WJCMS

Wasit Journal of Computer and Mathematics Science

Journal Homepage: https://wjcm.uowasit.edu.iq/index.php/WJCM

e-ISSN: 2788-5879 p-ISSN: 2788-5879

Credit Card Fraud Identification using Logistic Regression and Random Forest

Wang Yundong[1,*], Alexander Zhulev[2] and Omar G. Ahmed[3]

[1]Institute of Media, Social Sciences and Humanities, South Ural State University, 454080 Chelyabinsk, Russia
[2]Department of System Programming, South Ural State University, 454080 Chelyabinsk, Russia
[3]Department of Electric Drive, Mechatronics and Electromechanics, South Ural State University, Chelyabinsk, 454080, Russia

*Corresponding Author: Wang Yundong

DOI: https://doi.org/10.31185/wjcm.184
Received: July 2023; Accepted: Septemper 2023; Available online: Septemper 2023

ABSTRACT: Fraud is an ancient yet ever-changing profession. Because of the digitization of money, financial transactions, banks, fraudsters now have a limitless number of possibilities to perpetrate crime from behind a screen, anywhere around the world. Fraud has a broad influence, with direct ramifications for business and the economy. It is of great worry to cybercrime organizations as recent studies have proven that ML algorithms may successfully be utilized to identify fraudulent transactions in massive amounts of payment data. Such techniques may identify fraudulent transactions in real time, which human auditors may miss. In this research, we apply supervised ML algorithms to the issue of fraud identification by analyzing simulated financial transaction data that is available to the public. Our aim is to show how supervised ML methods may be utilized to successfully identify data with extreme class disproportion. By way of example, we show how exploratory analysis may be utilized to identify fraudulent from real purchases. We also show that Random Forest outperform Logistic Regression when applied to a clearly distinguished dataset.

Keywords: Fraud, ML, Random Forest, Transaction, Logistic Regression

1. INTRODUCTION

Fraud is an increasingly prevalent occurrence. Computer network fraud is a common topic in the news, and in some situations, it takes some time before the problem is discovered. Fraud has a significant effect that directly affects the economy and company. Therefore, fraud detection methods are crucial. Effective tools to help corporate operations are provided by statistical and data mining techniques. One of them is fraud detection, and there are programs that look for instances like credit card fraud or hacking of computer systems. The main cause of financial losses is fraud, which is more prevalent than ever, particularly in the modern Internet era [1]. Transaction fraud cost the economy over $30 billion in 2020, $32 billion in 2021, and over $36 billion in 2023 [2]. Cyber security and cybercrime teams work hard to stop online financial scams [3]. This is an important part of their jobs. Most financial companies and banks have teams of researchers whose sole job is to create automatic systems that look at transactions made with their products and find ones that might be fake. In order to be better equipped to handle cybercrime situations, it is crucial to investigate the method for resolving the issue of spotting fraudulent entries/transactions in vast volumes of data. Businesses and organizations in a variety of sectors are very concerned about financial fraud. Traditional rule-based systems and manual investigations often fall short of keeping up with fraudsters' increasingly sophisticated strategies. As a consequence, a lot of businesses are using ML algorithms to identify financial fraud [4].

ML employs advanced algorithms to evaluate massive quantities of data, identify trends, and detect anomalies that

may indicate fraud. By automating the process, ML models can significantly enhance the efficiency and accuracy of fraud detection, enabling organizations to identify and respond to fraudulent behavior in real-time. The aim of this research is to explore differ-entML techniques that can be applied to financial fraud detection. By utilizing historical transactional data, these techniques aim to identify fraudulent patterns, detect anomalies, and predict potential fraudulent activities. This study will discuss several ML approaches commonly employed in financial fraud detection, including supervised learning, unsupervised learning, anomaly detection, neural networks, ensemble methods, and feature engineering [5]. Each technique will be examined in terms of its applicability, advantages, and potential limitations. Furthermore, this research will emphasize the importance of continuous model updates and retraining to ensure that fraud detection systems remain effective against evolving fraud tactics. Additionally, it will discuss the need for a comprehensive approach that combines ML with other fraud detection techniques, such as rule-based systems and human expertise, to achieve optimal results [6].

The three primary goals of this research are as follows:

1) To investigate financial fraud detection literature to comprehend the many facets of the issue.

2) To use supervised ML methods to find financial fraud using a sample dataset that is publicly accessible.

3) To compare multiple approaches to determine which is most suitable for this application.

The reminder of the article is organized as below. Brief literature study is described in Section 2. Section 3 describes the framework. The method and result analysis are presented in Section 4. Section 5 discusses the conclusion and future activities.

2. LITERATURE REVIEW

Decision trees, Least Squares Regression, Logistic Regression (LR), and SVMare used to identify suspicious activities in live data sets. Two techniques under random forests [7] are used to teach the behavioral features of regular and abnormal transactions. Both CART and random trees are used in these random forests. There are still problems with unbalanced data, despite the fact that random forest obtains outstanding performance with small data sets. The focus of future work will be on resolving the aforementioned problem. The random forest algorithm might need some tweaks. KNN, Nave Bayes, and LR are tested on highly unbalanced credit card fraud data, and meta-classifiers and meta-learning strategies for handling such data are investigated [7].In certain situations, supervised learning techniques for fraud detection may not be successful. To separate out outliers from regular behavior, we build a model using a deep Auto-encoder and a constrained Boltzmann machine [8]. Credit card fraud may be detected via a hybrid technique suggested by Olena et al. [9], which uses random forest and isolation forest to identify anomalous transactions. The suggested paradigm of the author rests on two main components. One of them raises doubts about unsupervised learning-enabled anomaly detection. The second theory considers the unusual occurrences category. Supervised learning is used. The speed of data that works well with the hybrid model on real-time data [10] is the key issue of the proposed research. The technology was tested to see whether it could detect the location of customers making purchases and utilize that information for detection purposes. The magnitude of abnormality is not the basis for this hybrid model. On the other hand, it depends on what kind of oddity it is. The geolocation-based transaction anomaly detection technology is used to uncover fraud. However, privacy and secrecy are not sufficiently secured because of the usage of real-time data to identify fraudulent transactions. The author divided the detection of fraudulent activity into three stages: user verification; a fuzzy clustering approach; and an ANNclassification phase. The approach helped achieve a 93.90 percent success rate and a 6.10 percent error rate in classification [11]. Credit card fraud risk methods for higher-dimensional data have been suggested by Rtayli et al. [12], who offer a combination of random forest (RF) classifier and SVMclassification approaches. The idea came from seeing how fraudsters pick out their features in an unbalanced big dataset. Since fraudulent transactions are rare, finding them might be challenging. Recall, Accuracy, and AUC are only few of the assessment criteria the author has utilized to assess the model's performance.

A ML-based method for detecting credit card fraud has been recommended by [13], which makes use of hybrid method with majority voting and Ada Boost techniques. To help the approach along, they included random noise of 10%-30% into their hybrid models. Based on the 30% noisier sample data, multiple voting methods were given a score of 0.942. Therefore, they concluded that the voting system was the most efficient approach in the existence of disruption. To learn the differences between typical and abnormal buying and selling habits, researchers turned to the RF presented in [14]. In this analysis, we compare the effectiveness of the classifiers in these two RF in detecting credit card fraud. This study looks at how well these two models based on RF can spot credit card scams. These two RF models were evaluated using information from a Chinese e-commerce firm. Although the proposed RF may perform admirably on modest samples, issues like as imbalanced data prevent them from being at the same level of excellence as larger samples [15].

3. PROPOSED FRAMEWORK

This article used the standard ML method. The tagged class variable in the discovered dataset was utilized as the prediction variable in ML methods.

1. Using exploratory analysis, we thoroughly evaluated the data set and found potential fraud predictors.
2. We observed the separation of fraud and non-fraud transactions using different visualization approaches.
3. We tested with two supervised ML algorithms to tackle the fraud detection problem.
4. We also attempted under-sampling to solve the dataset's class imbalance.
5. The models were built using cross-validation to minimize overfitting and provide consistent performance.
6. AUC and Confusion Matrix were utilized to compare the performance of the various models.

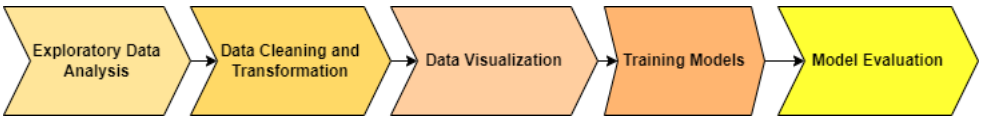

FIGURE 1. Recommended framework workflow

This study was carried out in Python utilizing a Jupyter notebook. The ML algorithms were run utilizing built-in libraries and methods. Functions were developed when required to ease certain analysis or visualizations. The Fig. 1 figure depicts in detail the whole procedure described in the study. The work was completed completely in Python, with the analysis recorded in a Jupyter notebook. Several analyses were carried out using standard Python libraries (sklearn, pandas, and seaborn). These libraries are discussed more below.

3.1 DESCRIPTION OF DATASET

This investigation makes use of a dataset of synthetically created digital transactions generated using an emulator called PaySim [16]. It replicates mobile money transactions based on a sample of genuine transactions collected from one month of financial logs from an African country's mobile money service. It generates a synthetic dataset by aggregating anonymized data from the private dataset and then injecting fraudulent transactions. The dataset contains nearly 6 million transactions as well as 11 variables. There is a variable called 'isFraud' that represents the transaction's real fraud status. This is the class variable for our investigation. The number 1 implies fraud, whereas the value 0 shows non-fraud.

4. METHODS AND RESULT ANALYSIS

4.1 CLASSIFICATION MODELS FOR FRAUD DETECTION

Recall is a valuable indicator for measuring model performance. High-class imbalance datasets generally result in low recall, despite great accuracy. Accuracy will also be considered, since lower accuracy suggests that the organization attempting to identify fraud would pay more costs in screening the transactions. Alternatively, we might use the ROC and AUC. This will not be an acceptable measure if the model correctly detects the bulk of fraudulent transactions. As a result, we utilize this to validate the model's performance. Cross-validation is also needed to make sure that the models don't make too many assumptions about the training data. We do this with stratified 5-fold because we need to make sure that the class mismatch is still there in the validation sets.

4.2 RANDOM FOREST METHOD

Random Forest (RF) is a popular ML algorithm that belongs to the family of ensemble methods. It combines the predictions of multiple decision trees to make accurate and robust predictions. The algorithm was introduced by Leo and Adele in 2001 and has since become widely used in various domains. RF have gained popularity due to their ability to handle complex problems, provide insights into feature importance, and achieve high accuracy with robustness against overfitting. In this part, we useRF model, and compute the mean recall score.

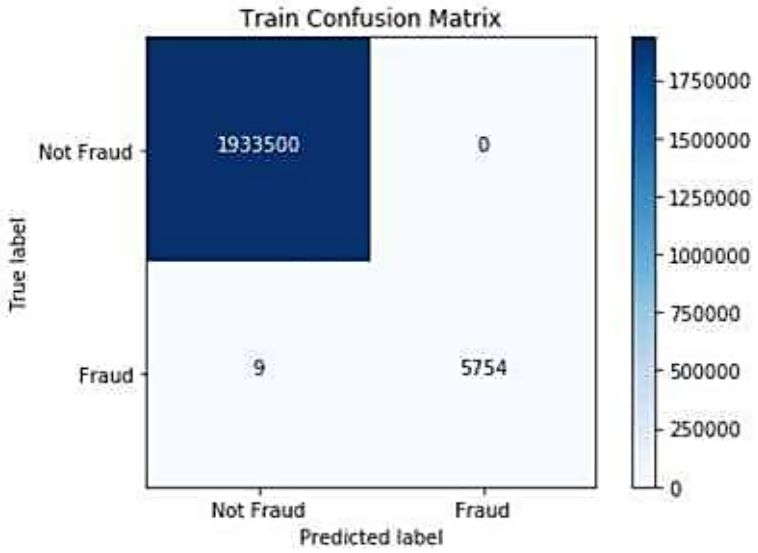

FIGURE 2. **Train Confusion Matrix**

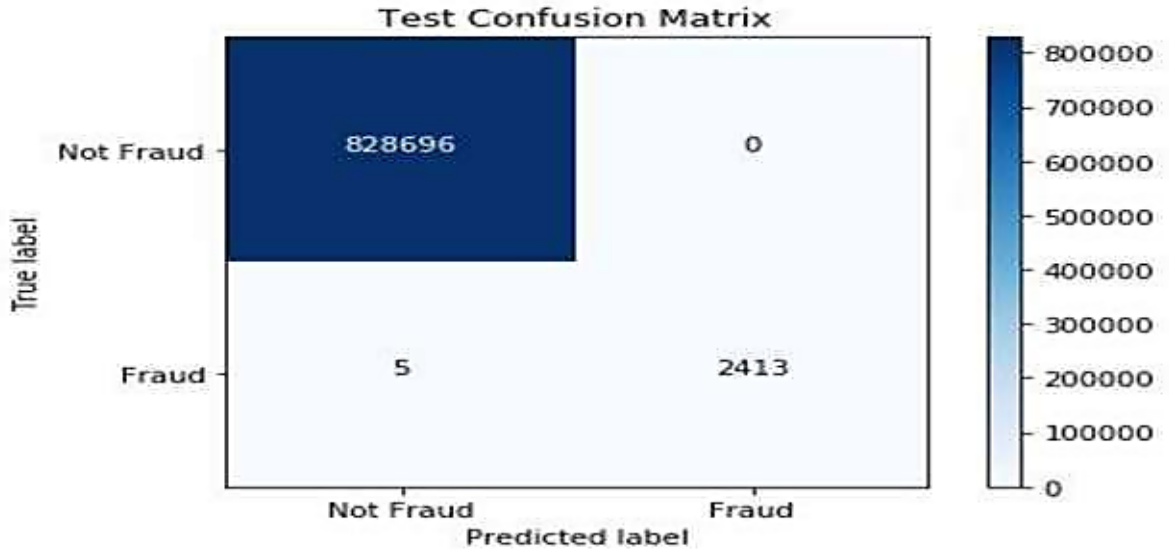

FIGURE 3. **Test Confusion Matrix**

4.3 LOGISTIC REGRESSION METHOD

LR is a widely used statistical and ML algorithm for binary classification tasks. It is a linear model that predicts the probability of an event occurring based on input features. LR was developed by statistician David Cox in the 1950s and has since been extensively applied in various fields. It is a widely used algorithm for binary classification tasks. It provides interpretable results, can handle large datasets efficiently, and serves as a good baseline model for many classification problems. Here, we train the LR model and compute the mean recall score. We draw the confusion matrices for the logistic regression model's train and test datasets and assess the recall and precision in each scenario.

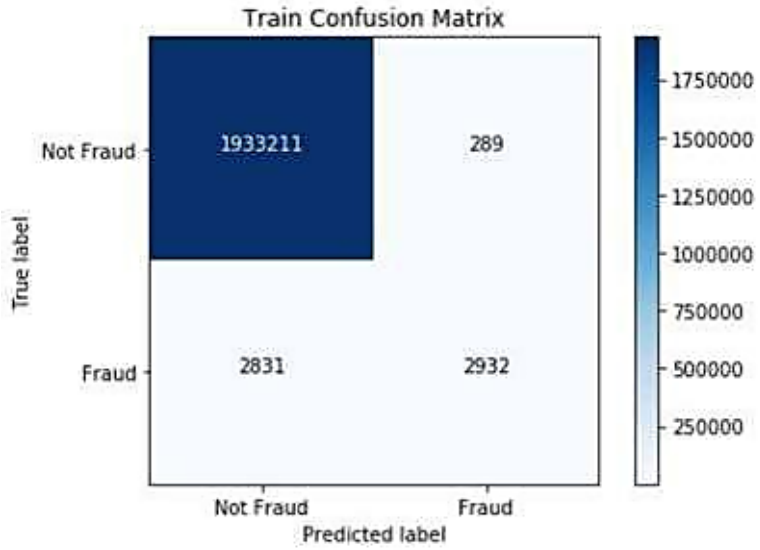

FIGURE 4. Train Confusion Matrix

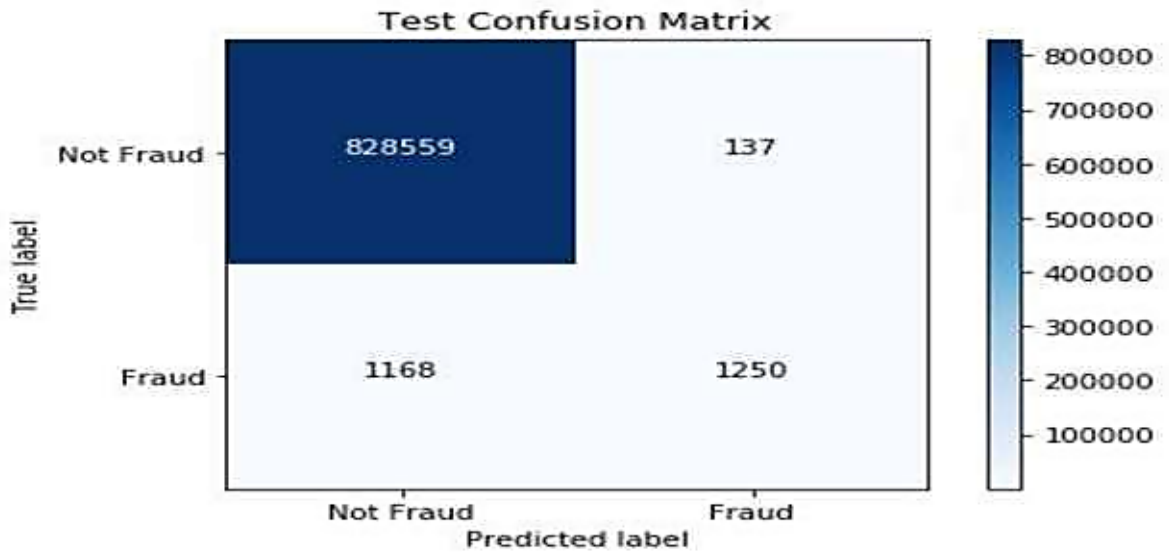

FIGURE 5. - Test Confusion Matrix

There are two outcomes from the preceding results:

1. There is no overfitting and the testing and training datasets are consistent
2. High accuracy and poor recall suggest that applying the method on data with a high degree of class imbalance will not provide great results.

The RF algorithm produces near-perfect outcomes. When memory scores are compared to LR, RF performs much better in identifying fraud. Furthermore, the RF model's performance is constant across training and testing datasets. As a result, there is no overfitting.

The outcomes of the two methods are compared in the table below:

Table 1. Comparison of Logistic Regression and Random Forest Results

ML Method	Precision		Recall	
	Test (%)	Train (%)	Test (%)	Train (%)
Random Forest	100	100	99.79	99.84
Logistic Regression	90.12	91.03	51.7	50.88

Even if the results of the RF model are good, we should try to improve the results of LR by changing the parameters and correcting class imbalance. In the next part, we'll talk about these plans. We train the LR model on a part of the original training sample. To make an under-sampled training dataset, we keep all of the fraud cases and pick a random number of non-fraud cases that are the same size. We can now find the best LR model for the under-sampled dataset by changing the 'Cost function' and 'Regularization factor' factors. In the result below, the memory scores for different combos of the punishment function and the cost function are shown. So, the best LR model with undersampling (l1 penalty and C = 100) and a recall of 50% has a recall of 50%. The LR method is not as good as the standard RF model.The model uses ten trees (n_estimators) in the forest, but the highest depth is not limited. When the results of cross-validation are good, overfitting is no longer a problem. The picture below shows how important each element is to the RF model. The picture below shows which factors have the biggest effect on the fraud predictions.

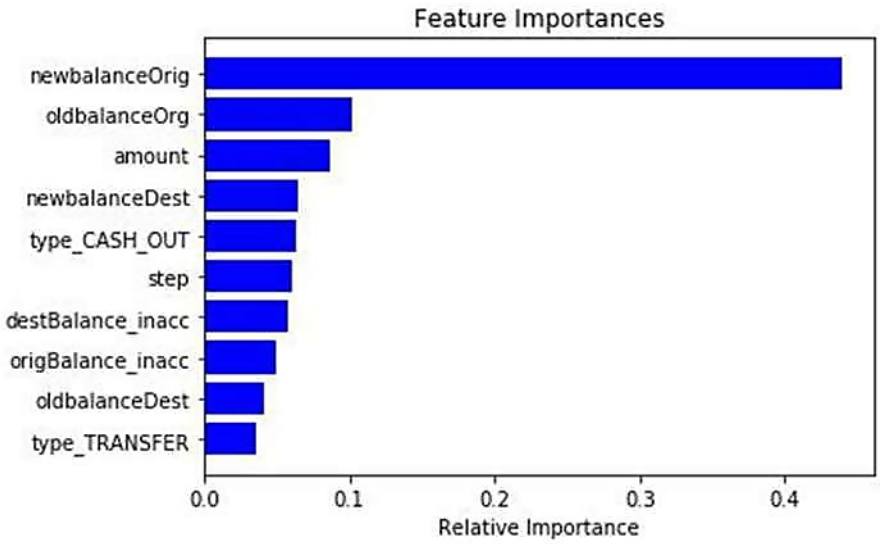

FIGURE 6. RF Method Feature Importance

Compared to all the other factors, the "newbalanceOrig" trait is the most important one for making the forecast. The picture below shows how the AUC and ROC curve were calculated for this model.

Summary of the Analysis

To find frauds, we looked at data on banking transactions and made an ML model. Part of the study was to clean up the data, do qualitative analysis, and use prediction models. During the data cleaning process, we checked for missing values, changed the data types, and made a summary of the factors in the data. In predictive models, we tried out both the LR and RF methods. We found that RF works best for this application because its accuracy and memory rates are almost 100%. We tried to improve the LR results by under sampling, but since a lot of data was left out, the results were the same. We used cross-validation to make sure the models weren't too good. From our labeled dataset, we can say that detecting theft in financial activities works, and RF is the best way to do this.

Limitations of the Research

In this work, we assessed the efficacy of employing certain supervised ML approaches to tackle the issue of financial transaction fraud detection. The following are the limitations of the methodologies used in this study:

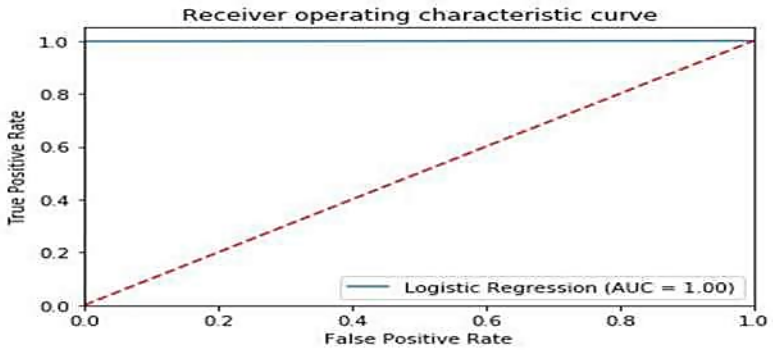

FIGURE 7. RF ModelROC curve

1. To train the algorithms, we utilized a pre-labeled dataset. However, it is sometimes difficult to locate labeled data, making the use of supervised ML approaches impractical. In such circumstances, we should consider unsupervised approaches that are beyond the scope of this work.

2. This study considers digital transaction data such as the quantity transmitted, the time of the transaction, and the balances of the recipient and the originator. These characteristics of detecting fraud may not be applicable to other types of financial transactions.

3. We have utilized RF and LR method.Even though the results of the research employing these algorithms are promising, additional strategies must be evaluated to determine which algorithm is most effective for this particular use case.

4. Due to the vast magnitude of the data, we were restricted in our computational power to investigate other strategies for parameter tuning and the SMOTE sampling methodology. These strategies may aid in enhancing the study's findings.

5. CONCLUSION

At last, we created a method for spotting suspicious financial activities in databases. This framework will help you better understand the many facets of fraud detection, such as the generation of derived variables to assist in class separation, the adjustment of class imbalance, and the choice of the most appropriate ML algorithm. Two ML algorithms, LR and RF, were compared and contrasted. Since RF outperformed LR, we may infer that tree-based algorithms work well with transaction data that has well-defined classes. This emphasizes the need for thorough exploratory research to be conducted before ML models are developed. A few indicators emerged from our exploratory analysis that served to further differentiate the groups from the raw data. The study's results will be useful for firms and organizations looking to establish or improve their financial fraud detection systems utilizing ML methods. Organizations may strengthen their defenses against fraud, reduce financial losses, and safeguard their stakeholders from possible damage by using these modern technologies.

FUNDING

None

ACKNOWLEDGEMENT

None

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

[1] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit card fraud detection: A novel approach using aggregation strategy and feedback mechanism," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3637–3647, 2018.

[2] R. B. Sulaiman, V. Schetinin, and P. Sant, "Review of machine learning approach on credit card fraud detection," *Human-Centric Intelligent Systems*, vol. 2, no. 1-2, pp. 55–68, 2022.

[3] A. Karthikeya, Y. B. Sai, S. Hariharan, A. C. Rao, J. D. Jignash, and A. B. Prasad, "Prevention of Cyber Attacks Using Deep Learning," *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, vol. 1, pp. 1332–1336, 2023.
[4] A. Ali, S. Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, and M. Nasser, "Financial fraud detection based on machine learning: a systematic literature review," *Applied Sciences*, vol. 12, no. 19, pp. 9637–9637, 2022.
[5] X. Lei, U. H. Mohamad, A. Sarlan, M. Shutaywi, Y. I. Daradkeh, and H. O. Mohammed, "Development of an intelligent information system for financial analysis depend on supervised machine learning algorithms," *Information Processing & Management*, vol. 59, no. 5, pp. 103036–103036, 2022.
[6] Z. Zhao and T. Bai, "Financial Fraud Detection and Prediction in Listed Companies Using SMOTE and Machine Learning Algorithms," *Entropy*, vol. 24, no. 8, pp. 1157–1157, 2022.
[7] H. Ye, L. Xiang, and Y. Gan, "Detecting financial statement fraud using random forest with SMOTE," *IOP Conference Series: Materials Science and Engineering*, vol. 612, pp. 52051–52051, 2019.
[8] M. A. Sharma, B. G. Raj, B. Ramamurthy, and R. H. Bhaskar, "Credit Card Fraud Detection Using Deep Learning Based on Auto-Encoder," *ITM Web of Conferences*, vol. 50, 2022.
[9] O. Vynokurova, D. Peleshko, O. Bondarenko, V. Ilyasov, V. Serzhantov, and M. Peleshko, "Hybrid machine learning system for solving fraud detection tasks," *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pp. 1–5, 2020.
[10] A. K. Rai and R. K. Dwivedi, "Fraud detection in credit card data using unsupervised machine learning based scheme," *2020 international conference on electronics and sustainable communication systems (ICESC)*, pp. 421–426, 2020.
[11] S. K. Majhi, S. Bhatachharya, R. Pradhan, and S. Biswal, "Fuzzy clustering using salp swarm algorithm for automobile insurance fraud detection," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 3, pp. 2333–2344, 2019.
[12] N. Rtayli and N. Enneya, "Selection features and support vector machine for credit card risk identification," *Procedia Manufacturing*, vol. 46, pp. 941–948, 2020.
[13] K. Randhawa, "Credit Card Fraud Detection Using AdaBoost and Majority Voting," *IEEE Access*, vol. 6, pp. 14277–14284, 2018.
[14] G. Liu, J. Tang, Y. Tian, and J. Wang, "Graph Neural Network for Credit Card Fraud Detection," *2021 International Conference on Cyber-Physical Social Intelligence (ICCSI)*, pp. 1–6, 2021.
[15] F. Carcillo, Y. A. L. Borgne, O. Caelen, and G. Bontempi, "Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization," *International Journal of Data Science and Analytics*, vol. 5, pp. 285–300, 2018.
[16] T. . Ntnu, "Synthetic Financial Datasets for Fraud Detection." https://www.kaggle.com/ntnu-testimon/paysim1.