10th International Conference on Control Engineering &Information Technology (CEIT-2025)
Proceedings Book Series –PBS- Vol 23, pp.170-178

Machine Learning on Ethereum Transactions: Addressing Imbalanced Data for Fraud Detections

Inass A. Husien[#1], Faraj O. Ehtiba[#2], Haitham S. Ben Abdelmula[#3], Akram Mohammed Imbarak[#4]

[#1,2]*Department of Computer Science, School of Basic Science, Academy for Postgraduate Studies Misurata, Libya*
[#3,4]*Department of Computer Networks Research, Libyan Center for Programming, Electronic Systems and Aviation Research, Libya*
[#2]*Research and Consulting Center – Misurata University, Libya*
[1]Inass.husien@gmail.com, [2]f.ehtiba@lam.edu.ly, [3]hsaa8383@gmail.com, [4]akram.ambark@epc.ly

***Abstract*— Fraudulent transactions pose a significant threat to the security of blockchain networks. This study investigates the effectiveness of a Random Forest model for fraud detection on imbalanced Ethereum transaction data. The inherent class imbalance, with far fewer fraudulent transactions compared to legitimate ones, can hinder model performance. To address this challenge, we employed data pre-processing techniques including Principal Component Analysis (PCA) for dimensionality reduction and random under-sampling to balance the class distribution. The Random Forest model achieved good performance on both imbalanced and balanced datasets. Although a slight decrease in metrics was observed on the balanced data, this trade-off enhances a model reliability, making it more suitable for real-world fraud detection applications.**

***Keywords*— Ethereum, Fraud Detection, Machine Learning, Data Imbalance, Random Forest**

I. INTRODUCTION

Blockchain is a decentralized, distributed digital ledger that is shared across a network of Peer-to-Peer computers. It stores information in a secure and transparent way, where data cannot be changed without affecting the entire record of subsequent blocks. This technology has become increasingly popular, especially in finance, because it enables secure online transactions and new investment methods [1]. Platforms like Ethereum have further enhanced blockchain by introducing programmability through smart contracts, enabling the development of decentralized applications (DApps), applications that run on a peer-to-peer network [2]. Ethereum is a leading blockchain platform that has enhanced the blockchain technology by introducing programmability through a Turing complete programming language called Solidity, enabling the creation of smart contracts [2]. Blockchain technology has evolved beyond simple transactions, enabling the creation of complex and decentralized digital systems and applications.

Blockchain technology has seen rapid adoption in recent years. Between 2016 and 2020, the number of blockchain wallets increased significantly, reaching 40 million. Investment in blockchain technology has also grown substantially, with approximately $2.9 billion invested in 2019, an 89% increase from the previous year. It is projected that corporate investment in blockchain will reach $12.4 billion by 2022. A recent survey suggests that blockchain is already widely used in 38% of US businesses, with an additional 44% anticipating widespread adoption within the next three years. [4].

Despite blockchain's inherent security features, such as cryptographic techniques and distributed consensus mechanisms, which have enabled its widespread adoption, the growing use of this technology has also surfaced concerns about its susceptibility to fraudulent activities. The distributed nature of blockchain networks can make it challenging to centrally monitor and detect suspicious transactions, further exacerbating the challenge of imbalanced data in fraud detection. In 2017, attackers hacked the Coin Dash ICO website and replaced the legitimate ICO contract address with their own. When users sent ether to the fraudulent address during the ICO's short 30-minute window, the attacker was able to steal $7.4 million before the real address could be restored [3].

One of the key challenges in combating blockchain-related frauds is the imbalanced nature of the data involved. Typically, fraudulent transactions on a blockchain network are far rarer than legitimate ones, posing

ISSN: 2961-6611

a significant challenge for effective fraud detection [4]. This imbalance in the data distribution can make it difficult for traditional fraud detection algorithms to accurately identify and flag suspicious activities, as they may be overwhelmed by the sheer volume of legitimate transactions.

The inherent imbalance between fraudulent and legitimate transactions on blockchains presents a significant hurdle for accurate fraud detection using traditional machine learning algorithms. To address this challenge, this paper investigates the effectiveness of a Random Forest model employed in conjunction with various data preprocessing techniques for imbalanced Ethereum transaction data.

The remainder of this paper will describe the methodology employed to answer these research questions. The dataset of Ethereum transactions used for the analysis, the specific implementation of the Random Forest model, and the application of Principal Component Analysis (PCA) and under sampling techniques for data preprocessing will all be described. Subsequently, the evaluation metrics used to assess the model's performance will be presented and the obtained results will be discussed in detail. Finally, conclusions will be drawn regarding the effectiveness of each data preprocessing technique in conjunction with the Random Forest model for fraud detection on imbalanced Ethereum transaction data.

II. FRAUDULENT TRANSACTIONS

Fraudulent transactions in blockchain refer to transactions that are intentionally designed to deceive or manipulate the blockchain network. These transactions can take various forms and have significant economic and trust implications for the network. Here are some key types of fraudulent transactions in blockchain:

A. Malicious Smart Contracts

Attackers create smart contracts with subtle bugs or exploits that allow them to steal funds or manipulate the blockchain. For example, a contract that records a deposited amount in one variable but the withdraw function references a different, uninitialized variable that always returns 0, preventing users from withdrawing their funds [1].

B. Phishing and Malicious Websites

Fake websites or phishing sites are created to steal user credentials or private keys, allowing attackers to access and steal funds. Fake clones of popular wallets could be made like My Ether Wallet that appear at the top of search results to phish for users' private keys [1].

C. Honeypot Attacks

Honeypot attacks lure victims into interacting with vulnerable smart contracts that then trap their funds [1].

D. Mempool Manipulation

Attackers exploit the mem-pool, the pool of unconfirmed transactions, to reorder transactions and extract value. The attacker would place a buy order slightly above the current market price. The victim then places a large market order, causing the price to rise temporarily. The attacker then quickly sells their position at the inflated price, profiting from the price movement caused by the victim's trade. This sandwich attack then allows the attacker to extract value from the victim's trade without taking on significant risk [1].

E. Initial Coin Offering (ICO) Scams

ICO scams involve replacing legitimate ICO contract addresses with fraudulent ones, allowing attackers to steal funds during the ICO period. The Coin Dash ICO attack had had attackers hack the website and replace the legitimate ICO contract address with their own, stealing $7.4 million before the real address could be restored [1].

F. Authentication Fraud

Low-volume, distributed login attempts spread over a long period, intended to evade detection.

G. Blockchain Accounting Fraud

Fraudulent practices in blockchain accounting, including the manipulation of financial records. The American Association of Certified Fraud Examiners (ACFE) reported fraudulent transactions valued at US$4 trillion in 2017 [5].

ISSN: 2961-6611

III. RELATED WORK

A. Fraud Detection in Blockchain

Fraud detection in blockchain systems using machine learning has garnered significant attention due to the escalating sophistication of fraudulent activities in cryptocurrency networks [6] addressed the challenge of detecting fraudulent transactions within the Ethereum blockchain by employing machine learning and deep learning approaches; Decision Trees, Logistic Regression, Gradient Boosting, XGBoost, and an innovative hybrid model that melds random forests with Deep Neural Networks (DNN). Their work underscores the importance of utilizing advanced technologies to effectively combat fraudulent activities. [7] also emphasized the significance of using machine learning models for anomaly detection in blockchain networks, demonstrating a comprehensive strategy to identify and mitigate anomalies effectively.

A new framework was introduced by [8] for fraud detection in Bitcoin transactions through an ensemble stacking model of Decision Tree, Naive Bayes, K-Nearest Neighbours, and Random Forest, showcasing the scalability and efficiency of machine learning in handling large volumes of transaction data for fraud detection purposes. Researchers have found that tree-based machine learning algorithms perform best on these imbalanced cryptocurrency datasets [9]. For instance, the use of the random forest algorithm has been proposed as an efficient method for fraud detection in datasets with a smaller number of fraud occurrences [10].

Additionally, [2] utilized Random Forest (RF), among other machine learning techniques, to detect fraudulent accounts in the Ethereum network, emphasizing the applicability of Random Forest in enhancing fraud detection capabilities within blockchain ecosystems.

This research [10] integrated blockchain technology and machine learning algorithms to identify fraudulent transactions in the Bitcoin network, demonstrating the effectiveness of combining these technologies for developing robust fraud detection systems in cryptocurrency transactions. It compared the effectiveness of XGBoost and Random Forest algorithms for transaction classification within blockchain data, highlighting the suitability of Random Forest for fraud detection in blockchain systems.

The paper [11] also proposed a machine learning based method for detecting fraudulent accounts on the Ethereum blockchain. The authors compared three classifiers: Random Forests, Support Vector Machines (SVM), and XGBoost. Expectingly, Random Forests achieved the best results in terms of recall and false positive rate, making it the most effective model for detecting fraudulent accounts.

Similarly, the Random Forest algorithm demonstrated in [12] an accuracy rate of 99.84% in detecting anomalies in the Ethereum blockchain over three other ML algorithms, K Nearest Neighbours (KNN), Gaussian Naive Bayes (Gaussian NB), Stochastic Gradient Descent (SDG), showcasing its robustness in identifying fraudulent activities within blockchain systems.

B. Handling Imbalanced Data

Blockchain data, particularly cryptocurrency transaction data, often suffers from a class imbalance problem. This means there is a significant disparity in the number of samples between the majority class (normal transactions) and the minority class (fraudulent or anomalous transactions). Ethereum blockchain data encounters class imbalance due to the limited known labels of illicit activities in the network. Conventional machine learning algorithms tend to be biased towards the majority class, making it difficult to accurately identify the minority class (fraudulent transactions). Additionally, imbalanced data can mislead the detection process, leading to misclassification issues. Real transaction datasets often face imbalanced problems, necessitating the use of advanced artificial intelligence approaches to address these challenges [13].

Feature engineering plays a crucial role in fraud detection tasks, especially in scenarios where the features obtained from industries are limited. Applying feature engineering methods and reforming the dataset are essential steps to enhance the performance of fraud detection models [14]. The deployment of robust machine learning models, coupled with dynamic feature selection techniques, has been proposed to enhance anomaly detection in cybersecurity tasks, leveraging comprehensive datasets with multiple attack types [15].

Furthermore, an approach to handling imbalanced data in machine learning could involve equalizing the dataset to address the issue of imbalanced data, ensuring that the model is trained on a balanced representation of both fraudulent and legitimate transactions [1].

ISSN: 2961-6611

In addition, resampling techniques like oversampling and under-sampling have been applied to these imbalanced blockchain datasets to improve classification performance. Under-sampling methods have achieved over 99% accuracy by removing noisy data points [16].

Employed XGBoost and Random Forest algorithms in [10] for fraudulent transaction classification, the paper addressed fraud and anomalies in the Bitcoin network. And for added enhancement, data balancing techniques used to generate synthetic malicious data points through Synthetic Minority Oversampling Technique (SMOTE) to achieve better results. [17] also used SMOTE and got better results and a more generalized mode.

While various machine learning approaches have demonstrated effectiveness in detecting fraudulent activities within blockchain systems, the challenge of imbalanced data poses a significant hurdle for accurate identification of fraudulent transactions. To address this challenge, this paper proposes a methodology that leverages the strengths of Random Forest, a tree-based algorithm known for its performance on imbalanced datasets. We will investigate the impact of data preprocessing techniques, specifically Principal Component Analysis (PCA) for dimensionality reduction and under sampling to balance the class distribution, on the performance of the Random Forest model. By comparing the effectiveness of these techniques, this study aims to identify an optimal approach for enhancing fraud detection accuracy in imbalanced Ethereum transaction data.

IV. METHODOLOGY

A. Data Collection

The data for this study was obtained from the Kaggle repository [18], a public platform for sharing datasets. The dataset comprises 9,841 Ethereum transactions, which is relatively small compared to other datasets of the same concept. But the dataset was chosen due to its small size since it would allow for faster processing.

Each tuple is characterized by 50 features, including information about transaction value, sender and receiver addresses, and potential ERC-20 token involvement. The data includes 7,662 legitimate transactions and 2,179 fraudulent transactions, reflecting the class imbalance typically encountered in fraud detection tasks involving blockchain data as shown in Fig 1. Some initial cleaning steps were performed to address missing values, and feature scaling will be applied during the data preprocessing stage to ensure all features are on a similar scale for model training.

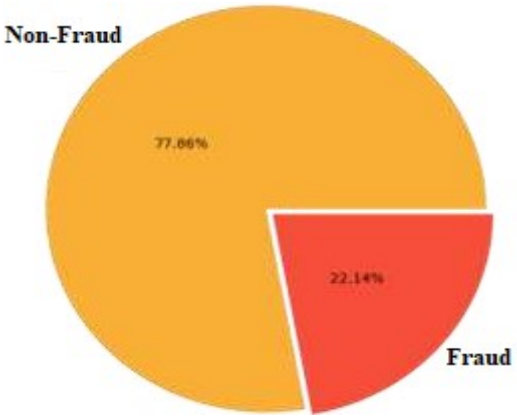

Fig. 1 Imbalanced Data Target Distribution

B. Dataset Preprocessing

1) Redundant Feature Elimination: The initial data contained a column designated "Unnamed: 0," likely an artifact of the data creation process. This column was eliminated due to its irrelevance to fraud detection analysis. Additionally, an automated approach identified and eliminated features solely containing zero values. These features were deemed redundant and unlikely to contribute meaningfully to the subsequent analysis. Furthermore, columns like "Address" and "Index" were removed as they were not considered directly relevant for characterizing fraudulent transactions.

ISSN: 2961-6611

2) Missing Value Imputation: The dataset exhibited missing values within certain numerical features. To address this, a technique known as mean imputation was employed. This involved calculating the mean (average) value for each numerical feature (excluding those related to ERC-20 tokens) and utilizing this value to impute the missing entries. This approach assumes that missing values are likely distributed around the average value for that specific feature. For categorical features containing missing values ("ERC20 most sent token type" and "ERC20_most_rec_token_type"), mode imputation was implemented. This involved replacing missing values with the most frequently occurring value ("0") within each respective feature. This approach assumes that missing token types are less prevalent than existing ones within the dataset.

3) Feature Categorization and Correlation Analysis: The features were categorized based on their data types, differentiating numerical features from categorical features. Subsequently, correlation analysis was conducted to identify features exhibiting high correlation coefficients (exceeding a threshold of 0.8). Features with high correlation can introduce redundancy and potentially impede model performance. Pairs of features demonstrating high correlation were identified, and one feature from each pair was eliminated. This process aimed to retain informative features while mitigating redundancy within the dataset.

4) Categorical Data Standardization: Inconsistencies were observed in the representations of missing values within the "ERC20_most_rec_token_type" feature ("None", " ", None). To ensure data uniformity, these variations were all replaced with a consistent value ("0").

C. Handling Imbalanced Data

The imbalanced nature of the dataset, with a significantly lower number of fraudulent transactions compared to legitimate ones, presented a challenge for effective model training. To address this challenge, two data pre-processing techniques were employed: Principal Component Analysis (PCA), and random under-sampling.

1) Principal Component Analysis (PCA): Principal Component Analysis (PCA) is a dimensionality reduction technique that can be beneficial for both feature engineering and improving model performance. This process involves identifying a new set of features, called principal components (PCs), that capture the maximum variance in the original data. By reducing the dimensionality, PCA can potentially mitigate the impact of irrelevant features and improve the computational efficiency of the model training process.

2) Random Under-Sampling: Given the significant class imbalance, a technique called random under-sampling was employed. This technique involves randomly removing data points from the majority class (legitimate transactions) to achieve a more balanced representation of both fraudulent and legitimate transactions in the training data. This helps to ensure that the model is not biased towards the majority class and can learn effectively from the limited number of available fraudulent transactions.

D. Random Forest Model

Following the data pre-processing steps, a Random Forest classification model was employed to classify the Ethereum transactions within the imbalanced dataset. Random Forest is a machine learning algorithm known for its effectiveness in handling imbalanced data due to its inherent ensemble nature. It operates by creating multiple decision trees, each trained on a random subset of features and data points with replacement (bootstrapping). The final prediction is based on the majority vote of these individual trees, reducing variance and improving model robustness. In this study, the Random Forest model was implemented using the scikit-learn library in Python. The model was configured with 100 decision trees and a fixed random state to ensure reproducibility of the results.

V. RESULTS AND DISCUSSION

A. Balancing Data

The initial dataset exhibited a significant class imbalance, with legitimate transactions constituting approximately 78% (7,662 transactions) and fraudulent transactions representing the remaining 22% (2,179 transactions). This imbalance can pose challenges for models tending to be biased towards the majority class. To address this issue, we employed two data pre-processing techniques: Principal Component Analysis (PCA) and random under sampling. PCA was applied to reduce the dimensionality of the data while preserving the

ISSN: 2961-6611

most important information for fraud detection. This not only improved computational efficiency but also potentially mitigated the impact of irrelevant features that might have exacerbated the class imbalance. Following PCA, random under-sampling was implemented to balance the class distribution. This technique involved randomly removing data points from the majority class (legitimate transactions) until the number of transactions in each class (fraudulent and legitimate) was approximately equal (around 50%) as illustrated in Fig 2.

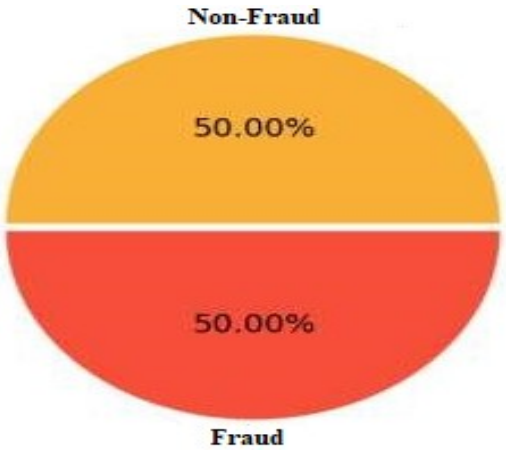

Fig. 2 Balanced Data Target Distribution

The resulting balanced dataset ensured that the model was not biased towards the majority class and could learn effectively from the limited number of available fraudulent transactions.

The effectiveness of these pre-processing steps is further supported by the good performance achieved by the Random Forest model, as discussed in the following section.

B. Model Evaluation

To assess the performance of the Random Forest model, the pre-processed data was split into training and testing sets using a stratified split. This technique ensures that the class distribution (proportion of fraudulent and legitimate transactions) is maintained in both the training and testing sets. A 70/30 split was used, allocating 70% of the data for training 30% for testing. The trained model was then evaluated on the unseen testing data to assess its generalizability and ability to accurately predict fraudulent transactions on new data.

The performance of the Random Forest model was evaluated using two key metrics: accuracy and classification report, which will be calculated by equations 1 to 4. Accuracy is a basic metric that represents the overall proportion of correct predictions made by the model. The classification report provides a more detailed breakdown of the model's performance, including precision, recall, F1-score, and support for each class (fraudulent and legitimate transactions). These metrics offer insights into the model's ability to identify both fraudulent and legitimate transactions effectively, especially in the context of imbalanced data.

$$Precision = \frac{T_P}{T_P + F_P} \quad (1)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (2)$$

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad (3)$$

$$F1 - Score = 2.\frac{Precision\ .\ Recall}{Precision + Recall} \quad (4)$$

ISSN: 2961-6611

Where are, T_P = number of true positives, F_P = number of false positives, F_N = number of false negatives and T_N = number of true negatives.

The results are revealed in Table I, which presents the key evaluation metrics: Accuracy, Precision, Recall, and F1-Score. The imbalanced data show exceptionally high performance across all metrics at approximately 0.99. While this seems promising, it's crucial to consider the potential impact of class imbalance on these metrics. In imbalanced datasets, models can achieve high accuracy simply by predicting the majority class (legitimate transactions in this case). Therefore, a more nuanced analysis of the other metrics, particularly Recall, is necessary.

TABLE I
PERFORMANCE EVALUATION

Random Forest	Accuracy	Precision	Recall	F1-Score
Imbalanced Data	0.99	0.99	0.99	0.99
Balanced Data	0.97	0.97	0.97	0.97

The balanced data results, achieved through PCA and random under-sampling, demonstrate a slight decrease in performance compared to the imbalanced data. All metrics are now around 0.97. However, this decrease is relatively minor, and the balanced data offers a more reliable evaluation of the model's ability to accurately detect both fraudulent and legitimate transactions.

Overall, the results indicate that the Random Forest model performs well in detecting fraudulent transactions. While addressing class imbalance led to a minor decrease in overall metrics, the balanced data offers a more reliable measure of the model's effectiveness for real-world fraud detection.

VI. CONCLUSIONS

This study investigated the effectiveness of a Random Forest model in conjunction with data pre-processing techniques for fraud detection on imbalanced Ethereum transaction data. The key findings are summarized below:

- **Impact of Imbalanced Data:** The imbalanced nature of the dataset, with a significantly lower number of fraudulent transactions, presented a challenge. This could be attributed to a bias towards the majority class (legitimate transactions).
- **Effectiveness of Pre-processing:** The employed data preprocessing techniques, including Principal Component Analysis (PCA) and random under-sampling, successfully reduced dimensionality while preserving important information and balancing the class distribution.
- **Model Performance:** The Random Forest model demonstrated good performance on both the imbalanced and balanced datasets. While a slight decrease in metrics was observed on the balanced data, this is a fair trade-off for achieving a more robust model.

REFERENCES

[1] S. Sh. Taher, S. Y. Ameen, and J. A. Ahmed, “Advanced Fraud Detection in Blockchain Transactions: An Ensemble Learning and Explainable AI Approach,” Engineering, Technology and Applied Science Research., vol. 14, no. 1, pp. 12822–12830, 2024.
[2] A. Sallam et al., “Fraudulent Account Detection in the Ethereum’s Network Using Various Machine Learning Techniques,” International Journal of Software Engineering and Computer Systems., vol. 8, no. 2, pp. 43–50, 2022.
[3] (2024) The eMarketer website. [Online]. Available: https://www.emarketer.com/insights/blockchaintechnology-applications-use-cases/
[4] N. Deepa et al., “A Survey on Blockchain for Big Data: Approaches, Opportunities, and Future Directions,” Future Generation Computer Systems, vol. 131, pp. 209–226, Sept. 2020.
[5] U. Mahtani, “Fraudulent practices and blockchain accounting systems,” Journal of Accounting, Ethics and Public Policy., vol. 23, no. 1, pp. 97–148, 2022.
[6] S. Siddamsetti and M. Srivenkatesh, “Efficient Fraud Detection in Ethereum Blockchain Through Machine Learning and Deep Learning Approaches,” International Journal on Recent and Innovation Trends in Computing and Communication., vol. 11, no. 11, pp. 71–82, 2023.

ISSN: 2961-6611

[7] S. Hisham, M. Makhtar, and A. A. Aziz, “Combining Multiple Classifiers using Ensemble Method for Anomaly Detection in Blockchain Networks: A Comprehensive Review,” International Journal of Advanced Computer Science and Applications., vol. 13, no. 8, 2022.
[8] N. Nayyer, N. Javaid, M. Akbar, A. Aldegheishem, N. Alrajeh, and M. Jamil, “A New Framework for Fraud Detection in Bitcoin Transactions Through Ensemble Stacking Model in Smart Cities,” IEEE Access., vol. 11, pp. 90916– 90938, 2023.
[9] I. Alarab and S. Prakoonwit, “Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques,” Data Science and Management., vol. 5, no. 2, pp. 66–76, Jun. 2022.
[10] T. Ashfaq et al., “A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism,” Sensors, vol. 22, no. 19, p. 7162, Sept. 2022.
[11] M. Ostapowicz and K. Żbikowski, “Detecting Fraudulent Accounts on Blockchain: A Supervised Approach,” the 20th International Conference on Web Information system Engineering, pp. 18–31, 2019.
[12] N. T. Anthony, M. Shafik, F. Kurugollu, and H. F. Atlam, “Anomaly Detection System for Ethereum Blockchain Using Machine Learning,” the 19th International Conference on Manufacturing Research, Sept.2022.
[13] D. Choi and K. Lee, “An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation,” Security and Communication Networks, vol. 2018, pp. 1–15, Sept. 2018.
[14] Y. Yazici, “Approaches to Fraud Detection on Credit Card Transactions using Artificial Intelligence Methods,” in Computer Science & Information Technology, AIRCC Publishing Corporation, pp. 235–244, Jul. 2020.
[15] M. Ahsan, R. Gomes, M. M. Chowdhury, and K. E. Nygard, “Enhancing Machine Learning Prediction in Cybersecurity Using Dynamic Feature Selector,” Journal of Cybersecurity and Privacy, vol. 1, no. 1, pp. 199–218, Mar. 2021.
[16] T. A. Borges and R. F. Neves, “Ensemble of machine learning algorithms for cryptocurrency investment with different data resampling methods,” Appl Soft Computing, vol. 90, p. 106187, May. 2020.
[17] R. M. Aziz, M. F. Baluch, S. Patel, and P. Kumar, “A Machine Learning Based Approach to Detect the Ethereum Fraud Transactions with Limited Attributes,” Karbala International Journal of Modern Science, vol. 8, no. 2, pp. 139–151, 2022.
[18] (2024) Ethereum Fraud Detection Dataset. [Online]. Available: https://www.kaggle.com/datasets/vagifa/Ethereum fraud detection-dataset/data.

ISSN: 2961-6611