International Journal of Management, IT & Engineering
Vol. 14 Issue 8, August 2024
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Fraud Detection in Banking Using Machine Learning Techniques

Gaurav Anand

Abstract- The banking sector has undergone a generational transformation over the last few decades. Modern-day banks are significantly larger and more specialized in providing safe and seamless banking experiences to their customers. Machine learning (ML) and artificial intelligence (AI) techniques have been instrumental in accurately identifying and mitigating fraudulent activities within banking infrastructure. In this paper, we have leveraged multiple ML techniques to detect fraudulent activity in financial transaction data. Based on our findings, techniques like Gradient Boosting and Random Forest deliver better accuracy than legacy techniques such as Regression and Decision Trees. This research contributes to the broader goal of protecting financial transactions from fraud by providing valuable insights into the potential of AI and ML in enhancing fraud detection. The study also emphasizes the importance of model selection and hyperparameter tuning for improving the detection rate and minimizing the false positive rate. To enhance the robustness of fraud detection frameworks in future studies, we will consider incorporating real-time systems and exploring deep learning technologies

I. INTRODUCTION

The banking sector has always been central to the growth of the economies as it has provided facilities to both the consumer and the business world. However, the very fast pace of the digitization of financial activities has caused more challenges for businesses, particularly concerning security issues. Fraud operations in the banking space, such as identity theft, phishing, and other operations without the client's consent, have increased significantly, thereby creating real threats to financial security and customer confidence. Thus, there is an even greater need to establish strong and efficient mechanisms of fraud detection.

Until recently, the practice of controlling fraud by the banks primarily depended on the implementation of rules based approaches and the use of time-consuming, labor- intensive methods. While these methods are somewhat effective, they are increasingly less so in the face of growing and more diverse and rapidly changing fraudulent activities. Rule-based systems are based on a set of patterns and criteria, which are relatively simple to bypass by professionals in the field of fraud. This has demanded a rise in the usage of modern advanced techniques such as machine learning based algorithms. AI and ML based techniques have provided a pragmatic solution for effective financial fraud detection and mitigation. With the help of these technologies, it is possible to analyze big volumes of transaction data in real time and use the results, to identify trends and deviations from them, which can indicate fraud. Ability of ML based models to accurately learn from historical trends and quickly adapt to any change in pattern is very useful in an era where fraud trends are bound to change from one day to the other.

The research carried out in the last few decades have revealed how successful ML and AI techniques, are in detecting financial fraud. Among these models logistic regression, decision trees, random forests, and neural networks have been used to assess data of transactions featuring various benefits. For instance, the decision tree and the random forest are more flexible when modeling and can accommodate interactions between the data while being more qualified than the logistic regression providing a very simple but straightforward model to the client. Of all the techniques, deep learning models that are computed using neural networks adapt better to pick out hard-to-detect faults in an organization's books since they reveal hidden information from large data sets.

While the findings regarding the use of ML and AI to address the problem of fraud are encouraging, there are some challenges associated with the application of these approaches. First challenge is the inequality in fraud detection datasets where the number of genuine transactions significantly outweighs the fraudulent ones. Unless addressed, models can start developing certain biases that would lead to a predisposition of results which would lean more towards non-fraudulent activities. Methods like oversampling, sampling, and the use of niche algorithms like SMOTE can be used to overcome this problem and guarantee more accurate results when it comes to identifying fraud.

The other issue is transparency of decisioning in ML and AI models. Most of ensemble and network based algorithms tend to work as a "Black box", and hence does not allow end user to figure out why some decisions were made. While in some industries this might be an issue, it does pose a potential roadblock in banking industry, given regulations around fair lending and banking.

Recent developments in the creation of interpretative models and techniques like SHAP and LIME can give the understanding of decision-making by complex models.

This paper focuses on the application of ML methodologies in identifying fraudulent transactions within the context of banking systems. Looking at the results of the accuracy, precision, recall, and F1 score levels of the logistic regression, decision tree, random forest, gradient boosting, and neural network models with the data set, have been compared. The paper outlines the need and the implication of these technologies in the fight against financial crimes while pointing out the need to consider and address the issues that come with the implementation of these technologies.

II LITERATURE REVIEW

The approaches based on machine learning (ML) and artificial intelligence (AI) for fraud detection have emerged today as a critical research area, mainly due to the necessity to cope with constantly evolving fraud schemes in the banking context. This literature review is aimed at presenting a historical analysis of anti-fraud measures, and an overview of the present ML and AI schemes in the framework of fraud detection, alongside the discussed challenges and changes.

A. Evolution of Fraud Detection Methods

The aspect of performing a manual analysis of all flagged transactions is tremendously time-consuming, and error-prone, and thus is not preferred by financial organizations. In the past,

fraud detection especially in the field of banking has been done through a traditional method and through rule-based systems [1]. These are a conventional approach that uses well-defined guidelines and limits such as reporting any transaction greater than a fixed amount or any account activity that deviates from the normal. Although they are useful to a certain degree, such systems are not very efficient because they cannot be programmed to adapt to emerging fraud schemes. Bolton and Hand in their paper, explained that a rule-based system is too simplistic and easy for fraudster to bypass.

B. Emergence of Machine Learning and Artificial Intelligence

Due to the limitations of conventional fraud detection approaches, industries have started adopting novel techniques such as ML and AI: These technologies cover the possibility to work with a large amount of data, note the patterns, and adjust to the new fraud trend [2]. ML and AI models as pointed out, can acquire newer knowledge from past data, and their performances get better with time. This capability is important in a context that is dynamic and the fraud techniques are bound to change at some point.

In fraud detection, several modeling techniques can be used; each with its own advantages as well as disadvantages [3]. Logistic regression is one of the most basic and easy-to-interpret types of models that has seen application in binary classification problems, including fraud detection. It has its drawback in that it is rigid in its structure and may not be able to accommodate all the relationships embedded in the data set. Meanwhile, decision trees and random forests have more possibility and may include nonlinear terms between the variables. Random forests are regarded as very stable and accurate for the same fact that, rather than building a single decision tree to solve a problem, composite decision trees are developed that considerably minimize overfitting.

Artificial neural networks, particularly deep learning algorithms, have proved to give very satisfactory results in finding hidden and complex relationships in large datasets [4]. Feature learning claims that deep learning models can learn the right features for a task without prior guidance. Such characteristics make them ideal for use in fraud detection because fraudulent activities may be buried in a sea of genuine activities.

C. Current Applications and Techniques

Various papers have proven that usage of ML and AI would pave the future in fraud detection [5]. The recurrent neural networks (RNNs) can identify credit card fraud with relatively higher accuracy as well as recall value. The technique uses the fact that the transaction data is ordered, which most other techniques cannot consider.

Other techniques also include ensemble methods in which more than one model is generated to work simultaneously to tackle the issue and reduce the rate of fraud [6]. Random forests, gradient boosting, and logistic regression were used to propose an ensemble ML strategy for recognizing fraudulent transactions. Their approach demonstrated how performance by aggregating different models is better than the stand-alone models‟ performance, and the importance of model diversification.

Also, the incorporation of unsupervised learning has been applied in the prevention of fraud performance [7]. Other techniques, which are in the unsupervised-learning category including clustering and anomaly detection can also be used to determine extraordinary patterns in the data without constructing labels from training samples. Researchers have came up with the Isolation Forest algorithm where anomalies are isolated through consecutive severances of the data [8]. This has helped in the analysis of transaction data to identify the outlier using an approach that enhances supervised learning

D. Challenges in Fraud Detection

There are still some issues with ML and AI based techniques in modern banking context [9]. The first challenge is that almost all fraud detection datasets are imbalanced. The problem of fraud detection is usually characterized by class imbalance since in most of the financial datasets, number of fraudulent transactions are relatively small compared to the legitimate ones. This imbalance can cause the model to favor predicting non- fraudulent activity in its classification. To handle this problem, oversampling and sampling and other forms of data generation like SMOTE have been used.

Another issue is related to the explainability of ML and AI models [10]. The most accurate ones for the present range of applications are neural networks and methods of ensemble learning, but the training mechanisms of these models contain certain components that work as a "black box," which makes it challenging to understand how exactly the model came up with a particular decision. This can be quite an issue in banking where regulatory authorities expect the institution to be able to explain various analytic findings, especially in detecting fraud. Work is ongoing to come up with some new and more understandable models and methods thus; SHAP (Shapley Additive exPlanation) as well as LIME (Local Interpretable Model-agnostic ExPlanations) which gives data about how the complicated models arrive at the decisions.

Other common considerations in the use of ML and AI in the provision of fraud identification are data confidentiality and protection [11]. It is expected from the financial institutions that they will properly manage their customer's data and would ensure its not leaked in firm data breach etc. Companies must adhere to data protection laws that are currently in existence like the GDPR of Europe or the CCPA of the United States, for customers' trust as well as to avoid legal consequences. Such issues as privacy are addressed by techniques like federated learning which enables models to be trained on the decentralized data without necessarily sharing some of the sensitive information.

E. Future Directions and Advancements

The future of fraud detection in banking is closely connected with such novelties as ML and AI [12]. One of the current areas of focus is the application of deep learning models, including RNNs and CNNs, which show significantly high accuracy in the analysis of patterns in transactions. Also, transfer learning which trains models on large databases before fine-tuning them on some tasks could improve fraud detection models given that knowledge from other related domains can be utilized.

Another promising technique for fraud detection is reinforcement learning (RL)[13]. RL models can choose strategies that can be used to detect fraud since the model itself interacts

with the environment to get feedback in terms of reward or penalty. This technique could be capable of evolving into one stop shop for fraud detection in banking industry.

In addition, the application of ML and AI with blockchain technology is a new way to prevent fraud [14]. The peer-to-peer registration characteristic of the blockchain paradigm also effectively creates a more secure environment for financial transactions since it is almost hard for the fraudster to alter data. The integration of blockchain with ML and AI models can offer a reliable environment for real- time detection and prevention of fraud.

III.METHODOLOGY

The methodology for this study on fraud detection in banking using machine learning (ML) and artificial intelligence (AI) involves several key steps: acquisition of data, data preprocessing, feature selection, model selection and building, performance assessment and comparison.

A. Data Collection

The dataset applied in this study is obtained from a large financial transactions dataset from Kaggle Website. It comprises parameters like transaction amount, originating/destination accounts and, before/after balance figures. Dataset also contains binary indicator, to determine if a transaction is fraud or not.

B. Data Preprocessing

Data preprocessing includes multiple steps such as missing value treatment, conversion of categorical variables into numerical via techniques like one hot encoding, data standardization and normalization etc. These basic steps ensure that the data is ready to be used for subsequent steps. Once data is processed, next step is to ensure data is not biased. If the data is biased techniques like synthetic generation of samples such as the Synthetic Minority Over-sampling Technique (SMOTE) helps to deal with the class imbalance

C. Model Selection and Training

Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and Neural Networks are the machine learning models applied in this research with the aim of determining the best method of fraud detection [20]. Again, every model has different features which include simple or complex, flexible, or not, and accurate and the models may be highly complex or less complex to articulate [16]. Every model is trained with pre-processed data. Grid search was used for hyper parameter tuning.

C. Model Evaluation and Comparison

The models are then tested and their performances are compared with the help of a test dataset that is not used in the training phase. Accuracy, precision, recall, and F1 score were used for measuring model performance

In addition to evaluating each model's accuracy, confusion matrices are used to display where on the distribution of true positives, true negatives, false positives, and false negatives they were correctly identified. Furthermore, the models are compared regarding how well they conform to the high-level objective of accurate abnormal event detection while balancing false positive rate.

IV.RESULTS

Each model was trained and tested rigorously with a view of establishing the model's capacity to detect fraud while reducing both the false positive and the false negative rates as shown in Figure 1.

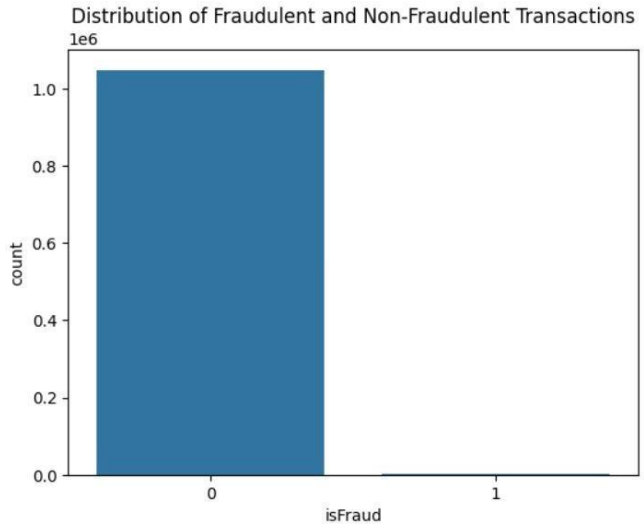


Fig 1: Distribution of Fraudulent and Non-Fraudulent Transactions

Based on model results we interpreted that models such as regression based linear models are most simplistic and are easiest to train. We saw a promising 99.91% of accuracy. However, given sensitive nature of fraud use case, this performance is not at par. Lack in performance was primarily driven by model's linear conception; it was impossible to capture the intricate correlation that may be tied to fraud activities. Therefore, the percentage of correct identification of fraudulent transactions among all identified frauds was high, but it had a rather low recall, which is 27.5%, meaning it failed to identify all actual fraudulent transactions. These limitations are also reflected in the F1 score that considered the balance between precision and recall and was 42.4%.

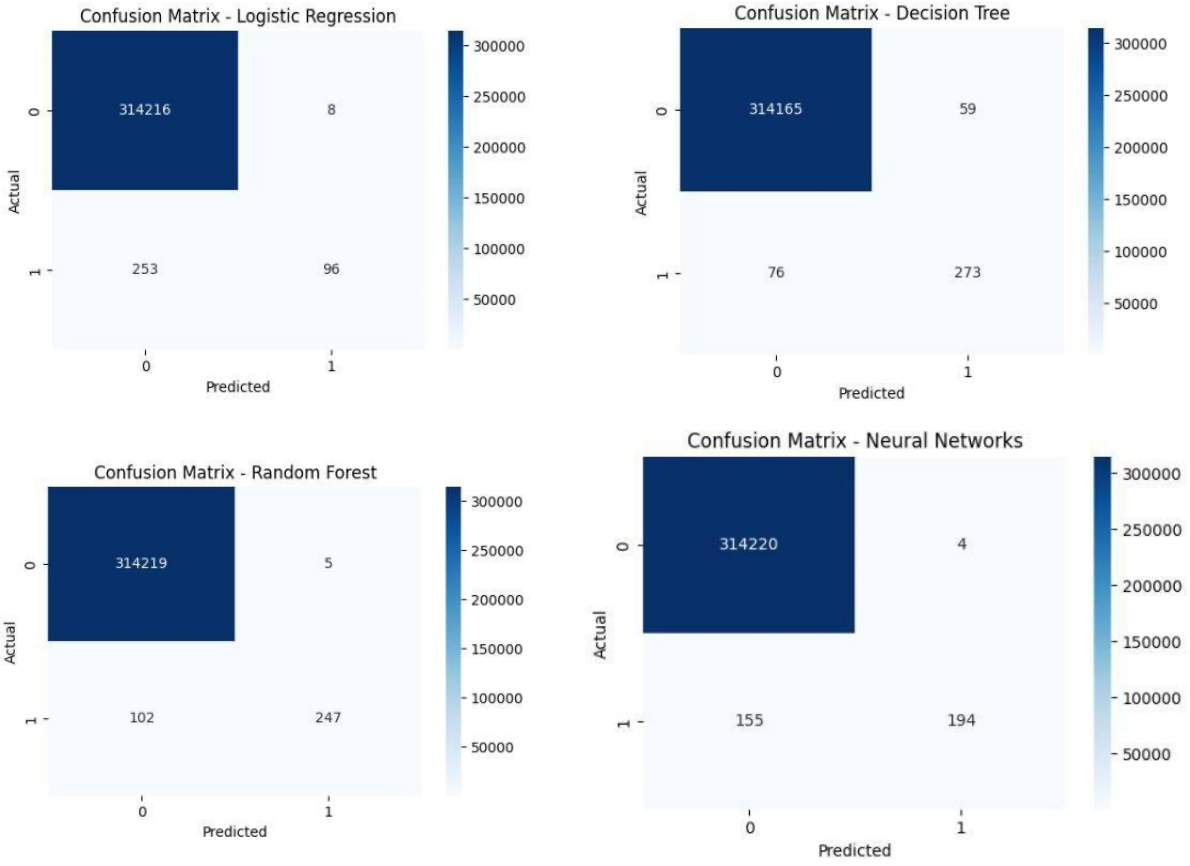
The decision tree model gave more insights into the data than the linear model. Since partitioning the dataset into various branches based on the feature values was possible, the decision tree model could handle nonlinearity. The decision tree model offered a better view of the data in comparison to the initial analysis. Model had 99.95% accuracy, precision of 81.8% and a recall of 78.5% leading to an F1 score of 80.1%.

Another relevant method, Random Forest, which combines decision trees outcomes, demonstrated significant advancements in performance [9]. Random forest had an accuracy of 99.96%. The precision of the model was 98.0% and the recall of the model was 71.4% resulting in the F1 score of 82.6%. With its ensemble approach, the overfitting factor was controlled, and the overall accuracy improved making it a preferred practical solution to fraud detection. Therefore, the precision and recall data provide a good sign since the model was able to categorize a large component of the approximately 34,000 fraudulent transactions while producing minimal false positives.

Another ensemble technique is gradient boosting, and it also yielded satisfactory results for accuracy and recall with accuracy ranging to 99.96%, with recall of 68.8% and F1 score of 78.3%. This process of progressive improvement made it especially useful in capturing the intricate relations related to fraudulent activities with the use of gradient boosting.

The neural networks showed good results with an accuracy of 99.95%, with precision of 97.8% and a recall of 51.9% which indicated an F1 score of 67.8%. In their structure and usability, Neural Networks can be described as black boxes for all their elaborate parameters and extensive data processing capability, the former can be used straightforwardly for fraud detection due to its ability to discern complex patterns.

As a next step, the confusion matrices were used to compare the results of model performances, and key performance indicators were used [1]. The concept of confusion matrices gave a clear distinction in Figure 2 in such a way that each of the models was evaluated appropriately. Moreover, the use of overall accuracy, precision, recall, and F1 score provided qualitative results in addition to defending the effectiveness of the presented models.



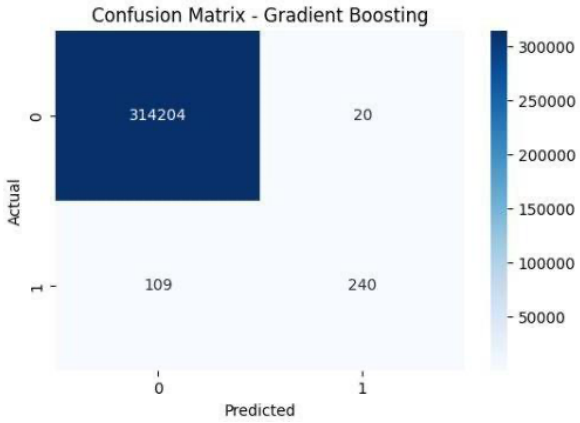


Fig. 2: Confusion Matrices of Different Models

The comparison of different approaches in Figure 3 demonstrates that the ensemble of the algorithms, as random forest, and gradient boosting, are more effective in comparison with the more conventional models as logistic regression and decision trees. Neural networks also succeeded especially in the case of dealing with great and challenging data sets.

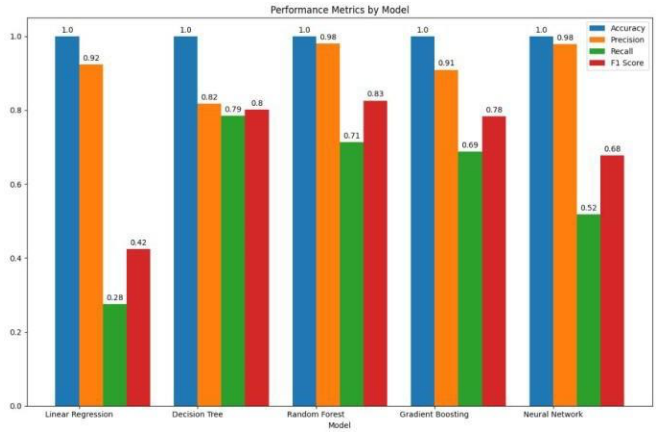


Fig. 3: Performance Metrics of Different Models

v. DISCUSSION

The paper focused on a comparative analysis of the machine learning models applied to fraud detection for banks, where the advantages and drawbacks were identified [18]. It was found in Table 1, that newer techniques such as random forest and gradient boosting outperformed the conventional models such as logistic regression and decision tree. Neural network samples gave high- performance levels with an accuracy of 99.95%, they can be effectively implemented in real-time fraud detection. However, the structure of neural networks is intricate and that results in issues related to the interpretability of the results. This can result in over-separation and risky models, especially for minority class; such an issue can be solved using SMOTE [19]. However, this has always been a challenge in trying to aim at ensuring that, the values of sensitivity and specificity are balanced. The

deployment of machine learning along with artificial intelligence in the detection of frauds and scams needs to be added to current structures and should be monitored.

Model	Accuracy	Precision	Recall	F1 Score
Linear Regression	99.91%	92.30%	27.50%	42.38%
Decision Tree	99.95%	81.79%	78.51%	80.12%
Random Forest	99.96%	98.03%	71.35%	82.59%
Gradient Boosting	99.96%	90.91%	68.77%	78.30%
Neural Network	99.95%	97.83%	51.86%	67.79%

VI.CONCLUSION

The study then narrows down to the way ML and AI could be useful in the banking industry to identify fraud. Like what has been discussed previously, gradient boosting and random forests also benefit from the nonpareil accuracy and precision of ensemble methods. Still yet, when there is an imbalance of the data or difficulties in terms of analyzing the data, then again, a way could be met. As for the former, methods such as SMOTE can be used, whereas, for the latter, interpretability tools can be employed. Based on future trends, it showed that blockchain would be adopted together with ML and AI.

REFERENCES

[1] Abdolrasol, M. G., Hussain, S. S., Ustun, T. S., Sarker, M. R., Hannan, M. A., Mohamed, R., ... & Milad, A. (2021). Artificial neural networks-based optimization techniques: A review. *Electronics*, *10*(21), 2689.

[2] Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., ... & Saif, A. (2022). Financial fraud detection based on machine learning: a systematic literature review. *Applied Sciences*, *12*(19), 9637.

[3] Ashtiani, M. N., & Raahemi, B. (2021). Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review. *Ieee Access*, *10*, 72504-72525.

[4] Bello, H. O., Idemudia, C., & Iyelolu, T. V. (2024). Integrating Machine Learning and Blockchain: conceptual frameworks for real-time fraud detection and prevention. *World Journal of Advanced Research and Reviews*, *23*(1), 056-068.

[5] Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Artificial intelligence in organizations: Current state and future opportunities. *MIS Quarterly Executive*, *19*(4).

[6] Chen, J., Huang, H., Cohn, A. G., Zhang, D., & Zhou, M. (2022). Machine learning-based classification of rock discontinuity trace: SMOTE oversampling integrated with GBT ensemble learning. *International Journal of Mining Science and Technology*, *32*(2), 309-322.

[7] Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable Artificial Intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *11*(1), e1391.

[8] Găbudeanu, L., Brici, I., Mare, C., Mihai, I. C., & Șcheau, M. C. (2021). Privacy

intrusiveness in financial- banking fraud detection. *Risks*, *9*(6), 104.

[9] Heigl, M., Anand, K. A., Urmann, A., Fiala, D., Schramm, M., & Hable, R. (2021). On the improvement of the isolation forest algorithm for outlier detection with streaming data. *Electronics*, *10*(13), 1534.

[10] Hilal, W., Gadsden, S. A., & Yawney, J. (2022). Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems with applications*, *193*, 116429.

[11] Iwendi, C., Khan, S., Anajemba, J. H., Mittal, M., Alenezi, M., & Alazab, M. (2020). The use of ensemble models for multiple class and binary class classification for improving intrusion detection systems. *Sensors*, *20*(9), 2559.

[12] Karthik, V. S. S., Mishra, A., & Reddy, U. S. (2022). Credit card fraud detection by modeling behavior patterns using hybrid ensemble model. *Arabian Journal for Science and Engineering*, *47*(2), 1987-1997.

[13] Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, *3*(1), 91-99.

[14] Mytnyk, B., Tkachyk, O., Shakhovska, N., Fedushko, S., & Syerov, Y. (2023). Application of artificial intelligence for fraudulent banking operations recognition. *Big Data and Cognitive Computing*, *7*(2), 93.

[15] Razaque, A., Frej, M. B. H., Bektemyssova, G., Amsaad, F., Almiani, M., Alotaibi, A., ... & Alshammari, M. (2022). Credit card-not-present fraud detection and prevention using big data analytics algorithms. *Applied Sciences*, *13*(1), 57.

[16] Saarela, M., & Jauhiainen, S. (2021). Comparison of feature importance measures as explanations for classification models. *SN Applied Sciences*, *3*(2), 272.

[17] Sánchez-Aguayo, M., Urquiza-Aguiar, L., & Estrada- Jiménez, J. (2022). Predictive fraud analysis applying the fraud triangle theory through data mining techniques. *Applied Sciences*, *12*(7), 3382.

[18] Singh, V., Chen, S. S., Singhania, M., Nanavati, B., & Gupta, A. (2022). How are reinforcement learning and deep learning algorithms used for big data-based decision making in financial industries–A review and research agenda. *International Journal of Information Management Data Insights*, *2*(2), 100094.

[19] Taha, A. A., & Malebary, S. J. (2020). An intelligentapproach to credit card fraud detection using anoptimized light gradient boosting machine. *IEEE access*, *8*, 25579-25587.

[20] Tatineni, S. (2020). Enhancing Fraud Detection in Financial Transactions using Machine Learning and Blockchain. *International Journal of Information Technology and Management Information Systems (IJITMIS)*, *11*(1), 8-15.