7208	2810
Test#13	6844	7924	3228
Test#14	6780	7190	2942
Test#15	7131	6902	2719
Test#16	7801	7202	3033
Test#17	6312	7922	3272
Test#18	6146	7698	2875
Test#19	7490	6444	2911
Test#20	7708	6187	2694

1996

On the credit card dataset, Table 3 shows a comparison of proposed outlier identification approach to the proposed outlier detection approach. On the credit card dataset, the conventional procedures had a shorter runtime than the proposed solution, as seen in the table.

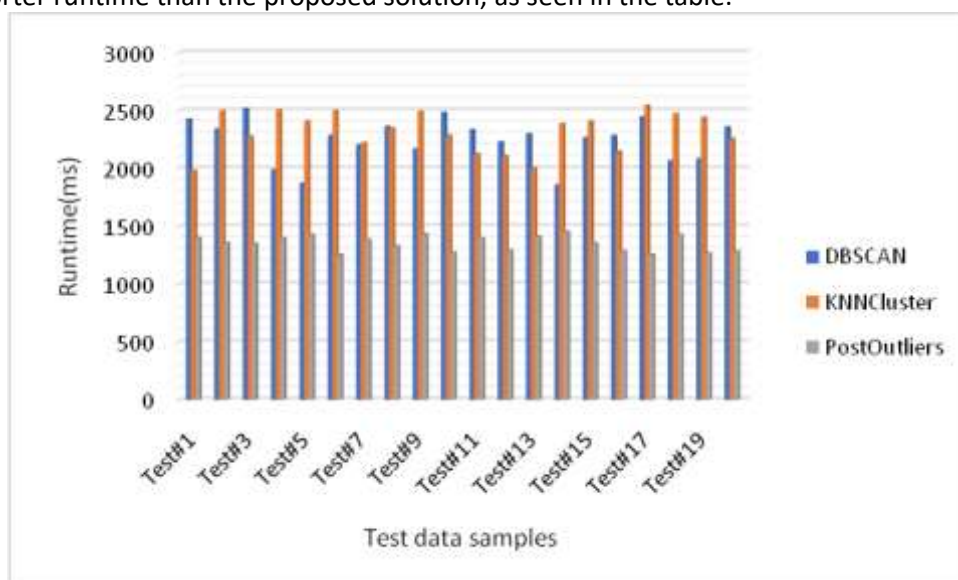


Figure 2: Comparative analysis of proposed post outlier runtime to the conventional models



Table 4: Comparative analysis of anomaly detection algorithm in terms of runtime (ms) on test credit card data.

Test Data	SVM	RF	Proposed Ensemble
Test#1	1304	1799	900
Test#2	1522	1529	926
Test#3	1891	1424	839
Test#4	1300	1611	902
Test#5	1363	1407	752
Test#6	1458	1265	853
Test#7	1554	1387	856
Test#8	1420	1714	739
Test#9	1557	1651	897
Test#10	1431	1294	952
Test#11	1500	1721	705
Test#12	1398	1773	950
Test#13	1751	1249	748
Test#14	1583	1836	698
Test#15	1525	1527	845
Test#16	1333	1511	970
Test#17	1400	1779	880
Test#18	1272	1610	836
Test#19	1590	1801	972
Test#20	1861	1724	728

1997

On the credit card dataset, Table 3 shows a comparison of proposed classification approach to the proposed classification approaches. On the credit card dataset, the proposed model has better runtime (ms) than the traditional approaches on credit-card fraud detection process.

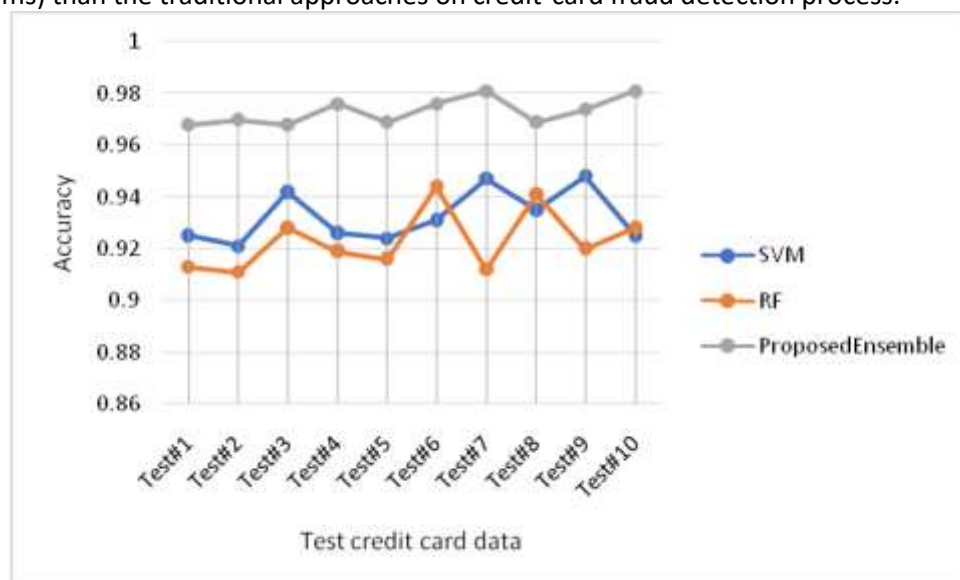


Figure 3: Performance results of advanced boosting classifier accuracy to the conventional approaches on the credit card dataset.



On the credit card dataset, Figure 3 shows the results of advanced boosting classifier accuracy versus conventional approaches. On the credit card dataset, the current model has better accuracy than the traditional approaches, as shown in the figure.

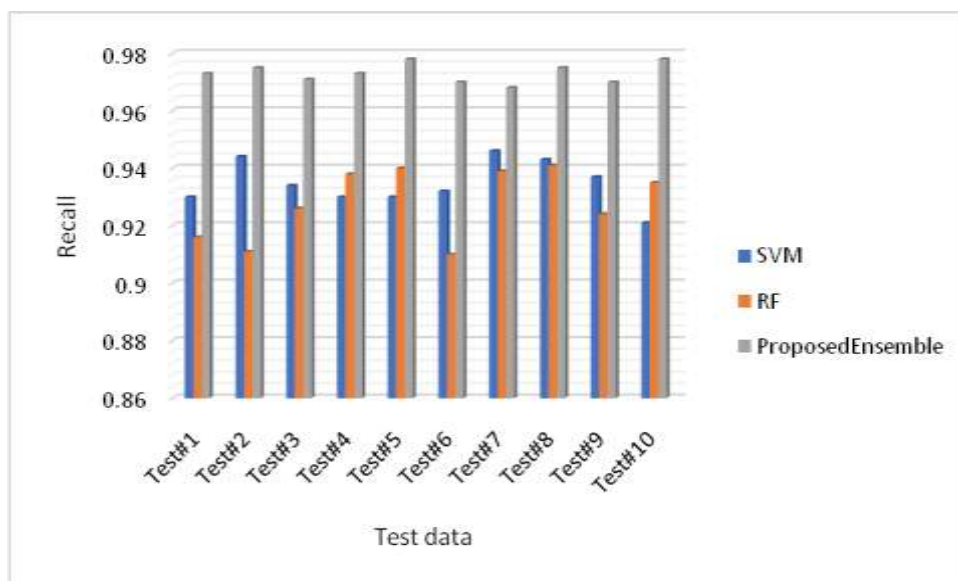


Figure 4: Performance results of advanced boosting classifier recall to the conventional approaches on the credit card dataset

On the kdd dataset, Figure 4 shows the results of advanced boosting classifier recall versus conventional approaches. On the credit card dataset, the current model has a higher recall than traditional approaches, as shown in the figure.

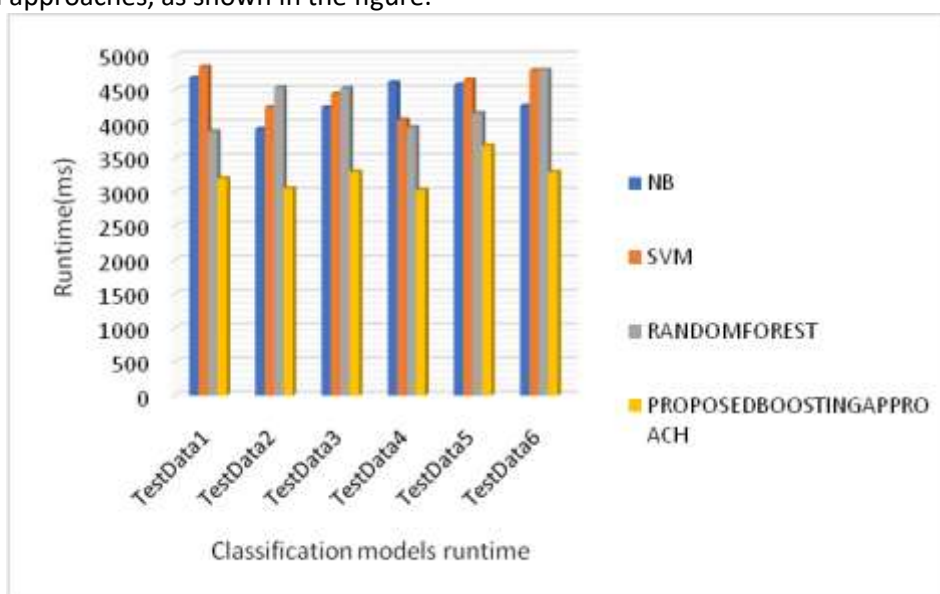


Figure 5: Performance results of advanced boosting classifier runtime (ms) to the conventional approaches on the credit card dataset.

On the credit card dataset, Figure 5 shows the results of advanced boosting classifier runtime versus conventional approaches. On the credit card dataset, the current model has a faster runtime than the traditional approaches, as shown in the figure.



Table 5: Performance results of advanced boosting classifier AUC to the conventional approaches on the credit card dataset.

Test Data	SVM	RF	Proposed Ensemble
Test#1	0.933	0.939	0.967
Test#2	0.942	0.912	0.968
Test#3	0.947	0.919	0.978
Test#4	0.946	0.943	0.971
Test#5	0.947	0.927	0.97
Test#6	0.922	0.917	0.968
Test#7	0.949	0.915	0.97
Test#8	0.933	0.942	0.977
Test#9	0.929	0.936	0.977
Test#10	0.926	0.917	0.977

On the credit card dataset, Table 5 shows the results of advanced boosting classifier AUC versus conventional approaches. On the credit card dataset, the current model has a better AUC than the conventional approaches, as shown in the table.

1999

Test Data	SVM	RF	Proposed Ensemble
Test#1	0.938	0.94	0.974
Test#2	0.931	0.923	0.971
Test#3	0.933	0.935	0.982
Test#4	0.933	0.916	0.98
Test#5	0.948	0.939	0.981
Test#6	0.937	0.929	0.97
Test#7	0.941	0.916	0.978
Test#8	0.95	0.93	0.969
Test#9	0.936	0.928	0.979
Test#10	0.937	0.922	0.979

On the credit card dataset, Table 6 shows the results of advanced boosting classifier precision versus conventional approaches. On the credit card dataset, the current model has better precision than the conventional approaches, as shown in the table.

hybrid boosting classifier is applied to the filtered anomaly datasets (ms). On anomaly databases, experimental results show that the booting classifier is nearly 5% more accurate than traditional classification models.

Venkata Ratnam Ganji

Conclusion

For better decision making, advanced machine learning approaches are applied to anomaly databases in this paper. Because most conventional approaches are unaffected by outliers or data size, the proposed model performs better in terms of outliers, filtering, and data classification. On the anomaly databases, a hybrid outlier detection and data transformation approach is used in this study. Finally, to improve accuracy and runtime, a





Venkata Ratnam Ganji Obtained M.Tech(CSE) from University College of Engineering, JNTUK, Kakinada in 2011. He is currently pursuing Ph.D degree with the Department of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India. He is doing research on credit card fraud detection. He has published several research articles in various international and national journals. His research interests are Data Mining, Big data and Machine Learning.

Aparna Chaparala



APARNA CHAPARALA received her Ph.D. degree in Computer Science & Engineering from JNTUH, Hyderabad in 2014. She is currently working as Professor in CSE department of RVR & JC College of Engineering, Guntur. She has 18 years of experience in teaching engineering students.

She has published several studies and scientific articles in well-known international journals and conferences. She is a professional member of ACM India. She is currently serving as Supervisor for many Postgraduate and Ph.D students. Her research interests include Machine Learning, Cyber Security, Digital Forensics, and Optimization Techniques.

References

- [1]M. Bahrami, M. Pourahmadi, A. Vafaei, and M. R. Shayesteh, "A comparative study between single and multi-frame anomaly detection and localization in recorded video streams," *Journal of Visual Communication and Image Representation*, vol. 79, p. 103232, Aug. 2021, doi: 10.1016/j.jvcir.2021.103232.
- [2]A. G. C. de Sá, A. C. M. Pereira, and G. L. Pappa, "A customized classification algorithm for credit card fraud detection," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 21–29, Jun. 2018, doi: 10.1016/j.engappai.2018.03.011.
- [3]Z. Li, M. Huang, G. Liu, and C. Jiang, "A hybrid method with dynamic weighted entropy for handling the problem of class imbalance with overlap in credit card fraud detection," *Expert Systems with Applications*, vol. 175, p. 114750, Aug. 2021, doi: 10.1016/j.eswa.2021.114750.
- [4]S. Misra, S. Thakur, M. Ghosh, and S. K. Saha, "An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction," *Procedia Computer Science*, vol. 167, pp. 254–262, Jan. 2020, doi: 10.1016/j.procs.2020.03.219.
- [5]Y. Zhou, H. Ren, Z. Li, and W. Pedrycz, "Anomaly detection based on a granular Markov model," *Expert Systems with Applications*, vol. 187, p. 115744, Jan. 2022, doi: 10.1016/j.eswa.2021.115744.
- [6]D. Lakhmiri, R. Alimo, and S. Le Digabel, "Anomaly detection for data accountability of Mars telemetry data," *Expert Systems with Applications*, vol. 189, p. 116060, Mar. 2022, doi: 10.1016/j.eswa.2021.116060.
- [7]V. Van Vlasselaer et al., "APATE: A novel approach for automated credit card



transaction fraud detection using network-based extensions,” *Decision Support Systems*, vol. 75, pp. 38–48, Jul. 2015, doi: 10.1016/j.dss.2015.04.013.

[8]Q. Yu, M. Kavitha, and T. Kurita, “Autoencoder framework based on orthogonal projection constraints improves anomalies detection,” *Neurocomputing*, vol. 450, pp. 372–388, Aug. 2021, doi: 10.1016/j.neucom.2021.04.033.

[9]F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, “Combining unsupervised and supervised learning in credit card fraud detection,” *Information Sciences*, vol. 557, pp. 317–331, May 2021, doi: 10.1016/j.ins.2019.05.042.

[10]E. N. Osegi and E. F. Jumbo, “Comparative analysis of credit card fraud detection in Simulated Annealing trained Artificial Neural Network and Hierarchical Temporal Memory,” *Machine Learning with Applications*, vol. 6, p. 100080, Dec. 2021, doi: 10.1016/j.mlwa.2021.100080.

[11]M Dileep Kumar and KV Ramana, “Cardiovascular disease prognosis and severity analysis using hybrid heuristic,” *Multimedia Tools and Applications*, ISSN 1380-7501, DOI 10.1007/s11042-020-10000-w

[12]A. Rb and S. K. Kr, “Credit card fraud detection using artificial neural network,” *Global Transitions Proceedings*, vol. 2, no. 1, pp. 35–41, Jun. 2021, doi: 10.1016/j.gltp.2021.01.006.

[13]M Dileep Kumar and KV Ramana, “Left Ventricle Of Cardiovascular Image Segmentation Using T-Segnet Hybrid And Extended Buffalo Optimization,” *European Journal of Molecular & Clinical Medicine*, ISSN 2515-8260 Volume 07, Issue 08, 2020

[14]V. N. Dornadula and S. Geetha, “Credit Card Fraud Detection using Machine Learning Algorithms,” *Procedia Computer Science*, vol. 165, pp. 631–641, Jan. 2019, doi: 10.1016/j.procs.2020.01.057.

[15] M Dileep Kumar and KV Ramana, “Cardiac Segmentation from MRI images using Recurrent & Residual Convolutional Neural Network based on SegNet and Level Set

methods”, *Annals of R.S.C.B.*, ISSN:1583-6258, Vol. 25, Issue 3, 2021, Pages. 1536 - 1545

[16]S. Panigrahi, A. Kundu, S. Sural, and A. K. Majumdar, “Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning,” *Information Fusion*, vol. 10, no. 4, pp. 354–363, Oct. 2009, doi: 10.1016/j.inffus.2008.04.001.

[17]N. Rtayli and N. Enneya, “Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization,” *Journal of Information Security and Applications*, vol. 55, p. 102596, Dec. 2020, doi: 10.1016/j.jisa.2020.102596.

