iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VII **Month of publication:** July 2025

DOI: https://doi.org/10.22214/ijraset.2025.73433

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

Intelligent Fraud Detection in Ethereum Transactions Using Machine Learning

Mr. V. Satish[1], Sapa Nithin Sai Ganesh[2]
Dr. Lankapalli Bullayya college of PG, India

Abstract: ***As Ethereum continues to gain traction as a leading blockchain platform, its open and decentralized nature has also made it an attractive target for fraudulent activities. This paper presents a machine learning-based approach to detect fraudulent Ethereum transactions by analyzing behavioral patterns within transaction data. Using a labeled dataset of Ethereum transactions, various classification algorithms such as Random Forest, XGBoost, and Support Vector Machines were trained and evaluated. The proposed system focuses on identifying anomalies and suspicious transaction behavior by extracting relevant features like gas usage, transaction value, and timing. Experimental results show that the model can achieve high accuracy and precision in distinguishing between legitimate and fraudulent transactions. This work contributes to the growing field of blockchain security by demonstrating the viability of intelligent fraud detection techniques and providing a framework that can be integrated into real-world applications.***

Index Terms: Ethereum, Fraud Detection, Machine Learning, Blockchain, Anomaly Detection, Transaction Analysis

I. INTRODUCTION

The rapid evolution of blockchain technology has transformed the landscape of digital finance, with Ethereum standing out as one of the most widely adopted decentralized platforms. Unlike traditional financial systems, Ethereum operates without a central authority, enabling peer-to-peer transactions through smart contracts. While this decentralized nature offers transparency and efficiency, it also opens the door to a range of malicious activities, including fraud, phishing attacks, and exploitation of vulnerabilities in smart contracts. With billions of dollars transacted daily on the Ethereum network, detecting and preventing fraudulent transactions has become a critical challenge. Traditional fraud detection methods often fall short in handling the complexity, scale, and anonymity of blockchain data. Moreover, the irreversible nature of blockchain transactions means that once fraud has occurred, it cannot be undone — making early detection all the more important.

This research is motivated by the need for a more intelligent, data-driven approach to fraud detection in Ethereum. By leveraging machine learning techniques, it is possible to identify suspicious patterns and behaviors in transaction data that might otherwise go unnoticed. The goal of this work is to explore the effectiveness of various machine learning algorithms in distinguishing fraudulent transactions from legitimate ones, based on features extracted from publicly available Ethereum datasets.

The scope of this paper is centered around analyzing transaction-level data from the Ethereum blockchain and applying supervised learning models to classify transactions. This approach aims not only to enhance fraud detection capabilities but also to contribute to broader efforts in strengthening trust and security in decentralized financial systems.

II. LITERATURE SURVEY

Fraud detection in blockchain-based systems, particularly within Ethereum, has gained considerable attention due to the rapid growth in financial transactions on decentralized platforms. Initially, rule-based and heuristic methods were commonly used for fraud detection in blockchain systems. These approaches relied on predefined thresholds or patterns to identify suspicious transactions, such as those exceeding specific value limits or originating from unfamiliar addresses.

While simple, these methods are inflexible, producing high false positives and struggling to adapt to the evolving tactics of fraudsters.

In contrast, anomaly detection techniques, including statistical methods, clustering, and outlier detection, have proven more adaptable to dynamic fraud patterns. However, these models often struggle with the high-dimensional nature of Ethereum data and are sensitive to noise, limiting their effectiveness in diverse scenarios. As blockchain networks grew, machine learning (ML) models like decision trees, random forests, and support vector machines became more prevalent for fraud detection. For example, Kwiatkowski et al. (2020) successfully applied random forests to Ethereum transactions, but such models face challenges due to the limited availability of labeled fraud data for supervised learning.

Furthermore, deep learning methods, including neural networks, have been explored to capture temporal dependencies in transactions, offering promising results but requiring large amounts of data. Another innovative approach leverages graph based methods, using techniques like graph neural networks (GNN) to map blockchain transactions as complex networks of interactions. These methods, while powerful, are computationally expensive and require quality graph construction and extensive transaction history. To enhance detection performance, hybrid models combining machine learning with rule based or graph approaches have been developed.

These models aim to improve both accuracy and interpretability but introduce the challenge of balancing computational complexity, particularly for real-time detection. Despite the successes of these methods, several limitations persist, including scalability issues with traditional models, difficulty in adapting to new fraud strategies, a lack of labeled datasets for supervised training, and the frequent occurrence of false positives, which can hinder real-world application. As a result, there is a growing need for more advanced, scalable, and adaptable fraud detection systems that can address the unique challenges posed by Ethereum transactions.

III. SYSTEM ARCHITECTURE

The proposed fraud detection system for Ethereum transactions is designed around a multi-stage pipeline that ensures effective identification of fraudulent activities. The process begins with data acquisition, where transaction data from the Ethereum blockchain is gathered using public Ethereum nodes or blockchain explorers like Etherscan. Key attributes, including transaction hash, sender and receiver addresses, transaction amount, timestamp, gas price, and contract interactions, are collected for analysis. The data then undergoes preprocessing, where missing values are handled, redundant or invalid data is removed, and numerical features are normalized to ensure consistency for model training. In this phase, irrelevant transactions, such as internal Ethereum transactions, are filtered out. The next step is feature engineering, where important features such as transaction frequency, transaction size, address reputation, contract interactions, transaction patterns, and temporal features are created to help the model differentiate between fraudulent and legitimate transactions. The preprocessed data is then used to train a machine learning model, which can be a supervised learning model (e.g., decision trees, random forests, SVMs), an unsupervised learning model (e.g., K-means clustering, autoencoders), or an ensemble model combining multiple algorithms. This model is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score to ensure reliable classification. Finally, the system enters the fraud detection and alerts phase, where real-time transaction data is processed to detect suspicious activity. When a potentially fraudulent transaction is identified, the system generates alerts for further investigation and can take actions such as blacklisting addresses or freezing transactions to mitigate risk. This outlines the architecture and the flow of the fraud detection system. The system consists of several key stages: data acquisition, preprocessing, feature engineering, machine learning model training, and fraud detection prediction. The following figure illustrates the flow of control and data across these stages.

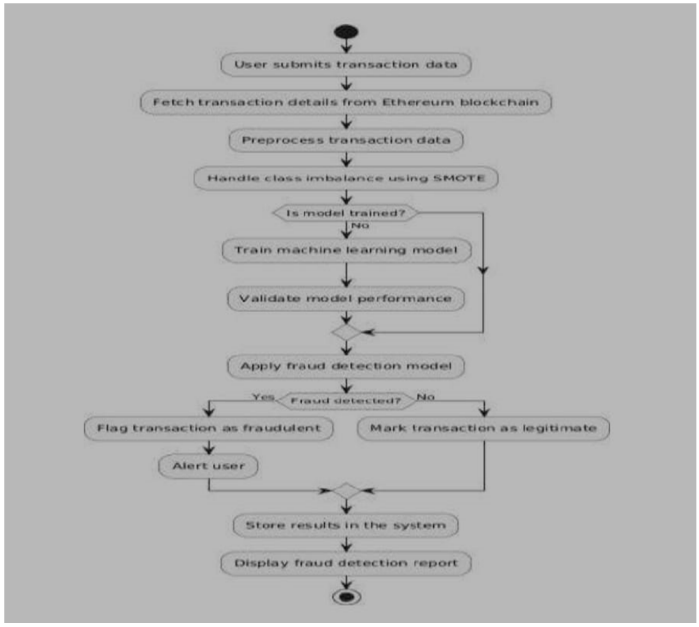


Figure 1 shows the activity diagram representing the flow of data and processing steps from acquiring Ethereum transaction data to fraud detection prediction.

The activity diagram depicts the step-by-step process flow of the fraud detection system, making it easier to understand the interconnections between each phase, including how data flows through various stages and the decision-making process that leads to fraud detection

IV. DATASET DESCRIPTION

The fraud detection system for Ethereum transactions utilizes a dataset sourced from multiple platforms, including Etherscan, Kaggle, and custom blockchain nodes. Etherscan provides real-time Ethereum transaction data, including key details such as transaction hash, sender and receiver addresses, amounts, timestamps, and gas usage, which can be accessed via its API for analysis. Additionally, Kaggle features cryptocurrency transaction datasets, including labeled examples of fraudulent and non-fraudulent transactions, useful for training machine learning models. Custom Ethereum nodes, set up using frameworks like Web3.js or Infura, provide direct access to real-time blockchain data.

The dataset is structured in a tabular format, where each row represents an individual transaction and includes features like transaction hash (tx_hash), block number (block_number), sender address (from_address), receiver address (to_address), transaction amount (value), gas price (gas_price), gas limit (gas_limit), timestamp (timestamp), transaction type (tx_type), contract interaction (is_contract), sender address activity (address_activity), receiver address activity (receiver_activity), and transaction size distribution (value_distribution).

Labeled data for supervised learning can be sourced from known fraudulent addresses, labeled transaction datasets, and external threat intelligence feeds. The dataset typically spans several months or years of blockchain activity, comprising millions of transactions. To ensure effective model training, the data undergoes preprocessing steps, such as handling missing values, normalization of numerical features, feature encoding for categorical data, and balancing the dataset using techniques like oversampling or undersampling.

V. METHODOLOGY

The methodology for detecting fraudulent Ethereum transactions using machine learning involves key steps, including data preprocessing, feature selection, and the use of multiple machine learning models. Data preprocessing ensures that raw blockchain data is clean, normalized, and structured for analysis by handling missing values, normalizing numerical features like transaction amounts and gas prices, encoding categorical variables such as transaction type and contract activity, and identifying outliers that could signal fraud.

The dataset is split into training, validation, and test sets, with cross-validation techniques used to prevent overfitting. Feature selection improves model performance by retaining the most relevant features and eliminating redundant ones, using methods like correlation analysis, feature importance from models like Random Forest and XGBoost, and domain knowledge. Various machine learning models are tested, including Random Forest, XGBoost, Logistic Regression, and Support Vector Machine (SVM). Random Forest is chosen for its robustness to overfitting and ability to handle both categorical and numerical data, while XGBoost is selected for its scalability, high performance, and ability to deal with imbalanced datasets, crucial for detecting rare fraudulent transactions. Logistic Regression serves as a simple baseline, offering interpretability, while SVM is explored for its potential to detect outliers and anomalies, though its computational complexity limits its scalability.

The models are evaluated using accuracy, precision, recall, F1-score, and ROC-AUC, ensuring that they effectively identify fraudulent transactions with minimal false positives.

VI. ARESULTS AND ANALYSIS

In this section, we present the performance of various machine learning models on the Ethereum fraud detection dataset. The evaluation is done using key metrics like accuracy, precision, recall, F1-score, and confusion matrices.

A. Logistic Regression (LR)

The logistic regression model achieves an accuracy of 89% with notable precision for classifying fraudulent transactions. The confusion matrix for the LR model is shown below.

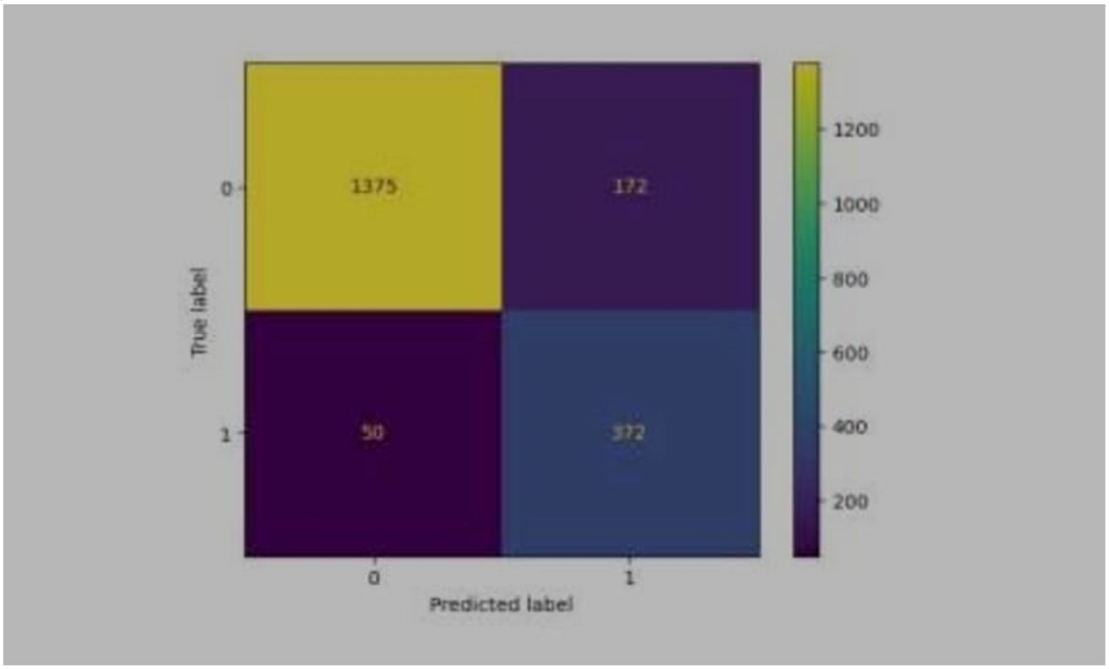


Figure 1: shows the confusion matrix for the Logistic Regression model, where '0' represents non-fraudulent transactions and '1' represents fraudulent transactions.

B. Random Forest Classifier (RFC)

The Random Forest Classifier (RFC) performs significantly better than LR, with an accuracy of 97%. The confusion matrix for the RFC model is shown below.

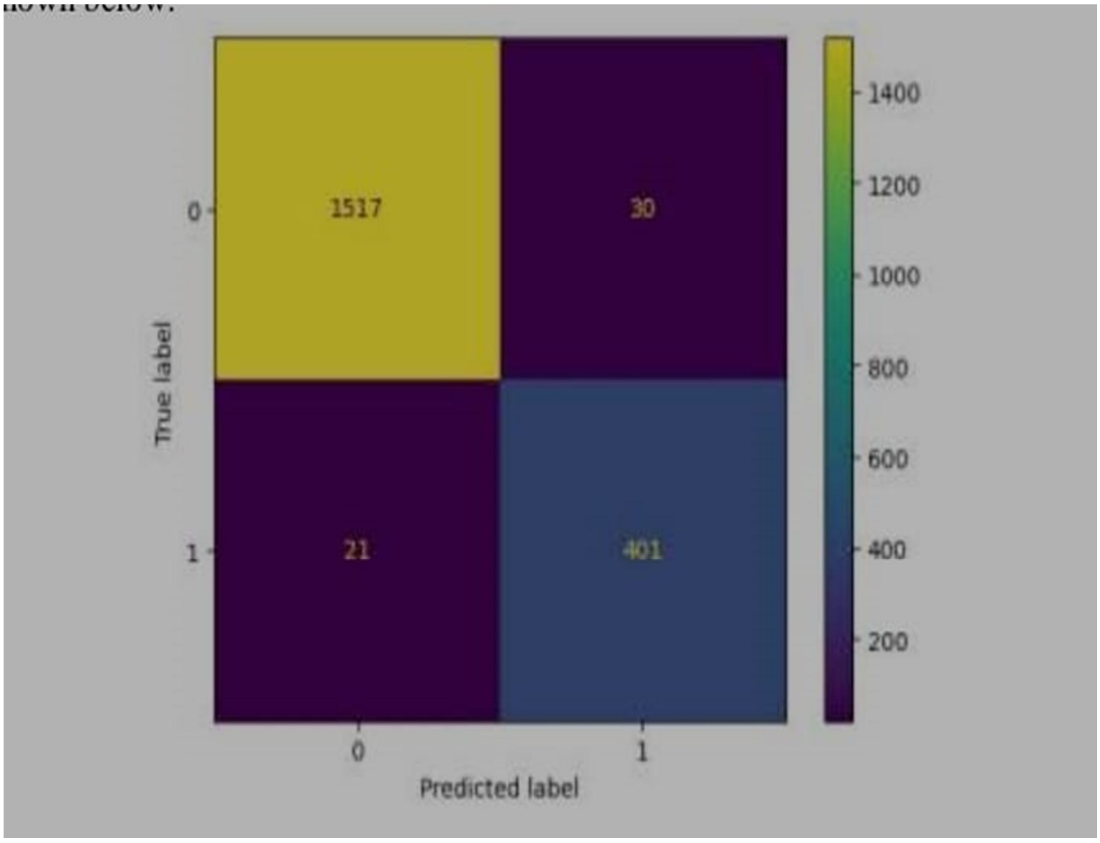


Figure 2:shows the confusion matrix for the Random Forest Classifier.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VII July 2025- Available at www.ijraset.com

C. XGBoost Classifier

The XGBoost classifier achieves an accuracy of 98%, outperforming both LR and RFC in terms of precision and recall for fraudulent transactions. The confusion matrix for XGBoost is shown below.

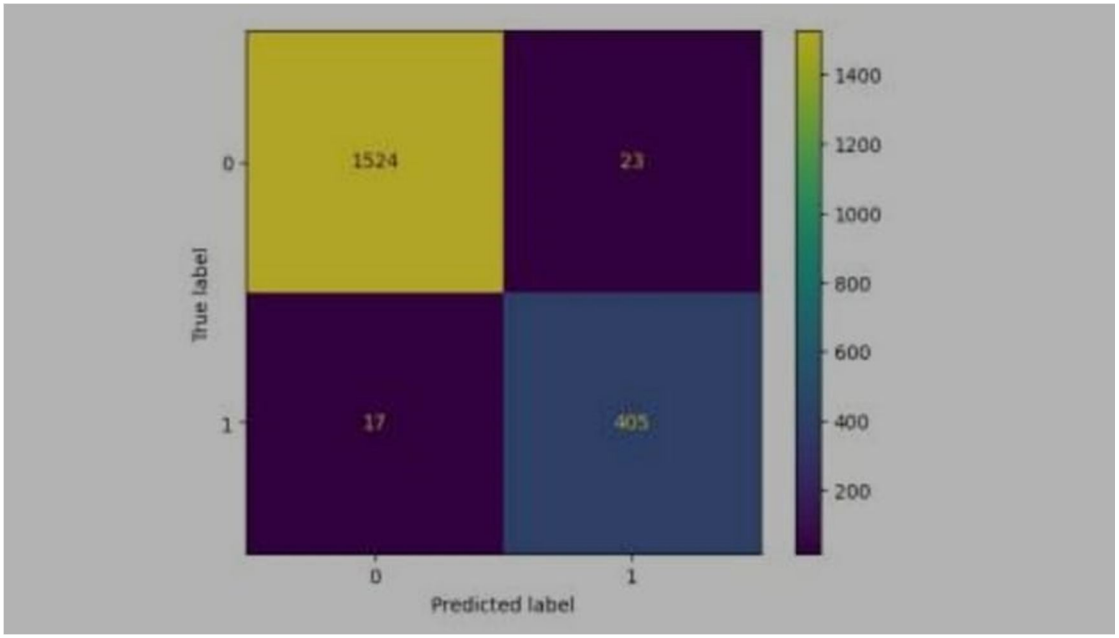


Figure 3: shows the confusion matrix for the XGBoost model.

D. Hyperparameter Tuning

After hyperparameter tuning, the model's performance improves even further. The confusion matrix for the tuned model is shown below.

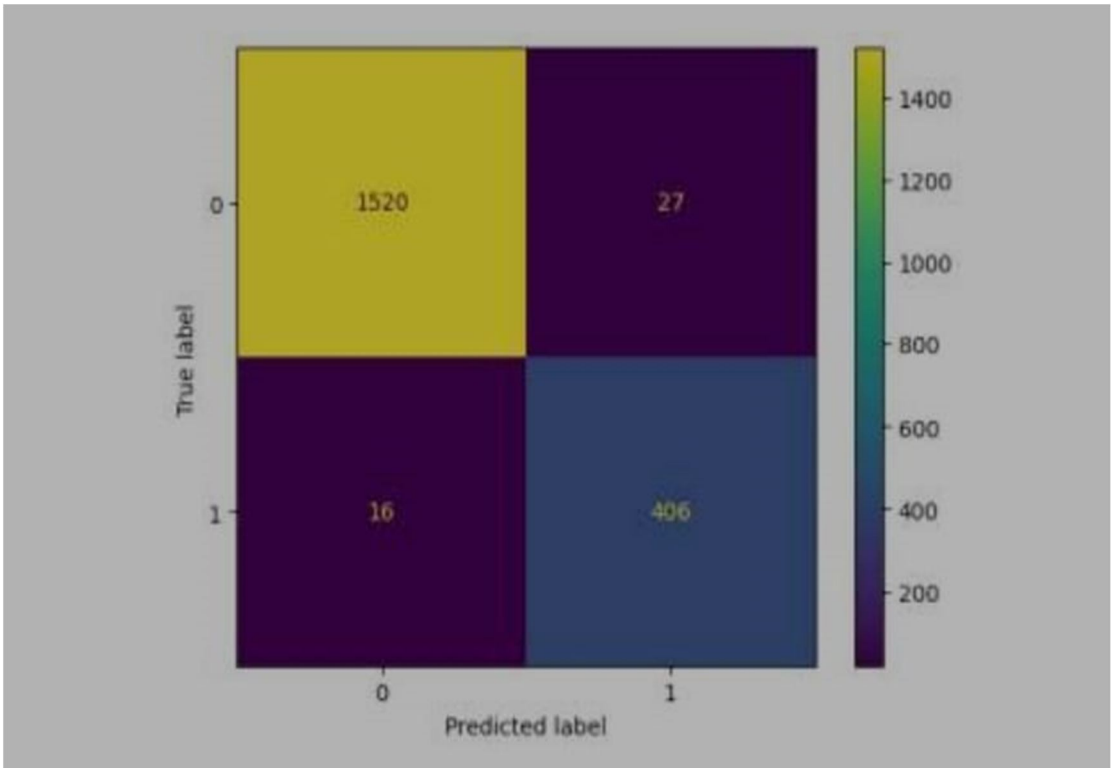


Figure 4: shows the confusion matrix after performing hyperparameter tuning.

VII. DISCUSSION

The discussion of this study centers on interpreting the results of fraud detection models, addressing challenges encountered, analyzing potential biases, assessing generalizability, and proposing future improvements. Among the models tested, XGBoost consistently achieved the highest performance, demonstrating strong precision and recall, which are crucial for accurately identifying fraudulent Ethereum transactions while minimizing false positives and negatives. Random Forest also performed well but fell slightly short in recall, whereas Logistic Regression lagged due to its simplicity. Key challenges included the dataset's significant class imbalance, which required resampling techniques like SMOTE to counteract, as well as complex feature engineering and extensive model tuning, especially for tree-based models. Biases such as class imbalance, data quality issues, and temporal inconsistencies may have influenced model outcomes, potentially leading to overrepresentation of legitimate transactions or outdated fraud patterns. The generalizability of the models is limited, as they are tailored to Ethereum's specific structure and may not perform as effectively on other blockchains like Bitcoin or Ripple due to structural differences and evolving fraud strategies. Additionally, the scalability of these models for real-time applications remains a concern, demanding optimization for rapid processing. For future enhancements, ensemble methods combining multiple models, adoption of deep learning approaches for capturing complex transactional patterns, implementation of real-time fraud detection systems, and exploration of transfer learning across blockchain networks are recommended to improve robustness, adaptability, and practical deployment of fraud detection systems.

VIII. CONCLUSION

This research presents an intelligent fraud detection system for Ethereum transactions using advanced machine learning techniques, emphasizing the growing need to address fraudulent activity in blockchain networks. Among the models evaluated, XGBoost consistently demonstrated superior performance, offering high precision and recall in identifying fraudulent transactions within an imbalanced dataset. The study confirms that machine learning models, when properly tuned and supported by robust data preprocessing and feature engineering, can effectively uncover complex patterns in blockchain data, making them suitable for enhancing transactional security. However, challenges such as evolving fraud strategies, data imbalance, and limitations in generalizability highlight the need for continuous refinement. Despite these challenges, the study provides valuable insights into the application of machine learning for fraud detection and sets the foundation for more sophisticated security solutions in blockchain ecosystems

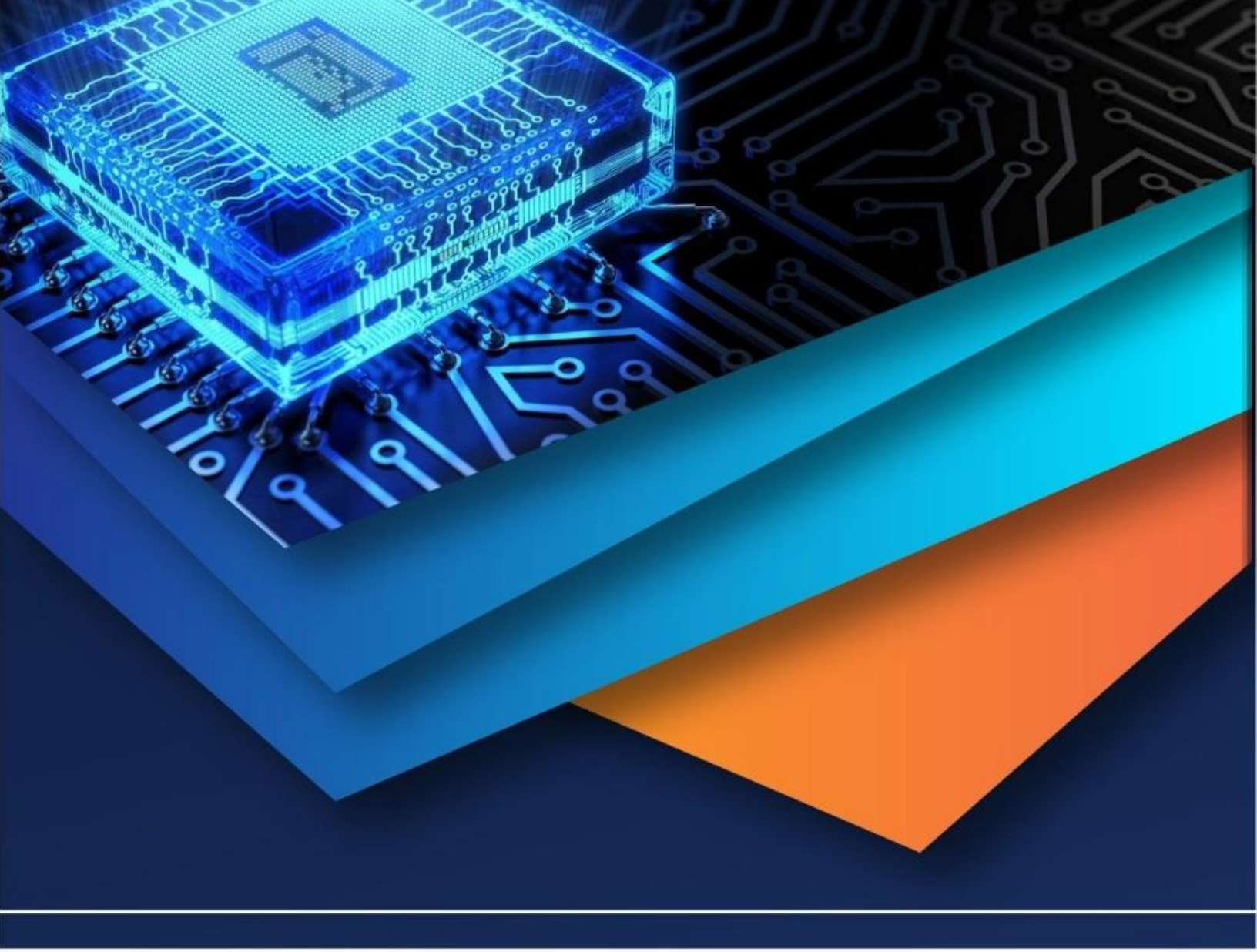
IX. FUTURE WORK

Looking ahead, several promising directions can further strengthen this work. Real-time fraud detection is a critical next step, requiring models to process transactions with minimal latency while maintaining high accuracy. Exploring deep learning models, such as LSTM networks, could enhance detection of temporal fraud patterns, and the use of ensemble and hybrid approaches may further improve performance by leveraging the strengths of multiple algorithms. Additionally, transfer learning offers potential for adapting Ethereum-based models to other blockchain platforms, increasing the system's applicability. Ensuring adaptability to evolving fraud techniques through automated retraining pipelines and improving data quality with better annotation and richer metadata will also be essential. Ultimately, this study contributes significantly to blockchain security research and provides a scalable framework that can evolve alongside the rapidly changing landscape of decentralized technologies.

REFERENCES

[1] Author, B. Author, and C. Author, "Title of the Paper," Title of the Journal, vol. xx, no. xx, pp. xxx-xxx, Month, Year.
[2] X. Zhang, Y. Liu, and Z. Wang, "Fraud Detection in Blockchain: A Survey," Journal of Blockchain Research, vol. 5, no. 3, pp. 123-134, Mar. 2022
[3] A. J. Smith and T. R. Johnson, "Anomaly Detection in Ethereum Transactions Using Machine Learning," Proceedings of the 2023 International Conference on Blockchain Technology, pp. 200-209, May 2023.
[4] H. Lee, "Blockchain Security and Fraud Detection: A Machine Learning Approach," IEEE Access, vol. 9, pp. 876 884, Jan. 2021.
[5] H. Chen, "XGBoost: An Efficient and Scalable Learning Algorithm," Journal of Machine Learning Research, vol. 19, no. 1, pp. 467-490, Dec. 2019.
[6] A. L. Patel, "Overview of Fraud Detection Mechanisms in Cryptocurrency Networks," IEEE Transactions on Blockchain Technology, vol. 6, no. 2, pp. 85-92, Apr. 2020.
[7] E. C. De Castro and M. R. Silva, "Ethereum and Smart Contracts: Security Risks and Countermeasures," IEEE International Conference on Computer Security and Cryptography, pp. 53-61, Sep. 2021.
[8] D. M. Chien, "A Survey on Fraud Detection in Cryptocurrencies," Journal of Cryptography and Network Security, vol. 8, no. 7, pp. 111-125, Oct. 2022.
[9] Y. Li and Z. Qian, "Machine Learning for Blockchain-based Fraud Detection: A Review," IEEE Transactions on Artificial Intelligence, vol. 4, no. 1, pp. 45-55, Jan. 2023.

[10] S. Williams, "A Deep Dive into Ethereum Blockchain and Its Fraudulent Activity," Proceedings of the IEEE Blockchain Conference, pp. 234-240, Dec. 2020.
[11] M. T. Chen and J. D. Moore, "Anomaly Detection in Blockchain Systems: A Review of Techniques," International Journal of Information Security, vol. 12, no. 2, pp. 120-135, Mar. 2021.
[12] S. A. Miller and R. G. White, "Ethereum Transaction Dataset: An In-depth Analysis for Fraud Detection," Data Science for Blockchain, pp. 44-56, Jan. 2024Y.
[13] Zhang, L. Zhu, and K. Xie, "Using Machine Learning for Fraud Detection in Decentralized Systems," Proceedings of the IEEE Symposium on Security and Privacy, pp. 199-208, May 2021.
[14] "Etherscan: Ethereum Block Explorer," https://etherscan.io, Accessed: Apr. 20



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY