International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-6, April 2020

Credit Card Fraud Detection Performance Improvement using Advanced Super Gradient Boosting Algorithm

V. Sudheer Goud, P. Premchand

Abstract: ***Credit card fraud introduces to the physical loss of a credit card or the destruction of sensitive credit card data. Several text mining procedures can be used for disclosure. This investigation reveals several algorithms that can be used to analyze transactions as a fraud or as a real background. This paper represents the possibility of fraudulent transactions in the prevalence and meaning of credit card usage also, Credit card fraud data collection was used in the investigation. Since the dataset was largely unbalanced, SMOTE (Synthetic Minority oversampling Technique) is applying for an overdose. In addition, jobs selected, and the data set divided into two parts, training data and test data. In this paper, The Advanced Super Gradient Boostingbased Text mining Algorithm (ASGB) suggested to detect the fraud transaction in Credit card transactions. ASGB is a Decision-Tree-Based Ensemble Text mining algorithm that utilizes a gradient boosting framework. In forecast difficulties, including unstructured data (Images, Text, etc.), artificial neural networks tend to exceed all other algorithms or structures. The proposed algorithms used in the experiment were the Hidden Markov Model, Random Forest, Gradient Boosting, and Enhanced Hidden Markov Model. The Experimental Results show that proposed algorithms, a well-tuned ASGB classifier outperforms all of them. And it presents better Precision is 99.1%, and Recall is 99.8%, F-measure is 99.5%.***

Keywords: Credit card fraud detection,Text mining, SMOTE, HMM, GB, Random Forest, and ASGB.

I. INTRODUCTION

Text mining techniques it has produced the most significant impact on fraud detection. It is reasonable since there are large amounts of digital data or can easily convert into statistics in the form of censuses and dimensions. It must also remember that the processing speed is of nature.

In particular, the problem lies in processing transactions, especially in communication and intervention records, where large amounts of reports prepared every day. However, they also practiced in credit cards, banking, and retail areas. The serious problem with standby paints is how useful devices are in detecting fraud, and the fraudulent problem is that it is generally not recognized how well-known counterfeit instances escape online.

Revised Manuscript Received on April 30, 2020.
***** Correspondence Author
V. Sudheer Goud*, Professor, Department of Computer Science in Holy Mary Institute of Technology and Science (HITS), (V) Bogaram, (M) Keesara, Medchal .Dist, Telangana, India.
P. Premchand, Professor, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, Telangana State, India

In the data, along with the time available for disclosure after the origins of fraud (in minutes, the number of transactions, etc.) should.

also be said. Examples of these functions are related to carrying the final discovery: in many cases, an account, a smartphone, etc. It must be used in many fraudulent transactions before it can be discovered as fraudulent so that many wrong and terrible classifications always madeDue to the rapid development of digital commerce, the use of credit cards for purchases has dramatically accelerated. Unfortunately, fraudulent or illegal use of credit cards has also provided attractive profits to scammers. Credit card frauds increase significantly due to security vulnerabilities that have discovered in traditional credit card processing systems due to a decrease in billions of bills every 12 months. The fraudsters of this time are very colorful and use the latest technology in credit card fraud. Fraudulent sports gifts present specific challenges to different banks and monetary institutions around the world that have difficulty playing credit cards.

There is a growing group of new international companies [1]. All these organizations strive to provide the best services for their clients. To achieve this, organizations process massive amounts of information daily. These records come from a wide range of assets and come in various forms. Also, this distinction contains some essential parts of the company's future company. This is why agencies need to keep them safe, store this information, technology, and what's vital. Without protecting information, many companies can use it to the fullest, or worse, they can steal.

In most cases, financial records are stolen, which could harm the organization or the whole person. There are many types of fraud [2]. Fraud verification takes place while a character fakes the test or can pay some money at a glance, understanding that there aren't always enough coins. Internet revenue is a scam when a fraudster sells counterfeit or counterfeit products or pays without delivering the products. There are other things like charitable fraud, identity theft, credit card fraud, debt consolidation, insurance fraud, and others. Due to the increasing popularity of non-cash transactions, the extraordinary maximum fraud is credit card fraud. Credit card fraud refers to a scenario in which the prankster uses a credit score card for their desires, even when the owner of the credit score card is unaware. In 2016, the number of fraudulent transactions with credit cards worldwide was 1.8 billion euros [3]. Although the volume of credit card transactions is increasing drastically, this is because of the latest fraud detection systems.

However, scammers are always finding new ways to steal information [4].

II. RELATED WORK

Credit card fraud detection has generated significant interest in investigations, and several techniques have been suggested, with particular emphasis on Text mining. Gosh and Reilly [5] It developed a fraud detection system with a neural network. Your system trained in a large sample of transactions for premium credit card accounts. These transactions include examples of lost cards, stolen cards, application fraud, fraudulent fraud, mail order fraud, and non-receiving issue status (NRI).

E. Aleskerov et al. [6] Present Carded, a database mining system used to detect credit card fraud. This system is based on the neuronal learning module and provides the interface for multiple trading databases.

Dorronsoro et al. [7] He suggested two unique features about fraud detection: the minimum period and a large number of credit card operations to be decided. They routinely removed false verbs using Fisher's distinction.

Syeda et al. [8] He used Parallel Neural Networks to improve the speed of text mining and knowledge discovery in detecting credit card fraud. A complete system was implemented for this.

Chen et al. [9] they subdivided a broad set of transactions into smaller subsections and then applied distributed Text mining to create user behavior models. Improves the detection accuracy of the resulting base models to create a meta-classifier.

A. Srivastava et al. [10] Model a series of credit card transactions using the Markov Hidden Model (HMM) and show how they can be used to detect fraud. HMM is initially implemented using the card holder routine. If a trained HMM does not accept a high-probability incoming credit card transaction, this is considered fraud. At the same time, they are also trying to ensure real transactions are not excluded.

The fraud detector was recently developed by Suvasini Panigrahi et al. [11], which include four extensions, specifically a lead candidate, a Dempster Schaefer, a transaction history database, and a Bayesian student. In a part that mainly depends on the teachings, they determine the level of suspicion of every transaction that comes from a very good employer-based mostly on the edge of deviation: the Dempster-Schaeffer theory used in combining many tests and calculating the initial belief.

In paper [12], neural network test in the European dataset is experiment included the neural network for more dispersive work with a wheel algorithm. The neural network has two input layers, 20 invisible layers, and two output layers. Due to the optimization algorithm, they achieved remarkable results in 500 test samples: 96.40% accuracy and 97.83% recovery.

Counterfeit operations are causing great harm, which prompted investigators to look for a solution that could detect and prevent fraud. Several methods have been suggested and tested. Some of them reviewed below. Classical algorithms such as Gradient Increase (GB), Support Vector Machines (SVM), Bit Tree (DT), LR, and RF have proven useful. A combination of GB, LR, RD, SVM, and individual classifications used in Document [13], resulting in the recovery of over 91% of the European dataset. Higher accuracy and improvement were achieved only after balancing the specified data by sampling the data. In Article [14], a set of European data also used and a comparison made between the LR, DT, and RF models. In all three models, the radio frequency shown to be optimal, with 95.5% accuracy, followed by DT with 94.3% and LR with 90% accuracy.

III. PROPOSED ASGBALGORITHM

Advanced Super Gradient Boosting (ASGB), determined by Chen et al., is one variety of GBM model. And it is an implementation of gradient boosting machines that strains the boundaries of computing power for advanced trees algorithms as it was developed and produced for the individual purpose of design representation and computational speed. Primarily, it managed to utilize every bit of memory and hardware devices for boosting. ASGBallows various high-level features for model tuning, computing conditions, and algorithm improvement. It is proficient in achieving the three primary forms of gradient boosting (Gradient Boosting (GB), Stochastic GB, and Regularized GB), and it is robust sufficient to support fine-tuning and enhancement of regularization parameters. Both ASGB and GBM support the principle of gradient boosting, but there are variations in modeling details. Mainly, ASGB employs a more regularized pattern formalization to check over-fitting, which gives a better representation. ASGB uses second derivative information, and ordinary GBM uses only first-order derivatives. ASGB models greatly enhance the traditional gradient boost model, and it is the fastest learning algorithm of the gradient boost algorithm. In this paper, we use Advanced Super Gradient Boosting (ASGB) algorithm to detect fake credit card transactions in a real-world (anonymous) dataset of European credit card transactions, and I compare the proposed algorithm to the previous one. What I do and I show is that the well-adjusted ASGB classifier performs better on them all.

ASGB is a decision tree-based Text mining algorithm that uses a gradient reinforcement framework. In predictive problems involving non-structured data (images, text, etc.), artificial neural networks outperform all other algorithms or frames. However, when it comes to small to medium structured/tabular data, the decision tree-based algorithm is currently considered the best in its class. It is an algorithm that has recently mastered Text mining skills applicable to structure or tabular data. ASGB is an implementation of Gradual Gradient trees designed for speed and efficiency.

Algorithm: Advanced Super Gradient Boosting

$$\boldsymbol{Input}: Training\ Set\ S = \{x_i, y_i\}, i = 1, \ldots . N; and\ y_i \in \mathbb{CC} = \{c_1, \ldots \ldots, c_m\}; T$$

$$number\ of\ iterations; I: Weak\ Leaner$$

$$\boldsymbol{Output}: Boosted\ Classifier \boldsymbol{Step1}: Different\ from\ GBM\ ,$$

ASGB tries to determine the step directly by solving

$$\frac{\partial L(y, f^{(m-1)}(x) + f_m(x))}{\partial f_m(x)} = 0$$

***Step2**: **for each** x in the data set. By doing second order Taylor expansion of the loss function around the current estimate $f(m-1)(x)$, we get*

$$L\left(y, f^{(m-1)}(x) + f_m(x)\right) \approx L\left(y, f^{(m-1)}(x)\right) + g_m(x) f_m(x) + \frac{1}{2} h_m(x) f_m(x)^2$$

***Step3**: **where** $g_{m(x)is}$ the gradient, same as the one in GBM, and $h_{m(x)is}$ theHessian (second order derivative) at the current estimate:*

$$h_m(x) = \frac{\partial^2 L(Y, f(x))}{\partial f(x)^2} \; f(x) = f^{(m-1)}(x)$$

***Step4**: **Then** the loss function can be rewritten as*

$$L(f_m) \approx \sum_{i=1}^{n} [g_m(x_i) f_m(x_i) + \frac{1}{2} h_m(x_i) f_m((x_i)^2] + const.$$

$$\propto \sum_{j=1}^{T_m} \sum_{i \in R_{jm}} [g_m(x_i) w_{jm} + \frac{1}{2} h_m(x_i) w_{jm}^2]$$

***Step5**: Letting G_{jm} represents the sum of gradient in region j and H_{jm} equals to the sum of hessian in regionj, the equation can be rewritten as*

$$L(f_m) \propto \sum_{j=1}^{T_m} [G_{jm} w_{jm} + \frac{1}{2} H_{jm} w_{jm}^2]$$

***Step6**: With the fixed learned structure, for each region, it is straightforward to determine the optimal weight*

$$w_{jm} = -\frac{G_{jm}}{H_{jm}}, j = 1 \ldots\ldots, T_{m.}$$

***Step7**: Plugging it back to the loss function, we get*

$$L(f_m) \propto -\frac{1}{2} \sum_{j=1}^{T_{m.}} \frac{G_{jm}^2}{H_{jm}}$$

***Step8**: Taking regularization into consideration, we can rewrite the loss function as*

$$L(f_m) \propto \sum_{j=1}^{T_m} \left[G_{jm} w_{jm} + \frac{1}{2} H_{jm} w_{jm}^2\right] + \gamma T_m + \frac{1}{2} \lambda \sum_{j=1}^{T_m} w_{jm}^2 + \alpha \sum_{j=1}^{T_m} |w_{jm}|$$

$$\sum_{j=1}^{T_m} [G_{jm} w_{jm} + \frac{1}{2} (H_{jm} + \lambda) w_{jm}^2 + \alpha |w_{jm}| + \gamma T_m]$$

A. SYSTEM ARCHITECTURE:

1. Raw Data: The collected input data is in the form of csv files.

2. Prepared Data: This Method is for cluster context for input records. Understand statistics for pre-processing and cleaning data sets. The "quantity" and "time" columns are no longer regular. The final columns standardized the use of the relevant Principal component analysis (C-PCA). The "quantity" and "timescale" functions are between -1 and 1, using standardized records as used in the Gaussian distribution.

3. Oversampling: Frauds are 492 unbalanced samples. Therefore, fraud cases dispensed with the use of artificial minority seizure technology of SMOTE.

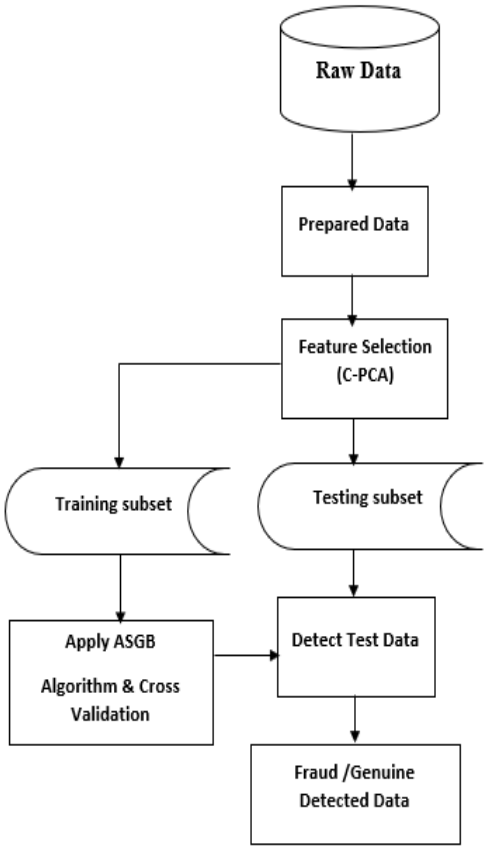


Fig.1 Proposed architecture

4. Training and Testing Subset: Since the dataset is unbalanced, many workbooks show bias to majority instructions. Minority capabilities are treated as noise and ignored. Therefore, miles suggested for choosing a set of style data.

5. Applying ASGB algorithm: Below are the classification algorithms used to test the sub form data set. Think of ASGB as a gradual increase in "steroids" (of something called "advanced super-gradient boosters"). It is the perfect mix of software and hardware optimization technologies to produce superior results with fewer computing resources in the shortest time possible.

6. Fraud Detection results: The subtest group applies to the trained model. The measures used are the degree of accuracy and freshness. The ROC curve is drawn, and the desired results obtained.

In this paper, the proposed ASGB algorithm is compared with previous fraud detection algorithms of Hidden Markov Model (HMM), RF (Random Forest), Gradient Boosting (GB), Enhanced Hidden Markov Model(EHMM).

B. EXPERMENTAL RESULTS AND DISCUSSION

The dataset contained 284,807 transactions carried out by credit cardholders in September 2013 for two days. There are 492 fraudulent transactions, and for this reason, the data set is not entirely balanced. The class (fraud) represents 0.172% of all trades. It consists of the best numerical input variables that can result from a C-PCA conversion. Unfortunately, due to confidentiality issues, unique functions and historical information are not given to the facts. The characteristics V1, V2, V28 are the main additions obtained with C-PCA, and the best functions that are not converted using C-PCA are "time" and "quantity." The "time" function includes the seconds that elapsed between each transaction and the first transaction in the data set. The attribute "Class" is the interaction variable and takes price 1 in the event of fraud or else equals 0. The Dataset taken from ULB (Université Libre de Bruxelles) (https://www.kaggle.com/mlg-ulb/creditcardfraud).It can be seen that the amounts in fraudulent transactions have always been less 2,500.

Data Pre-processing

At this point, after analyzing the dataset, it was interpreted that all columns, except for quantity and time, were measured using the C-PCA transformation technique. Therefore, time and quantity columns are measured using dimensional reduction technology to ensure uniformity.

Table.1 Scaled Amount and Time column in credit card dataset

scaled_ amount	scaled_ time	V1	..	V28	Class
1.783274	-0.994983	-1.359807	..	-0.021053	0
-0.269825	-0.994983	1.191857	..	0.014724	0
4.983721	-0.994972	-1.358354	..	-0.059752	0
1.418291	-0.994972	-0.966272	..	0.061458	0
0.670579	-0.994960	-1.158233	..	0.215153	0

$[5\ rows \times 31\ columns]$

IV. METRICS USED

The output of the metrics depends on the results obtained by$True\ positive\ (TP), True\ Negative\ (TN)$, False Positive (FP), False Negative (FN).

True Positive(**TP**): The transaction cases which are not fraud and the system model has predicted as not fraud

True Negative(**TN**): The transaction cases which are fraud and the system model has predicted as fraud

False Positive(**FP**): The transaction cases which are fraud and the system model has predicted as not fraud

True Negative(**TN**): The transaction cases which are not fraud and the system model has predicted as fraud

Table.2 Shows the ROC for the obtained readings

TP	FP	TN	FN	FPR	TPR
33	25	30	12	0.454545	0.733333
127	187	102	84	0.647059	0.601896
411	187	275	160	0.404762	0.71979
623	285	408	184	0.411255	0.771995
749	392	551	308	0.415695	0.708609
867	448	819	366	0.353591	0.703163
1034	522	1093	351	0.32322	0.74657

Also table shows the TPR and FPR readings.

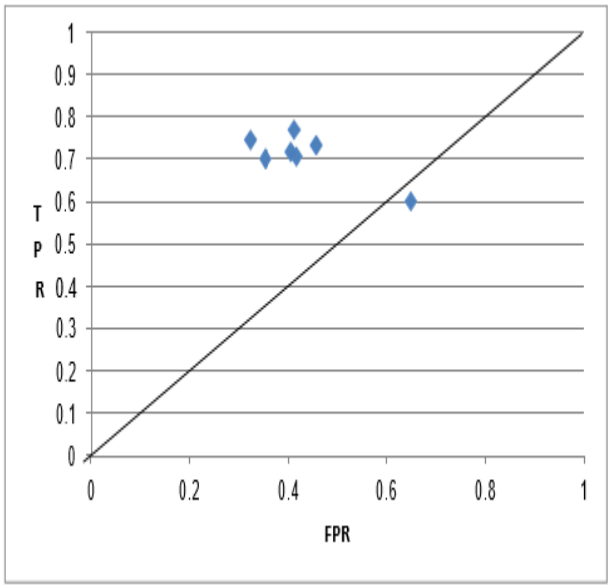


Fig.2 ROC- ASGB

As shown in Fig.2 the ROC- ASGBin this work, training samples are used in different proportions. Increasing the size of negative groups will not affect the outcome. In addition, all different samples predict a similar accuracy rate.

Most used metrics for determining the results of Text mining algorithms are accuracy, recall and precision. All of the mentioned metrics can be calculated from a Confusion matrix.

Our model achieved nearly 99.5% accuracy, with 99.1% precision (positive predictive value) and 99.8 recall (sensitivity). We can see there are only 6 false negatives. This means that the baseline model will be very hard to beat.

Table.3 Comparison of various algorithms in Accuracy, recall, and precision parameters

Metrics	HMM	RF	GB	EHMM	ASGB
Precision	89.25	88.68	86.15	95.15	99.10
Recall	92.64	93.26	90.24	97.26	99.80
F-measure	92.91	94.52	96.26	97.85	99.50

Precision:

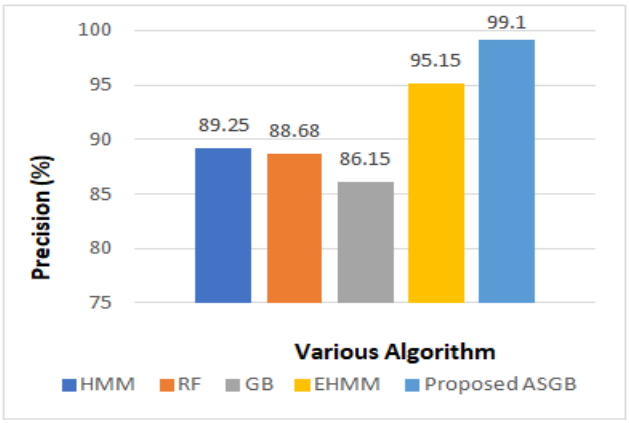


Fig.4 Precision (%) Vs Various Algorithms

Recall:

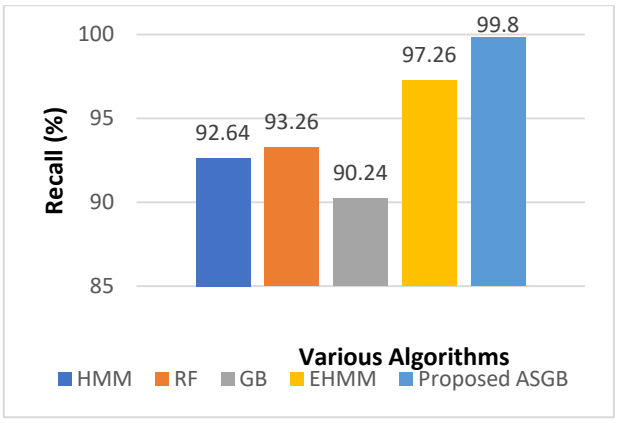
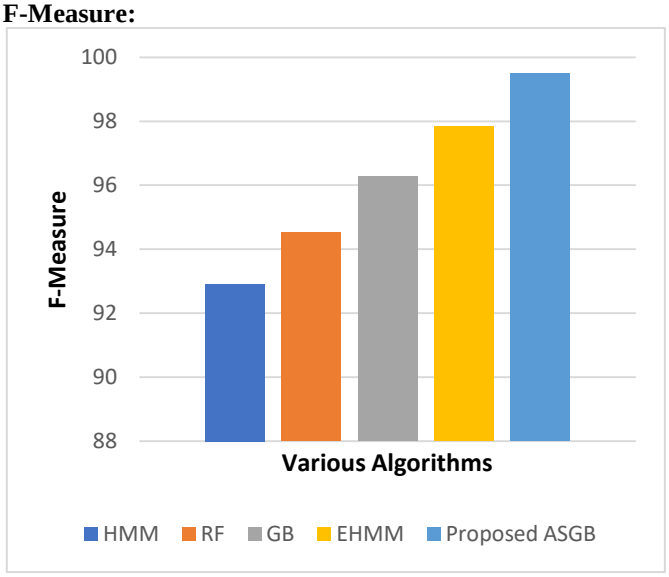


Fig.5 Recall Vs Various Algorithms

F-Measure:



As shown in Fig.3, Fig.4, and Fig.5, the Proposed ASGB experimental taken in parameters of Accuracy, Precision, and Recall calculated and compared with previous algorithms of HMM, RF, GB, EHMM, and with the analysis we can say that the proposed algorithm achieved high Accuracy rate compare to previous algorithms. The proposed algorithm got better accuracy by improved Precision is 99.1%, and Recall is 99.8%, and F-measure 99.5%,

V. CONCLUSION

Credit card frauds represent a severe business problem. These tricks can lead to huge losses, both commercial and personal. Therefore, companies invest more and more money in developing new ideas and methods that help detect and prevent fraud. The main objective of this work is to compare some of the text mining algorithms to identify fraudulent transactions. Therefore, a comparison made and the proposed ASGB algorithm provided the best results, that is, it better categorized whether the transactions were fraudulent or not. It was created using different metrics, such as callback, accuracy, and accuracy. For this type of issue, it is essential to have a high-value draw. The features selection and the balance of the dataset have proven critical to achieving important results. And the proposed algorithm compared with existed algorithms of Hidden Markov Model, Random Forest, Gradient Boosting, Enhanced Hidden Markov Models with the results we can say the proposed algorithm improve the better accuracy compared to previous algorithms. The proposed system is help to detect and before prevent the fraudulent transaction and activities, so to reduce the number of losses in financial industry.And it provided better Precision is 99.1%, and Recall is 99.8%, and F-measure is 99.5%.

REFERENCES

1. Global Facts, 2019, “Topic: Startups worldwide. [Online] Available at: https://www.statista.com/topics/4733/startups-worldwide/”.
2. Legal Dictionary, 2019. “Fraud - Definition, Meaning, Types, Examples of fraudulent activity”.
3. European Central Bank, 2018,” Fifth report on card fraud”,
4. En.wikipedia.org, 2019,.”Credit card fraud. [online] Available at: https://en.wikipedia.org/wiki/Credit_card_fraud [Accessed 24 Jan. 2019].
5. S.Ghosh, D.L.Reilly, 1994, “Credit card fraud detection with a neural–network”, pp. 621-630.
6. Aleskerov E,Rao B,1997, “CARDWATCH: a neural network-based database mining system for credit card fraud detection”, pp.220-226.
7. R J..Dorronsoro, Cruz C.S., 1997, “Neural fraud detection in credit card operations”, pp. 827-834.
8. M. Syeda, Y. Pan,2002, “Parallel granular neural networks for fast credit card fraud detection”, pp. 572-577.
9. P.K. Chan, W. Fan, A.L. Prodromidis, S.J. Stolfo, “Distributed Text mining in credit card fraud detection”, in: Proceedings of the IEEE Intelligent Systems, 1999, pp. 67-74.
10. A. Srivastava, Kundu A , 2008, “Credit card fraud detection using hidden markov model”.
11. Panigrahi Suvasini, A. Majumdar K., “Credit cardfraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning”.
12. Wang C., Wang, S. Pan, “Credit card fraud detection based on whale algorithm optimized BP neural network”, pp. 1-4.
13. A. Mishra, 2018, “Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques”, pp. 1-5.
14. S. V. S. S. Lakshmi, S. Kavilla D. “Text miningfor Credit Card Fraud Detection System”.

AUTHOR PROFILES

V. Sudheer Goud, Post Graduated in Master of Computer Application **(MCA)** from OU,1994 , Post Graduated in Master of Business Administration **(MBA**) from OU,2006, Post Graduated in Master of Computer Science & Engineering **(M.Tech)** from IETE , Hyderabad in 2013 and Pursuing Phd in Computer Science in ANU. He is currently working as an Associate Professor, Department of Computer Science in Holy Mary Institute of Technology and Science (HITS), (V) Bogaram, (M) Keesara, Medchal .Dist, Telangana, India. He has 25 years. Of Teaching Experience. His research interests include, Data Mining, Cloud Computing and Information Security.

Prof. P. Premchand, He is currently working as a Professor, Department of Computer Science and Engineering, University College of Engineering, Osmania University, Hyderabad, Telangana State, India