2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI), 17-18 December, Dhaka

Fraud Detection of Credit Card using Data Mining Techniques

Nishat Jahan Nishi*, Farhana Akter Sunny†, Sagor Chandro Bakchy‡
Dept. of Computer Science and Engineering, Green University of Bangladesh, Dhaka, Bangladesh*†
Dept. of Electrical and Computer Engineering, Rajshahi University of Engineering and Technology, Rajshahi, Bangladesh‡
Email: nishi@cse.green.edu.bd*, farhana@cse.green.edu.bd†, sagorchandro.10@gmail.com‡

***Abstract*—A credit card is an essential and daily used component in our life. Credit card fraud is a crime in our society. It causes many problems in our daily online transactions. Various deficiency of the previous method has been researched in the introduction chapter. The establishment of avoiding credit card fraud security is a big issue for small financial institutions. The goal of this paper is to propose a model for detecting credit card fraud and making a false alarm. For this research purpose, we have used four data mining techniques i.e. Decision Tree (DT), Random Forest (RF), Artificial Neural Networks (ANN), and Logistic regression (LR). Our motivation was how to reduce the false alarm so that the cardholder gets into less trouble and also the card provider gets more time for accurate alarming checking. After collecting the data from the Kaggle repository we look into the structure of the large dataset to check the performance. Datasets are filtered by the feature selection algorithms called Pearson correlation and chi-squared. We have introduced this model to get fewer false alarms in credit card fraud. From the proposed system, we can obtain about 99% accuracy (LR) and it also gives fewer false alarms. For tracking credit card fraud in the era of industry 4.0, our research work will ascertain the advancement and effectiveness of sustainable technologies.**

***Index Terms*—Credit card fraud, Data mining algorithms, Supervised data analysis, Feature selection, False alarm.**

I. Introduction

Industry 4.0 has fostered the accumulation of digital payment in E-commerce. At the same time, it is challenging to overcome fraudulent activities. From the business market to online technology, we are using online credit cards which is important for the new world. From the daily use of the internet, the information from the whole world and interlink hurdle was being interrupted [1]. Online fraud using credit cards can be executed in different ways. Different types of fraud are executed in different ways [2]. Online card transaction fraud has become the most popular and perilous fraud. Since online transaction using a credit card is becoming popular day by day in the modern world, crime has been established [3]. In the credit card industry, credit card fraud detection and avoiding fraud is the most important thing in the modern world for pitfall management [4]. From the credit card fraud study, we can say that many financial organizations are losing money from this fraud. And this kind of fraudulent losing money cannot be included in the dealer's bank account, and they do not get a chargeback and it is also a very big loss for the bank. This kind of loss adding to the bank association becomes very dangerous day by day and people are getting fraud. The most effective and also costly fraud from credit card fraud is Cardholder-not-present (CNP) fraud in this present time. In many countries, including Bangladesh, in 2012, this CNP-type fraud became more dangerous, and almost 118 US dollars are lost, which is more than in any other country [5].

The attempt of our research study is to find out the process of detecting online transactions in order to solve and minimize the increasing and effecting problems of credit card fraud [6]. We also discuss the fraud detection of credit cards using many data mining techniques. To increase fraud, we also can detect false alarm about the problem. Using data mining techniques we also can get detecting the false alarms of fraud detection.

Organization: We organize this paper as follows: **Section II** represents the related works. We have proposed our methodology and described the functionality in **Section III**. After that, **Section IV** illustrates the performance evaluation and analysis of the proposed system. Finally, in **Section V**, we terminate the writing with the value and future implications.

II. Related Works

Several authors have proposed various works for fraud detection in different sectors in recent years. This section presents some literature reviews of recent works:

To detect the fraudulent activities, an ensemble model using deep recurrent neural networks and a novel voting mechanism based on ANN have been proposed in [7]. For detecting suspicious transactions and also finding the outliers, Weston et al. declared the squint type analysis [8]. Also for detecting the wrongful online transaction, a genetic algorithm with the scatter search which can minimize the wrongful online transaction [1]. The genetic algorithm is also used to detect fraud of credit cards in [9]. A hybrid technique has been proposed to improve the accuracy of fraud detection in [10] and [11]. For detecting the fraud of the credit card researchers used training a dataset which is a multi-classifier for the distribution of the cardholder information [12]. Using the Hidden Markov Model (HMM), the researchers built a sequence using the dataset to train the dataset and also evaluate the normal behavior of the card holders and also proved how the fraud of the online transaction can be detected [13]. Also, the same perspective was proposed by the two other papers [14] [15]. An artificial immune system has been used for the

978-1-7281-6099-3/19/$31.00 ©2022 IEEE

fraud detection of online credit card transactions using the difference between consumer's information [6] [16]. Using association rules the researchers used to have the information about the fraud credit card holders to detect their fraudulent [17]. The cost-sensitive which is a decision tree that has been approved for detecting the online transaction of credit card fraud detection and cost sensitive decision tree is a tree where we discuss the cost amount of the sensitivity [5]. There is an effective algorithm for fraud detection which is Bayes minimum risk algorithm for the approaching of the fraudulent [18]. Another researcher has proved that the SVM algorithm is the best for detecting fraud in online transactions and they also used Gaussian classification with the kernel detection [19]. Many researchers used decision trees like CART, C4.5 and many more, they also used SVM like linear, sigmoid, polynomial and many more for fraud detection and they prove that the decision tree gives better accuracy than the SVM specially CART and CHAID algorithms [20]. Reverse K-Nearest-neighbour (KNN) was also used for the detection of fraud and they also used reversed KNN for the detection of the outlier detection of the data stream [21]. Day by day increasing online shopping fraud also increasing in our daily life. It is also a bigger threat as the business dealer has no such photo identification, or recheck confirmation, for searching the valid information of the consumers. There has no possibility for the detection of the card for physical verification. This makes the fraud for becoming more active and also makes the way easy for the fraud. The ratings for this problem have been increasing day by day. Many researchers use ANN with the combination of the outlier data stream for detecting credit card fraudulent [22]. For the customers' purposes, there are established fuzzy sets from the neural network for establishing the detection of the fraud using parallel dataset [23]. There is a comparative study among the SVM, logistic regression and ANN, they proved that SVM classification is giving better accuracy than two other algorithms [24]. There has been also used five different algorithms which have been CART, SVM with the classification kernel, LR, NN and Bayesian belief network and the SVM gives better result [25]. In a recent study, using a modified discrimination function, some linear algorithms' fraud detection became easier than any other algorithms [26]. In 2021, Rathore et al. used machine learning algorithms for comparative studies [27]. In 2022, There has been classification in fraudulant transaction Using MLP XGBoost algorithm [28].

III. Proposed Methodology

To investigate credit card fraud, we have analyzed data mining techniques and by using these techniques we have compared the performance. If the transaction is fraudulent, the genuine customer will get a notification as a consumer gets notified. The proposed work is shown briefly in Fig 1.

A. Data Collection

We have collected data from the website called Kaggle, UCI machine learning repository etc.

During the data collection, We found some kind of null data and also get some not important data which is not considered during the checking of fraudulent. Then the dataset is filtered by the feature selection algorithms called Pearson correlation and chi-squared. For Dataset 1, we have used the variables such as Cust Id, Average Amount/transaction/day, Transaction amount/day, Is Declined, Total Number of declines/day, Is Foreign Transaction, Is High Risk Country, Daily-charge-back-avg-amt, 6-months-avg-chbk-amt, 6-month-chbk-freq, Is Fraudulent (Data set 1 Link). For Dataset 2, we have used the variables such as Cust-Id, Balance, Balance frequency, Purchases, Installments purchases, Cash advance, Purchase frequency, Purchase installments frequency, Cash advance frequency, Purchases TRX, Cash advance TRX, Credit limit, Payment, Minimum payments, PRC full payments (Data set 2 Link).

After collecting datasets from the website, we merged it which is shown in table I. This helped us for getting better result and accuracy. And also a large information will give the better result for the false alarming reduction.

B. Feature Selection from the dataset

Our main study is how to predict credit card transaction fraud and stop fraud in the online era, which is a very composite abstract, which has to be diminished to the maximum dimensions. In this research, we have pre-owned two types of the feature selection algorithms. They are beneficial for getting better accuracy. Feature selection is the method of decreasing the number of independent variables when establishing an anticipating and accurate model.

There are some feature selection methods can be distinguished into three main types [29].

- Filter based: It defines some measure formed on that perception which is filtered. An instance of such a measured process could be the Pearson correlation, chi-squared.
- Wrapper-based: Wrapper systems accept the collection of a set of characteristics as an exploring problem, such as recursive feature elimination.
- Embedded: Embedded processes use techniques that have integrated attribute selection processes. Such as, Lasso and RF have their grant attribute-choosing methods.

We used filter-based method for feature selection in the research.

1) ***Pearson Correlation****:* It is a filter-formed system. It checks the total non-variable gain of the Pearson correlation between the selecting values and mathematical attributes in the dataset. We take the high numbers of attributes formed on these characteristics.

$$R = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}} \tag{1}$$

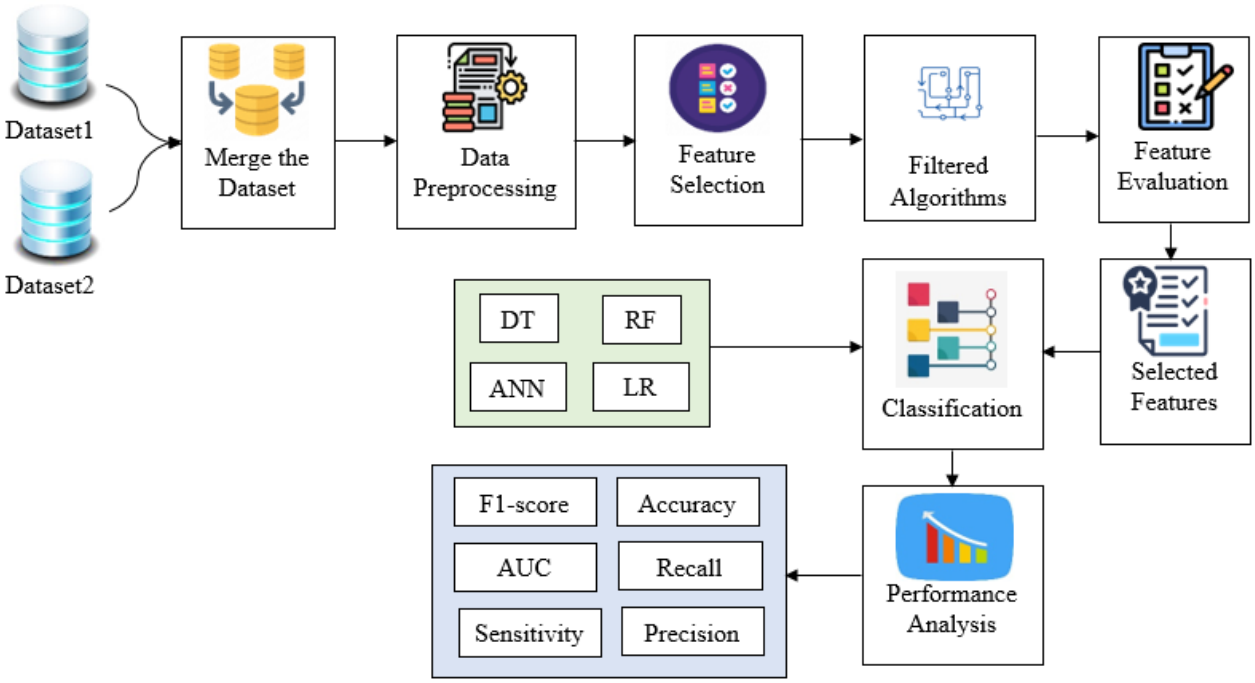

Fig. 1. Working Methodology

Where, X= Target Value for feature x, $\bar{X}$= Actual Value for feature x, Y= Target Value for feature y, $\bar{Y}$= Actual Value for feature y

2) ***Chi-squared****:* Chi-squared is also formed in the filtrate process. It checks in this process, we determine the chi-square measurement between the selecting values and the mathematical features attributes and it only selects the attributes with the greatest chi-squared values [30].

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{2}$$

Where, X= Chi-squared value, O= Observed value, E= Expected Value for the model. The chi-squared statistics also access in a hand-wavy method with positive numerical and classified features.

From the feature selection algorithm, we got 23-24 features from 26 features. And also evaluates the features.

From the feature selection algorithm, we got 20 features from 26 features. And also evaluates the features. And we prefer chi-squared over the Pearson correlation because chi-squared feature selection ability and performance are better than the Pearson correlation. The two algorithms gave almost the same features. So, we take further research using that modified dataset. From dataset 1 we have removed some features, because they did not give better performances.

C. Classification Methods

We have analyzed four data mining algorithms i.e. Decision Tree (DT), Random Forest (RF), Artificial Neural Networks (ANN) and Logistic regression (LR).

The C4.5 algorithm worn in Data Mining Techniques as a Decision Tree algorithm that can be professionalized to initiate a resolution, formed on a particular amount of experimental data.RF is consist of a great number of discrete decision trees that manage as a forest. Artificial Neural Networks (ANN) are the computerized model that is the subject confiscated by the human brain. Typically encumber represents the robustness of the inter correlations among neurons in the ANN. LR is a classifier-related data mining algorithm that is cast off to allocate monitoring to a linear set of features [19]. One of the examples of LR is online transactions Fraudster detection or not Fraudster. LR transforms its result casting off the logistic sigmoid function to come back with an anticipating value.

IV. Performance Evaluation

In this section, we have analyzed the performance of four data mining algorithms. Then the decision is making if the transaction is fraud or not.

A. Performance Metrics

Different kinds of performance metrics have been used to evaluate the performance of each classifiers.

- **Classification Accuracy:** CA calculates dataset accuracy that is the set of labeled predicted for a sample must exactly match the corresponding set of labels.

$$\text{CA} = \frac{\text{True Positive (TP)} + \text{True Negative (TN)}}{\text{Total number of observation(n)}} \times 100 \tag{3}$$

TABLE I
NEW DATASET BY USING THE CHI-SQUARED AND PEARSON CORRELATION ALGORITHM

Variable Name	Description
Cust_Id	The identity of the customer or user of the credit card.
Balance	The present balance of the credit card
Balance frequency	The average balance of the credit card normally
Purchases	The amount of the purchase from the credit card online transaction
Installments purchases	The installment amount of the daily purchases
Cash advance	The advance from the credit card
Purchase frequency	The frequency of the purchases from online transaction
Purchase installments frequency	The frequency of the purchase installments
Cash advance frequency	The frequency of the advance cash on online transaction of the credit card
Purchases TRX	Transfer amount of the purchase
Cash advance TRX	Transfer of the advance cash
Credit limit	The limit of the credit card
Payment	The payments of the products of the credit card
Minimum payments	The lowest amount of the online payments of using credit card
PRC full payments	The record of the full payments using credit card in online transaction
Is Foreign Transaction	Is it the transaction from another country or not. (Yes/NO)
Is High Risk Country	Is it the transaction from or to a high risk country or not. (Yes/No)
6_months_avg_chbk_amt	The amount of the average chargeback for the last 6 months
6_month_chbk_freq	The frequency of the 6 months chargeback amount
Is Fraudulent	Now the prediction of the online credit card fraudulent (Yes/No)

- **Sensitivity/Recall:** The Recall is the proportion of positives correctly categorized as given below:

$$Recall = \frac{\text{TP}}{\text{Total number of actual fraudulent}} \quad (4)$$

- **Precision:** The precision represents the division of examples categorized as positive that are actually fraudulent inspections. The best value of precision is 1 and the worst value is 0.

$$Precision = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (5)$$

- **F1-score:** The F1 score can be taken as an encumbered average of the precision and recall, where an F1 score outstretches its greatest value which is one and worst score which is zero.

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} \ + \text{recall})} \quad (6)$$

- **AUC:** Area under classification (AUC) is that, the system quality is established by the accuracy of the bulge to the upper right hand corner.
- **ROC:** The Receiver Operating Characteristic (ROC) is a diagram related curve of the algorithm that notifies the characteristics capability of a categorized system as its partisanship of the approach value is varied.

B. Performance Analysis

Firstly, We train and test the data set which is called data preprocessing using K-nearest neighbor (KNN). Fig. 2 plots the measurement of the fraud case of the credit card. Fig. 3 shows the training and testing accuracy of the dataset using KNN and it shows that the testing accuracy is getting better than the training accuracy.

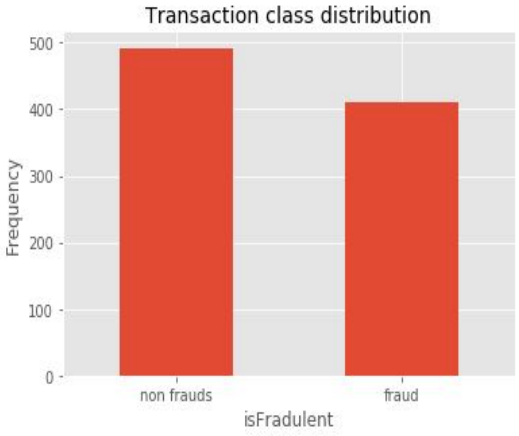

Fig. 2. Fraudulent versus frequency of transaction class distribution

In our research, in the beginning, we accessed a primary analysis from the website. After that, we would attempt to offer if there were any communications between the data description of the data set and how that type of cross-connections could be discussed. We would additionally attempt to disentangle and edit the dataset in order to gain a rather easy-to-acknowledge dataset.

TABLE II
PERFORMANCE OF DIFFERENT ALGORITHMS ON THE PROPOSED SYSTEM

Model	AUC	CA	F1	Precision	Recall	Accuracy (%)	Sensitivity (%)
Decision Tree (C4.5)	0.907	0.969	0.967	0.968	0.969	96.90%	96.90%
Random Forest	0.99	0.978	0.978	0.978	0.978	97.80%	97.80%
ANN (MLP)	**0.997**	0.986	0.986	0.986	0.986	98.60%	98.60%
Logistic Regression	**0.997**	**0.989**	**0.989**	**0.989**	**0.989**	**98.90%**	**98.90%**

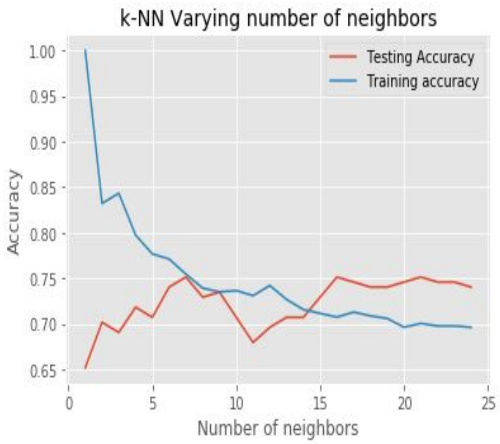

Fig. 3. Training versus testing accuracy using KNN

For the performance metrics we have used area under classification, classification accuracy, precision and recall. From the table II, we can say that the LR and ANN give better accuracy and sensitivity for the proposed system. Fig. 4 shows the ROC analysis for DT and LR and Fig. 5 shows the ROC analysis for RF and ANN. DT shows that the true positive rate or the sensitivity approaches the threshold value after getting some false positive rate. RF, ANN and LR show that the true positive rate or the sensitivity approaches the threshold value straight and which means the FP rate is quite low.

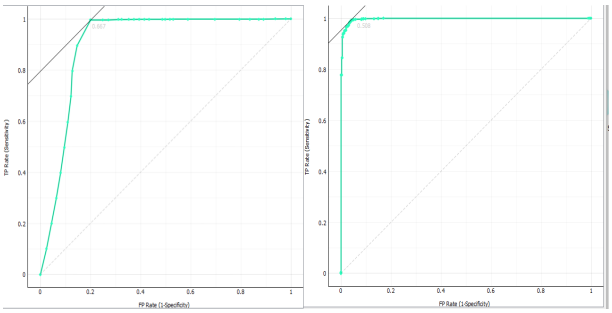

Fig. 4. ROC analysis for DT (C4.5) and LR

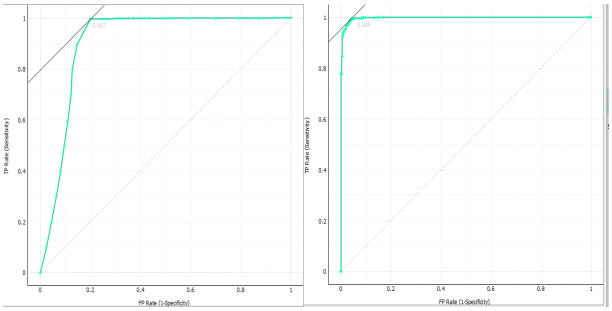

Fig. 5. ROC analysis for RF and ANN

Table III shows the confusion matrix of the of four algorithms. We can see that the actual and predicted result is quite accurate and it is 99.1%.

TABLE III
CONFUSION MATRIX OF THE FOUR CLASSIFIERS

Classifier	Actual	Predicted	
		Not Fraud(0)	Fraud(1)
DT	Not Fraud(0)	96.90%	3.20%
	Fraud(1)	3.10%	96.80%
RF	Not Fraud(0)	98.10%	5.00%
	Fraud(1)	1.90%	95.00%
ANN	Not Fraud(0)	98.90%	3.30%
	Fraud(1)	1.10%	96.70%
LR	Not Fraud(0)	99.10%	2.30%
	Fraud(1)	0.90%	97.70%

Fig. 6 shows the comparison between our proposed model and existing model.

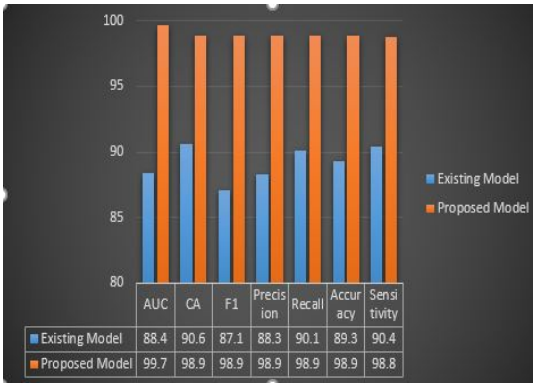

Fig. 6. Comparative study between existing and proposed model

C. Challenges

In this work, we have faced some challenges in data selection. There are many different types of data sets available

in the online platform, but we have to select the effective one which has more relevant information. Data filtering and feature selection techniques are the most effective way for this proposed model and reducing false alarm.

V. CONCLUSION

Fraudulent credit card is a difficult discussion for online commerce institutions along with normal peoples. In this research, we endeavor to find a stable infusion for reducing this problem. We contrast the presentation of four data mining techniques specified to credit card deceive perception and to classify their deficiency. We found that the LR, C4.5 decision tree algorithm, Random Forest and ANN are the greatest systems according to the presentation calculations (accuracy, recall, and precision). Our exploratory study proves that the perspectives normally cast off to solve false alarming problems may have displeasing conclusions when the customer's parameter is huge, such as bringing about a remarkable number of false positives. Since these procedures enhance the perspective, there are many deceptive cases that continue to go not finding. Many future demands are arranged that will be cared about in the next stage of this research concept. We will establish a system for the difficulties to discover a back-and-forth between recall and accuracy. We will establish a system that can also trace the fraudster and testify our system with a huge amount of data.

ACKNOWLEDGEMENT

This work was supported in part by the Center for Research, Innovation and Transformation (CRIT) of Green University of Bangladesh (GUB).

REFERENCES

[1] E. Duman and M. H. Ozcelik, "Detecting credit card fraud by genetic algorithm and scatter search," *Expert Systems with Applications*, vol. 38, no. 10, pp. 13 057–13 063, 2011.

[2] F. A. Sunny, M. I. Khan, M. S. Satu, and M. Z. Abedin, "Investigating external audit records to detect fraudulent firms employing various machine learning methods," in *Proceedings of the Seventh International Conference on Mathematics and Computing*. Singapore: Springer Singapore, 2022, pp. 511–523.

[3] A. Singh, R. K. Ranjan, and A. Tiwari, "Credit card fraud detection under extreme imbalanced data: a comparative study of data-level algorithms," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 34, no. 4, pp. 571–598, 2022.

[4] R. Van Belle, B. Baesens, and J. De Weerdt, "Catchm: A novel network-based credit card fraud detection method using node representation learning," *Decision Support Systems*, p. 113866, 2022.

[5] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5916–5923, 2013.

[6] A. Brabazon, J. Cahill, P. Keenan, and D. Walsh, "Identifying online credit card fraud using artificial immune systems," in *IEEE Congress on Evolutionary Computation*. IEEE, 2010, pp. 1–7.

[7] J. Forough and S. Momtazi, "Ensemble of deep sequential models for credit card fraud detection," *Applied Soft Computing*, vol. 99, p. 106883, 2021.

[8] D. J. Weston, D. J. Hand, N. M. Adams, C. Whitrow, and P. Juszczak, "Plastic card fraud detection using peer group analysis," *Advances in Data Analysis and Classification*, vol. 2, no. 1, pp. 45–62, 2008.

[9] P. Shimp and P. Kadroli, ""banking expert system"with credit card fraud detection using hmm algorithm," *International Journal Of Engineering And Computer Science*, 01 2016.

[10] 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, 2021.

[11] J. I.-Z. Chen and K.-L. Lai, "Deep convolution neural network model for credit-card fraud detection and alert," *Journal of Artificial Intelligence*, vol. 3, no. 02, pp. 101–112, 2021.

[12] J. Pun and Y. Lawryshyn, "Improving credit card fraud detection using a meta-classification strategy," *International Journal of Computer Applications*, vol. 56, no. 10, 2012.

[13] A. Srivastava, A. Kundu, S. Sural, and A. Majumdar, "Credit card fraud detection using hidden markov model," *IEEE Transactions on dependable and secure computing*, vol. 5, no. 1, pp. 37–48, 2008.

[14] A. Prakash and C. Chandrasekar, "A novel hidden markov model for credit card fraud detection," *International Journal of Computer Applications*, vol. 59, no. 3, pp. p35–41, 2012.

[15] L. Oghenekaro and C. Ugwu, "A novel machine learning approach to credit card fraud detection," *International Journal of Computer Applications*, vol. 140, no. 5, pp. 45–50, 2016.

[16] N. Wong, P. Ray, G. Stephens, and L. Lewis, "Artificial immune systems for the detection of credit card fraud: an architecture, prototype and preliminary results," *Information Systems Journal*, vol. 22, no. 1, pp. 53–76, 2012.

[17] D. Sánchez, M. Vila, L. Cerda, and J.-M. Serrano, "Association rules applied to credit card fraud detection," *Expert systems with applications*, vol. 36, no. 2, pp. 3630–3640, 2009.

[18] A. C. Bahnsen, A. Stojanovic, D. Aouada, and B. Ottersten, "Cost sensitive credit card fraud detection using bayes minimum risk," in *2013 12th international conference on machine learning and applications*, vol. 1. IEEE, 2013, pp. 333–338.

[19] V. N. Dornadula and S. Geetha, "Credit card fraud detection using machine learning algorithms," *Procedia computer science*, vol. 165, pp. 631–641, 2019.

[20] Y. Sahin and E. Duman, "Detecting credit card fraud by decision trees and support vector machines," in *World Congress on Engineering 2012. July 4-6, 2012. London, UK.*, vol. 2188. International Association of Engineers, 2010, pp. 442–447.

[21] T. Razooqi, P. Khurana, K. Raahemifar, and A. Abhari, "Credit card fraud detection using fuzzy logic and neural network," in *Proceedings of the 19th Communications & Networking Symposium*, 2016, pp. 1–5.

[22] H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah, and Q. Kang, "Optimizing weighted extreme learning machines for imbalanced classification and application to credit card fraud detection," *Neurocomputing*, vol. 407, pp. 50–62, 2020.

[23] M. Syeda, Y.-Q. Zhang, and Y. Pan, "Parallel granular neural networks for fast credit card fraud detection," in *2002 IEEE World Congress on Computational Intelligence. 2002 IEEE International Conference on Fuzzy Systems. FUZZ-IEEE'02. Proceedings (Cat. No. 02CH37291)*, vol. 1. IEEE, 2002, pp. 572–577.

[24] S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," *Decision support systems*, vol. 50, no. 3, pp. 602–613, 2011.

[25] K. Jegadeesan and S. Ayothi, "An empirical study of methods, metrics and evaluation of data mining techniques in credit card fraudulence detection," *Journal of Advanced Research in Dynamical and Control Systems*, vol. 12, 10 2020.

[26] N. Mahmoudi and E. Duman, "Detecting credit card fraud by modified fisher discriminant analysis," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2510–2516, 2015.

[27] A. S. Rathore, A. Kumar, D. Tomar, V. Goyal, K. Sarda, and D. Vij, "Credit card fraud detection using machine learning," in *2021 10th International Conference on System Modeling Advancement in Research Trends (SMART)*, 2021, pp. 167–171.

[28] S. Negi, S. K. Das, and R. Bodh, "Credit card fraud detection using deep and machine learning," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2022, pp. 455–461.

[29] M. Papík and L. Papíková, "Detecting accounting fraud in companies reporting under us gaap through data mining," *International Journal of Accounting Information Systems*, vol. 45, p. 100559, 2022.

[30] W. L. Al-Yaseen, A. K. Idrees, and F. H. Almasoudy, "Wrapper feature selection method based differential evolution and extreme learning machine for intrusion detection system," *Pattern Recognition*, vol. 132, p. 108912, 2022.