A SURVEY ON TRANSACTION FRAUD DETECTION USING DATA MINING TECHNIQUES

[1]N.Geetha, [2]Dr.G.Dheepa,

[1]Assistant Professor , [2]Assistant Professor ,

[1]Department of Computer Science, [2]Department of Computer Science,

[1]Gobi Arts & Science College, [2]P.K.R College of Arts & Science For Women,

TamilNadu, India.

Abstract : Nowadays,credit card plays a vital part in economy. Credit cards turn out to be a necessary element of global, business and household activities. Even though the credit cards give vast advantages while utilized attentively and correctly, considerable credit and economic harms may be caused as a result of fraudulent actions. Data mining has commonly obtained recognition in managing credit-card fraud owing to its efficient, and machine learning schemes that can be executed to identify or predict fraud by Knowledge Discovery from abnormal models obtained from collected data. Credit Card Fraud Detection (CCFD) system is extremely important, for the purpose of identifying fraud correctly and prior to the occurrence of fraud. Various algorithms like k-means clustering, Genetic Algorithms, neural networks and Hidden Markov Model are used for detecting fraudulent usage of credit cards. The benefits and difficulties of CCFD schemes are detailed and evaluated in this paper. Besides, evaluation conditions in literature are gathered and described. Accordingly, open complications for CCFD are clarified and the future ways of the CCFD industry is provided.

Keywords: Credit Card, Fraud Classification, Fraud Detection Techniques, Data mining.

I. INTRODUCTION

Credit card is the maximum utilized payment mode ,while comparing other modes like e-wallet and bank transfer [1]. Moreover, the vast transactional services are frequently watched by criminals to carry out fraudulent actions with the assistance of credit card services. Credit card fraud is described as the unofficial utilization of card, abnormal transaction activity, or dealings on an unused card. Credit card frauds are commonly classified into three types, merchant related frauds [2], online frauds, and conventional frauds. In recent past, credit card violations have been raising alarmingly. Consequently, it is essential to advance CCFD techniques as the preventive action to manage criminal activities [3]. Moreover, CCFD has been called as the method of recognizing whether transactions are genuine or fraudulent.

At the time of purchasing, the client uses the credit card, the fraudster find out the code word or user related vital details which is later used by him for fraudulent transaction [4, 5]. The credit card transaction is done by means of online or offline. In offline transaction,the cardholders fundamentally give the vital details like, card number ended date and card validation number via phone or web [6].

The numbers of credit card business are raised in each year. At this point, the technology is developed and obtains additional gain for the people, but other way it raises this credit card fraud cases [7]. The logical and numerical authentication schemes are applied in this credit card fraud cases. But the fraudsters usually conceal their details such as, identity and locality in the internet [8]. This credit card fraud complication influences both sides admin and user side. It influences the (a) issuer fees, (b) charges, (c) administrative charges, that are the fees are loss [9]. Therefore, the merchants make the conclusion that is elevated rate fix in goods or concessions are diminished.

Data mining is a process of detecting and separating knowledge from massive data sets .It is a constant process which is used for the purpose of scanning, throughout vast collections of data in an attempt to find out supportive details [11]. The fundamental operation of data mining is extracting the significant designs and associations in huge data sets. Artificial intelligence, arithmetic calculations, and machine learning approaches to identify designs are employed by data mining for the purpose of taking decisions, likely results and create actionable information.

The main motive of this article is to examine different machine-learning methods, like Hidden Markov Model (HMM), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB) and Multilayer Perceptron (MLP) with the aim of deciding which method is most appropriate for CCFD.

Here, the further sections of this research work are constructed as: in Section II provides a list of existing work in this field, Section III described the benefits and difficulties of the current fraud detection techniques, Section IV deals with the result and discussion. Lastly, concluding statements and future directions are described in Section V.

II. LITERATURE REVIEW

Fraudulent behaviors are the major reasons for loss, which encourages and to discover a result that possibly will identify and avoid frauds. Some techniques have beenpreviously formulated and analyzed. A few of them are reviewed briefly as follows.

The author, Whitrow et al [12] developed a structural arrangement for the purpose of transacting aggregation is taken into account and its efficiency is estimated against transaction-level detection, utilizing a various types of categorization techniques and a realistic cost-dependent performance evaluation measure. These techniques are implemented in two case studies with the assistance of actual data. Transaction aggregation is found to be useful in several but not every situation. Furthermore, the length of the aggregation time duration has a huge impact over performance. Aggregation looks especially efficient when a random forest is utilized for categorization. Furthermore, random forests were found to execute better when comparing to remaining categorization techniques, including logistic regression, KNN and SVMs. Aggregation also has the benefit of not involving accurately tagged data and might be additional robust to the outcomes of population drift.

Van Vlasselaer et al [13] introduced Anomaly Prevention with the help of Advanced Transaction Exploration (APATE), is a new way to identify fraudulent transactions performed during online storage. This type of process unites (1) intrinsic attributes obtained from the attributes of inward transactions and the consumer expenditure records with the assistance of the fundamentals of RFM (Recency - Frequency - Monetary); and (2) network-dependent characteristics through utilizing the network of credit card holders and merchants in addition gaining a time-dependent apprehensive rank for every object. Thus, the outcomes confirm that mutually intrinsic and network- dependent characteristics are two robustly entwined parts of the same image. The integration of these two categories of characteristics guides to the most excellent performing designs that attain AUC-scores in excess of 0.98.

Fashoto et al [14] formulated a design for fraud detection scheme which possibly will effort to successfully identify credit card fraud by making clusters and examining the clusters produced as a result of the dataset for anomalies. Here, the main intention of this investigation is to assess the working potential of two hybrid schemes based on the detection accuracy. In this approach, hybrid techniques are employed by utilizing the K-means Clustering scheme with MLP together with the HMM. The analysis discovered that the finding accuracy of "MLP together with K-means Clustering" is greater than the "HMM together with K-means Clustering" for 80% split however the overturn is the scenario when the "MLP together with K-means Clustering" is evaluated by means of the "HMM together with K-means Clustering" for 10 fold up cross-validation however the accuracy is similar in the two hybrid approaches for proportion split of 66%. In additional, wide-ranging assessment with further huge datasets is on the other hand needed to authenticate these outcomes.

Halvaiee et al [15] proposed the CCFD with the assistance of Artificial Immune Systems (AIS), and establish a novel model regarded as AIS-based Fraud Detection Model (AFDM). Here in this scheme, will employ an immune system stimulated method (AIRS) and develop it for the purpose of fraud identification. In this approach, enhanced the accuracy to 25%, decrease the cost nearly 85%, and reduce system response time to 40% when evaluated against the base method.

Dal Pozzolo et al [16] formulated a method of efficient fraud detection approaches is key for minimizing the losses, and more schemes depend on advanced machine learning schemes in order to help fraud investigators. In this article, offered some responses from the practitioner's perception by means of concentrating on three important problems: assessment, non-stationarity and unbalancedness. This investigation is made achievable by a real credit card dataset given through the industrial partner.

Zareapoor et al [17] put forward the bagging classifier in terms of decision tree, as the excellent classifier to create the fraud detection approach. On the other hand, the most frequently used schemes for fraud detection methods are KNN, SVM and NB. These methods can be utilized independently or in combination utilizing ensemble or meta-learning methods for the purpose of constructing classifiers. However, while considering the entire existing schemes, ensemble learning approaches are recognized as well-liked and standard scheme, not due to its relatively uncomplicated implementation, however as well owing to its excellent predictive potential on practical issues. Here, the performance assessment is executed on original credit card transactions dataset for the purpose of demonstrating the advantage of the bagging ensemble method.

Bahnsen et al [18] developed the transaction aggregation scheme, in addition recommended to generate a novel set of characteristics in accordance with examining the periodic ways of the transaction time by means of the von Mises distribution. Subsequently, with the assistance of an authentic credit card fraud dataset offered by huge European card processing organisation, evaluate state-of-the-art CCFD models, in addition assess how the various sets of attributes have an influence on the outcomes. Through the process of integrating the formulated periodic attributes into the techniques, the outcomes demonstrated a standard raise in savings of 13%.

Şahin et al [19] formulated a fraud avoidance systems, is the vital system and perhaps the excellent manner to end the fraud categories. Here, classification models in terms of decision trees and SVM are formulated and implemented on CCFD issues. This system is one of the firsts to evaluate the working potential of SVM and decision tree schemes in CCFD by means of a factual data set.

Srivastava et al [20] developed an approach, the series of processes in credit card transaction procedure with the assistance of a HMM and illustrate, how it can be employed for the discovery of frauds. Here, HMM is mainly instructed with the common activities of a cardholder. When an inward credit card transaction is not recognized by means of the trained HMM with sufficiently high possibility, it is regarded as fraudulent. On the other hand, attempt to make sure that factual transactions are not eliminated. Accordingly, the experimental outcome illustrates that the efficiency of this scheme and evaluate it with other schemes.

Jha et al [21] developed transaction aggregation approach for the purpose of identifying credit card fraud. Here, the integrated transactions to obtain user obtaining activities earlier than every transaction and utilized these aggregations for design evaluation to recognize fraudulent transactions. Here, employed original data of credit card transactions from a global credit card process.

Quah et al [22] concentrated on instantaneous fraud detection and provided a novel and inventive approach in recognizing spending behaviors to interpret potential fraud cases. It utilizes self-organization map for the purpose of interpret, filter and examine consumer actions for fraud detection. However, this has enabled it simpler for fraudsters to take part in novel and mysterious manners of doing credit card fraud over the Internet.

Abdulla et al [23] formulated a hybrid scheme engaged phases of pre-processing where unspecified transactions are eliminated, GA modeled for feature selection and SVM for categorization. In this article, explains an easy fraud detection scheme that can efficiently discover fraud with vast accuracy. It is pointed out that, each and every phase are successfully executed to the dataset and generated a better model for identifying fraud. SVM illustrates excellent accuracy in the formulated scheme by classifying the test data to fraud and legal correspondingly. Here, this accuracy is attained by the feature selection procedure that has been designed with the assistance of K Nearest Neighbour system.

Mhatre et al [24] developed an approach, the series of processes in credit card transaction procedure with the assistance of a HMM and illustrate, how it can be employed for the discovery of frauds. Here, HMM is mainly instructed with the common activities of a cardholder. When an inward credit card transaction is not recognized by means of the trained HMM with sufficiently high possibility, it is regarded as fraudulent. On the other hand, attempt to make sure that factual transactions are not eliminated. Accordingly, the experimental outcome illustrates that the efficiency of this scheme and evaluate it with other schemes.

Patidar et al [25] formulated a neural network together with the GA based model for discovery of credit card fraud. Here, the research technique fused, Genetic Algorithm and Neural Network (GANN), attempted to identify the credit card fraud productively. On the other hand, the concept of integrating NN and GA come from the information, that when a person is naturally extremely expert and he is trained well then possibilities of individual of achievement is extremely huge. GA are utilized for building the judgment regarding the network topology, amount of concealed layers, amount of nodes that will be utilized in the formulation of neural network for the issues of CCFD. Here, the outcomes point out that the formulated detection scheme gives better fraud detection method evaluate against the current detection scheme.

Baboo et al [26] provided a model with the assistance of HMM. In this article, with this authenticated security verify the information of transaction is fake or authentic. This method is extremely protected from illicit anomalous consumer utilizing credit card and neglect fraud handling of card by online transactions. Experimental outcomes illustrate the performance and efficiency of the expenditure pattern of the cardholders. Here, the scheme is also scalable for managing huge amounts of transactions.

Zheng et al [27] introduced a Logical Graph of Behavior Profiles (LGBP) that is overall order-dependent design to signify the rational association of qualities of transaction files. In accordance with LGBP and users' transaction files, can determine a path-dependent transition possibility from a feature to an extra one. On the other hand, described an information entropy-dependent diversity coefficient with the aim of describing the multiplicity of transaction performances of a consumer. Besides, describe a state conversion probability matrix for the purpose of capturing temporal characteristics of transactions of a consumer. Accordingly, can make a BP used for every consumer and subsequently employ it to confirm, if an inward transaction is a fraud/not. Thus, the researches over a factual data set show that this technique is superior to three standard schemes.

III. COMPARISON ANALYSIS

Table 1 shows the comparison analysis for different credit card detection techniques.

S. No	Author	Method	Merits	Demerits
1	Srivastava et al (2008)	Hidden Markov Model (HMM)	Easily function on huge databases	It can be responsive to loud data and outliers
2	Şahin et al (2011)	SVM and Decision Tree	The accuracy performance of SVM based models reach high than the performance of the decision tree based models.	High computational difficulty.
3.	Patidar et al (2011)	Genetic Algorithm and Neural Network (GANN)	GANN strategies rely only on the GA to find an optimal network; in these, no training takes place.	It can be responsive to loud data and outliers.
4.	Dal Pozzolo et al (2014)	EasyEnsemble based incremental Learning	This approach has showed better results than updating the models at a lower frequency (weekly or every 15days).	This algorithm not preferable for nonlinear data.
5.	Zareapoor et al (2015)	Bagging classifier using KNN, SVM and NB	It can manage thousands of input variables exclusive of variable removal.	It is not adequately potential to manage fraud detection at the time of transaction.
6.	Abdulla et al (2015)	Hybrid approach SVM and K Nearest Neighbour approach.	Functions fast and well on online huge datasets.	This scheme wants dependent and independent attributes.
7.	Van Vlasselaer et al (2015)	APATE	Using less memory. Computation is needed.	It is over fit for classification /regression tasks with louddataset

8.	Bahnsen et al (2016)	Transaction Aggregation Strategy	Work better on linear dataset.	Highly cost effective.
9.	Fashoto et al (2016)	Hybrid methods using the K-means Clustering algorithm with MLP and HMM.	It is a powerful classifier that functions well on both basic and more complex detection complication.	Excessive training need / costly.
10.	Zheng et al (2018)	Logical Graph of Behavior Profiles (LGBP)	High processing and detection speed/high accuracy.	It is good if dataset has plenty of input but little amount of errors.

Various methods such as, Genetic Algorithm and Neural Network (GANN) and Hidden Markov Model (HMM) however having elevated detection charges and provides better accuracy, they are especially costly to teach.

IV. RESULTS AND DISCUSSION

Here, this section shows the performance of the new methods formulated in CCFD. Nowadays, credit card data are usually unavailable. Even though there are few unrestricted data sets regarding CCFD like, the one in https://www.kaggle.com/dalpozz/creditcardfraud, the records have been altered, and it is not possible to identify the majority of their attributes [27]. Specifically, each consumer has numerous records in those data sets.

In this paper, evaluated various methods estimate the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) produced by a scheme and employ these in quantitative dimensions to assess and evaluate performance of various systems. TP is amount of transactions that were fraudulent and were also labeled as fraudulent by the system. TN is amount of transactions which were valid and were also categorized as genuine. FP is amount of transactions that were valid however were incorrectly categorized as fraudulent transactions. FN is amount of transactions that were fraudulent however were incorrectly categorized as valid transactions. Here, the different metrics are employed for assessment, specifically, F-measure, Precision, Recall and Accuracy.

1. Accuracy indicates the fraction of transactions which were properly categorized.

$$Accuracy=(TN+TP)/(TP+FP+FN+TN)$$

2. Precision also regarded as detection rate, that is the amount of transactions either genuine/fraudulent which were correctly categorized.

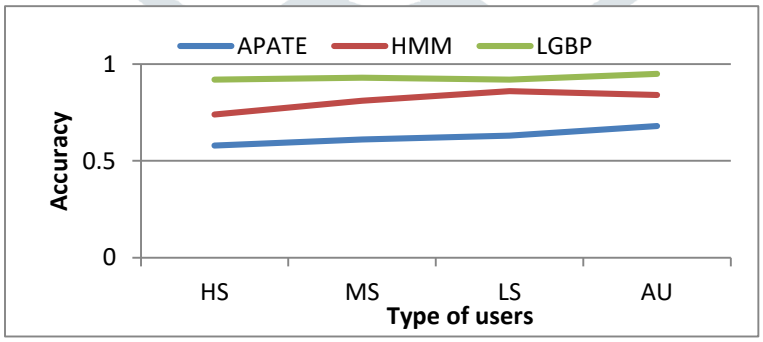
$$Precision=TP/(TP+FP)$$

3. Recall determines the fraction of uncharacteristic records correctly categorized by the system.

$$Recall=TP/(TP+FN)$$

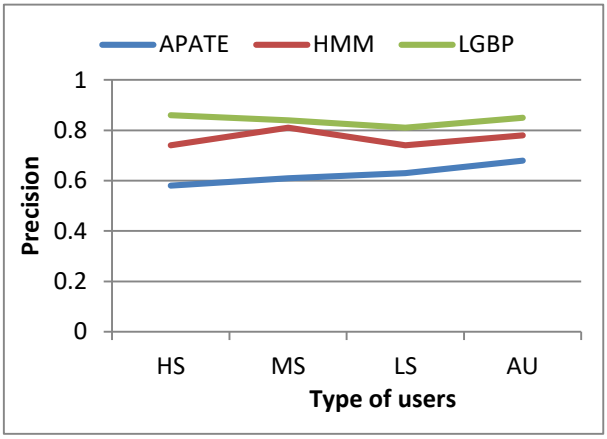
4. The F score, also known as the F1 score or F measure, is a measure of a test's accuracy. It is described as the weighted harmonic mean of the test's precision and recall.

$$F\text{-}measure=2*Precision*Recall/(Precision+Recall)$$



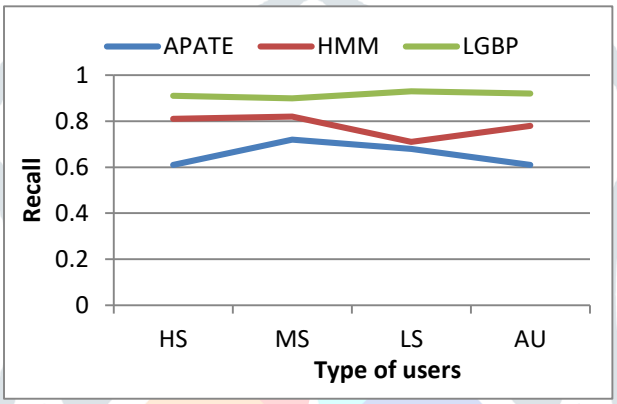
Figue.1. Accuracy comparison results of different fraud detection techniques

Here, the transaction records are categorized into three kind`s in accordance with transaction amount: low amount (LA), medium amount (MA), and high amount (HA). Accordingly, partition consumers into three different sets: high stability (HS), medium stability (MS), and low stability (LS) and All Users (AU) in view of the allocation of a user's transaction records in LA, MA, and HA. The figure 1 illustrates the accuracy comparison outcomes of the various methods for CCFD together with the different category of users. Therefore, the accuracy of the LGBP method is high when comparing against the existing fraud detection schemes.



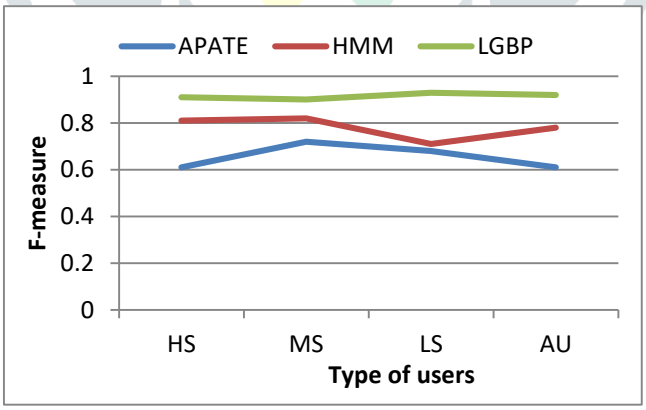
Figue.2. Precision comparison results of different fraud detection techniques

Here, the performance of the different CCFD methods is shown in Fig. 2. The outcomes illustrate that APATE and HMM are poorer to LGBP. Finally, it ends that the LGBP methods have better precision results where the idea of drift takes place.



Figue.3. Recall comparison results of different fraud detection techniques

The existing CCFD schemes were implemented to different type of consumers as explained in fig 3, and the LGBP was validated using actual financial transaction dataset. Furthermore, for more accurate validation performs the validation procedure with more dataset. The LGBP technique has better recall outcomes evaluated to the other schemes.



Figue.4. F-measure comparison results of different fraud detection techniques

Figure 4.illustrates the average of F-measure by means of different fraud detection methods in various consumers. Three kinds of classification algorithms were utilized and partition the outcomes with different users for more specified information. On behalf of every consumer the detection rate of the classification scheme was measured in terms of average detection rate of the previous fraud detection scheme.

V.**CONCLUSION**

Even though there are some fraud detection methods accessible at present however none is capable of identifying the entire frauds fully when they are really occurring, they typically notice it once the fraud has been happened. In this paper, this occurs because a very little amount of transactions from the entire transactions are literally fraudulent in character. Consequently, the main job is to construct a perfect, strict and rapid noticing CCFD method that can identify frauds occurring over the internet. Here, several of the current CCFD methods are reviewed together with the merits and demerits are presented. The major

disadvantage of the entire methods is that they are not sure to give the similar outcomes in all circumstances. They provide better outcomes with a specific kind of dataset and poor or inadequate results with additional type. Even though, a few methods such as, HMM and SVM provides outstanding results with small data sets however are not scalable to huge datasets. A few methods like aggregation strategy and support vector provides improved outcomes on sampled and pre-processed data while, a few systems such as, logistic regression and fuzzy systems provide improved accuracies with unprocessed ensample data. Finally, these gaps could be bridged by means of generating a hybrid of different methods that are previously utilized in fraud identification to remove their limitation in addition obtains improved performance.

REFERENCES

1. Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C., HienTran, P., & Le, T. M. H. (2018, February). Real time data-driven approaches for credit card fraud detection. In *Proceedings of the 2018 International Conference on E-Business and Applications* (pp. 6-9).
2. Mekterović, I., Brkić, L., & Baranović, M. I. R. T. A. (2018). A systematic review of data mining approaches to credit card fraud detection. *WSEAS Transactions on Business and Economics*, *15*, 437.
3. Puh, M., & Brkić, L. (2019, May). Detecting Credit Card Fraud Using Selected Machine Learning Algorithms. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1250-1255). IEEE.
4. Manlangit, S., Azam, S., Shanmugam, B., Kannoorpatti, K., Jonkman, M., & Balasubramaniam, A. (2017, December). An efficient method for detecting fraudulent transactions using classification algorithms on an anonymized credit card data set. In *International Conference on Intelligent Systems Design and Applications* (pp. 418-429). Springer, Cham.
5. Dhankhad, S., Mohammed, E., & Far, B. (2018, July). Supervised machine learning algorithms for credit card fraudulent transaction detection: a comparative study. In *2018 IEEE International Conference on Information Reuse and Integration (IRI)* (pp. 122-125). IEEE.
6. Mittal, S., & Tyagi, S. (2019, January). Performance evaluation of machine learning algorithms for credit card fraud detection. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)* (pp. 320-324). IEEE.
7. Awoyemi, J. O., Adetunmbi, A. O., & Oluwadare, S. A. (2017, October). Credit card fraud detection using machine learning techniques: A comparative analysis. In *2017 International Conference on Computing Networking and Informatics (ICCNI)* (pp. 1-9). IEEE.
8. Patil, S., Nemade, V., & Soni, P. K. (2018). Predictive modelling for credit card fraud detection using data analytics. *Procedia computer science*, *132*, 385-395.
9. Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., & Jiang, C. (2018, March). Random forest for credit card fraud detection. In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)* (pp. 1-6). IEEE.
10. Rushin, G., Stancil, C., Sun, M., Adams, S., & Beling, P. (2017, April). Horse race analysis in credit card fraud—deep learning, logistic regression, and Gradient Boosted Tree. In *2017 systems and information engineering design symposium (SIEDS)* (pp. 117-121). IEEE.
11. Zheng, L., Liu, G., Luan, W., Li, Z., Zhang, Y., Yan, C., & Jiang, C. (2018, March). A new credit card fraud detecting method based on behavior certificate. In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)* (pp. 1-6). IEEE.
12. Whitrow, C., Hand, D. J., Juszczak, P., Weston, D., & Adams, N. M. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data mining and knowledge discovery*, *18*(1), 30-55.
13. Van Vlasselaer, V., Bravo, C., Caelen, O., Eliassi-Rad, T., Akoglu, L., Snoeck, M., & Baesens, B. (2015). APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions. *Decision Support Systems*, *75*, 38-48.
14. Fashoto, S. G., Owolabi, O., Adeleye, O., & Wandera, J. (2016). Hybrid methods for credit card fraud detection using K-means clustering with hidden Markov model and multilayer perceptron algorithm. *Brit. J. Appl. Sci. Technol.*, *13*(5), 1-11.
15. Halvaiee, N. S., & Akbari, M. K. (2014). A novel model for credit card fraud detection using Artificial Immune Systems. *Applied soft computing*, *24*, 40-49.
16. Dal Pozzolo, A., Caelen, O., Le Borgne, Y. A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert systems with applications*, *41*(10), 4915-4928.
17. Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. *Procedia computer science*, *48*(2015), 679-685.
18. Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. *Expert Systems with Applications*, *51*, 134-142.
19. Şahin, Y. G., & Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines.
20. Srivastava, A., Kundu, A., Sural, S., & Majumdar, A. (2008). Credit card fraud detection using hidden Markov model. *IEEE Transactions on dependable and secure computing*, *5*(1), 37-48.
21. Jha, S., Guillen, M., & Westland, J. C. (2012). Employing transaction aggregation strategy to detect credit card fraud. *Expert systems with applications*, *39*(16), 12650-12657.
22. Quah, J. T., & Sriganesh, M. (2008). Real-time credit card fraud detection using computational intelligence. *Expert systems with applications*, *35*(4), 1721-1732.
23. Abdulla, N., Rakendu, R., & Varghese, S. M. (2015). A hybrid approach to detect credit card fraud. *International Journal of Scientific and Research Publications*, *5*(11), 304-314.
24. Mhatre, G., Almeida, O., Mhatre, D., & Joshi, P. (2014). Credit card fraud detection using hidden markov model. *Int. J. Comput. Sci. Inf. Technol.(IJCSIT)*, *5*(3).

25. Patidar, R., & Sharma, L. (2011). Credit card fraud detection using neural network. *International Journal of Soft Computing and Engineering (IJSCE)*, *1*(32-38).
26. Baboo, C. D. S. S., & Preetha, N. Analysis of Spending Pattern on Credit Card Fraud Detection. *IOSR Journal of Computer Engineering (IOSR-JCE) Vol*, *17*, 61-64.
27. Zheng, L., Liu, G., Yan, C., & Jiang, C. (2018). Transaction fraud detection based on total order relation and behavior diversity. *IEEE Transactions on Computational Social Systems*, *5*(3), 796-806.