Machine Learning-Based Transactions Anomaly Prediction for Enhanced IoT Blockchain Network Security and Performance

Nor Fadzilah Abdullah[1,2*], Ammar Riadh Kairaldeen[1], Asma Abu-Samah[1,2], and Rosdiadee Nordin[3]

[1] Department of Electrical, Electronic, and Systems Engineering, Faculty of Engineering & Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

[2] Wireless Research@UKM, Universiti Kebangsaan Malaysia 43600 UKM Bangi, Selangor, Malaysia

[3] Department of Engineering, School of Engineering and Technology, Sunway University, 47500 Bandar Sunway, Selangor, Malaysia

[e-mail: fadzilah.abdullah@ukm.edu.my, aaltotanje@gmail.com, asma@ukm.edu.my, rosdiadeen@sunway.edu.my]

[*]Corresponding author: Nor Fadzilah Abdullah

Received November 13, 2023; revised April 15, 2024; accepted June 30, 2024; published July 31, 2024

Abstract

The integration of blockchain technology with the rapid growth of Internet of Things (IoT) devices has enabled secure and decentralised data exchange. However, security vulnerabilities and performance limitations remain significant challenges in IoT blockchain networks. This work proposes a novel approach that combines transaction representation and machine learning techniques to address these challenges. Various clustering techniques, including k-means, DBSCAN, Gaussian Mixture Models (GMM), and Hierarchical clustering, were employed to effectively group unlabelled transaction data based on their intrinsic characteristics. Anomaly transaction prediction models based on classifiers were then developed using the labelled data. Performance metrics such as accuracy, precision, recall, and F1-measure were used to identify the minority class representing specious transactions or security threats. The classifiers were also evaluated on their performance using balanced and unbalanced data. Compared to unbalanced data, balanced data resulted in an overall average improvement of approximately 15.85% in accuracy, 88.76% in precision, 60% in recall, and 74.36% in F1-score. This demonstrates the effectiveness of each classifier as a robust classifier with consistently better predictive performance across various evaluation metrics. Moreover, the k-means and GMM clustering techniques outperformed other techniques in identifying security threats, underscoring the importance of appropriate feature selection and clustering methods. The findings have practical implications for reinforcing security and efficiency in real-world IoT blockchain networks, paving the way for future investigations and advancements.

Keywords: Anomaly prediction, Blockchain, Clustering algorithms, Fraud detection, Internet of Things (IoT), Legitimate transactions, Machine learning, Privacy, Security.

This research was supported in part by a research grant from the Malaysian Ministry of Higher Education [FRGS/1/2023/ICT08/UKM/02/1].

http://doi.org/10.3837/tiis.2024.07.014 **ISSN : 1976-7277**

1. Introduction

The rapid advancement of the Internet of Things (IoT) has revolutionised the interconnectedness and data-sharing capabilities of devices within a connected environment [1]. Equipped with sensors and software, IoT devices collect and transmit data for processing and analysis over the network [2]. Blockchain technology has also gained significant attention, with its decentralised and secure nature making it a promising platform for data exchange, decentralised identity management, and data provenance verification in a variety of domains [3], [4].

The main challenge in the implementation of IoT blockchain networks is the complex trade-off between security and performance. Most IoT devices have limited processing power, battery life and storage capacity. Validating the transactions from a massive amount of data generated by IoT devices can be slow and expensive. Trusting the decisions made by the system becomes difficult, particularly when it comes to making autonomous decisions based on data from IoT devices and information stored on the Blockchain. Additionally, performance is also critical, as IoT blockchain networks generate substantial data volumes that require efficient processing and storage capabilities. Scalability issues arise when the network experiences an influx of devices and transactions, potentially leading to delays and congestion [5]. Moreover, the choice of consensus algorithm directly affects transaction speed and overall performance specifically when focusing on integrating Blockchain into an IoT network [6]. Therefore, it is important to detect unexpected or anomalous transactions, so these can be isolated from further IoT blockchain network processing.

Extensive research is underway to improve the security and performance of IoT blockchain networks. This ongoing effort involves developing lightweight cryptographic protocols, efficient consensus algorithms and identity management systems, as well as data processing and storage optimisation techniques [7]. Furthermore, cutting-edge technologies such as machine learning and artificial intelligence are being explored to strengthen security measures, improve data integrity, detect anomalies, reduce security risks, and improve network performance [8].

One promising strategy for enhancing network performance is to utilise transaction representation techniques to reduce storage and computational requirements [9]. By intelligently encoding and structuring transactions, this can improve data processing speed, reduce latency, and increase scalability. Machine learning algorithms can analyse the vast amount of data generated by IoT devices to identify patterns and potential security threats, thereby strengthening the overall security of the blockchain network [10].

Although significant progress has been made in addressing security vulnerabilities or performance limitations in IoT blockchain networks, a crucial research gap remains in developing holistic solutions that simultaneously address both aspects. Therefore, an integrated solution that combines transaction representation techniques and machine learning algorithms is proposed to optimise data processing and storage, proactively identify security threats, and accurately detect anomalies. This approach helps to mitigate security vulnerabilities, enhance network performance, reduce costs, and improve overall efficiency, thereby facilitating the widespread adoption of IoT blockchain technology. A summary of the key contributions of this paper are outlined below:

1. **Integration of Blockchain IoT Environment and Machine Learning Techniques**: An IoT blockchain network was developed to integrate a transaction representation framework, machine learning (ML) clustering techniques, and prediction algorithms.

This proposed system model proactively detects and prevents malicious activities, identifies anomalies, and strengthens identity management systems. Based on the threat prediction output, user and data integrity checks are adopted only for normal transactions, while suspicious transactions will undergo further procedures for threat investigation. This ensures the integrity and confidentiality of data transmitted across the network.

2. **Performance Analysis and Experimental Evaluation**: The effectiveness and efficiency of the proposed solution were evaluated using real-world scenarios and two types of datasets (unbalanced and balanced). Furthermore, two datasets (unlabelled and labelled) were used to train and validate the system performance. The proposed approach can improve data processing speed, reduce latency, and enhance scalability within IoT blockchain networks with extensive data generated by IoT devices to identify patterns and potential security threats.
3. **Decision Support, Practical Implications and Future Directions**: The potential applications of the proposed solution across various domains were discussed, and future research directions to further enhance the security and performance of IoT blockchain networks were identified. Areas of exploration include advanced machine learning algorithms, large-scale deployment optimisation techniques, and robust consensus algorithms for IoT blockchain networks.

The rest of the paper is structured as follows: Section II provides a comprehensive literature review, summarising existing research and state-of-the-art approaches in IoT blockchain networks while highlighting challenges and limitations. Section III explains the proposed anomaly prediction system model, including the deployed dataset, experimental setup, and selection of machine learning algorithms for performance optimisation. Section IV presents a performance evaluation of the proposed system model based on the interpretation of the clustering and prediction analysis results. Finally, Section V concludes the research's key findings and contributions.

2. Related Works

The convergence of IoT and blockchain technologies has attracted substantial research attention due to their potential for enhancing security, privacy, and scalability in interconnected systems. Recently, machine learning (ML) has also emerged as a data-driven intelligent catalyst in this convergence for extracting meaningful patterns, identifying anomalies, predicting future outcomes and optimising processes.

Blockchain Internet of Things (BIoT) signifies a revolutionary era for secure, autonomous digital ecosystems. Blockchain technology introduces a layer of security and trust to the IoT, mitigating risks associated with data breaches and cyber-attacks by distributing data across a decentralized network [11]. This enhances scalability, allowing for a more robust infrastructure that can accommodate the exponential growth of IoT devices. Furthermore, BIoT enables the implementation of smart contracts, automating processes without the need for centralized authority, thereby reducing costs and increasing efficiency in operations like supply chain management. However, this integration is not without challenges. The resource constraints of IoT devices, such as limited processing power and energy, pose significant obstacles to the adoption of blockchain, which is computationally intensive [6]. Additionally, interoperability between different IoT platforms and blockchain systems remains a complex issue that requires standardization. An intriguing aspect worthy of discussion is the potential

of BIoT in creating fully autonomous and self-regulating systems, revolutionizing industries by making them more resilient, transparent, and efficient. This blend of technologies could pave the way for innovative applications, unlocking new avenues for sustainability and economic growth.

In the context of blockchains, ML algorithms have been proposed to enhance the security of IoT blockchain networks. Anomaly prediction methods, particularly those utilising deep learning models, can detect unusual patterns or behaviours in IoT data, indicating potential security risks. The authors in [12] presented a machine learning-based approach for anomaly prediction in IoT systems, leveraging blockchain for data integrity. Similarly, [13] proposed a blockchain-based framework that utilises machine learning algorithms to detect and prevent cyber-attacks in IoT networks. Research conducted by [14] demonstrates the effectiveness of machine learning in identifying anomalies in IoT blockchain networks. In [15], machine learning techniques for anomaly prediction using support vector machines (SVM) and neural networks have been explored to identify potential security threats in IoT blockchain networks.

Privacy and identity management of IoT blockchain has also received significant attention. A blockchain-based framework for secure and decentralised identity management of IoT devices to protect user privacy and ensure reliable authentication has been proposed [16], [17]. Authors in [18] proposed a privacy-preserving data-sharing scheme using blockchain, enabling secure and controlled data sharing among IoT devices. The zero-knowledge proofs, such as zk-SNARKs (zero-knowledge succinct non-interactive arguments of knowledge) [19], were proposed for transaction verification without revealing transaction details. This approach ensures data privacy and integrity while maintaining the transparency of a blockchain.

Blockchain transaction delves into sectors where it secures and validates data from IoT devices, revolutionizing processes from supply chain logistics to urban management, critical in industries where product authenticity and safety are paramount. The transaction terminology does not strictly refer to a financial exchange. It can be any transfer of data or information that is recorded on the IoT blockchain. In the context of IoT, a transaction could be the exchange of data between devices, such as temperature readings from sensors, GPS locations from trackers, or status updates from smart devices [20]. These transactions in all cases have numerical values that are secure, benefiting from blockchain's inherent properties like decentralization, transparency, and immutability.

Scalability is another critical aspect of IoT blockchain networks. In [21], the scalability of blockchain systems was reviewed, and various techniques to improve scalability were highlighted. Additionally, [5] investigated the integration of cloud computing and IoT, uncovering potential synergies between these technologies to address scalability concerns. Another approach involves integrating off-chain solutions to mitigate scalability limitations. The Lightning Network, introduced by [22], enables the execution of microtransactions off the main blockchain, enhancing throughput and reducing transaction fees. Leveraging payment channels and smart contracts provide a scalable and efficient solution for IoT blockchain networks.

Sharding techniques have emerged as a promising solution in IoT blockchain networks to improve performance [23], [24]. It involves dividing the blockchain into smaller partitions, known as shards, enabling concurrent and shared processing of transactions to achieve higher throughput and enhanced network scalability.

Advancements have been made in addressing security and performance issues in IoT blockchain networks. The use of consensus algorithms, such as the Proof-of-Stake (PoS), has been proposed by [25] to improve the scalability and energy efficiency of blockchain networks. The PoS algorithm achieves consensus by allocating mining rights based on participants' stake

in the network, reducing computational overhead and enabling faster transaction processing. The results show that PoS performs better than traditional Proof-of-Work (PoW) algorithms.

The existing research in the integration of ML, IoT and blockchain technologies has made significant contributions. The study in [28] proposes using ML for generating secure side chains in IoT-based blockchain systems, enhancing security without excessive computational complexity, and making it suitable for real-time applications. The authors in [29] propose enhancing data privacy in IoT-enabled blockchain through machine learning algorithms, ensuring secure data transactions in telehealth supply chains with improved security levels. The ML model processes the data management based on the number of transactions, time of transaction, and transaction confirmation time. A smart city intrusion detection system is proposed in [30] using an integration of ML and decentralized blockchain architecture to secure the fog computing layer vulnerability. A blockchain-based solution for secure and private IoT in smart homes, utilizing a Deep Extreme Learning Machine (DELM) was proposed in [31] by carefully evaluating the network reliability against security privacy, integrity, and accessibility goals. Transaction representation allows for the proactive identification of security vulnerabilities and the implementation of robust security measures. The work conducted by [32] showcased the effectiveness of transaction representation techniques in detecting fraud in financial systems. Similarly, the research undertaken by [25] demonstrated the potential of machine learning algorithms in optimising resource allocation and improving performance in IoT networks.

However, these works warrant further critical analysis of the current research. While machine learning-based approaches have shown promise in detecting anomalies and preventing cyber-attacks in IoT networks, there is a lack of performance evaluation of the algorithms using realistic datasets. Therefore, it is crucial to assess the performance of these algorithms in detecting security threats while considering the computational and resource requirements of using machine learning implementation in resource-constrained IoT devices. **Table 1** summarises the focus and limitations of previous works related to ML adoption in IoT blockchain networks. The approach presented in this paper extends our previous works in [26] and [27] by adding transaction representation, machine learning, and proactive security measures to tackle the security, performance, scalability, and cost-effectiveness challenges in IoT blockchain networks.

Table 1. Summary of Previous Works on ML-based IoT Blockchain Network

Source	Focus	Limitation
[28]	Proposed ML model for calculation for the age of side chains in IoT blockchain network, with high-security execution but generally less computational complexity, suitable for near real-time (NRT) applications.	Does not consider any anomaly detection in the proposed algorithm.
[29]	Enhancing data privacy in IoT-enabled blockchain through ML algorithms, ensuring secure data transactions in telehealth supply chains with improved security levels.	Scalable local ledgers compromise on peer validation of transactions.
[30]	Designing a smart city intrusion detection system architecture using machine learning and blockchain technology to secure the fog computing layer vulnerability.	Lack of sufficient training real dataset for labelling and validation of the model.

Source	Focus	Limitation
[31]	Introducing blockchain-based solution for secure and private IoT smart homes, utilizing Deep Extreme Learning Machine (DELM).	Highly demanding computational and time requirements, with marginal overhead creation.

3. System Model

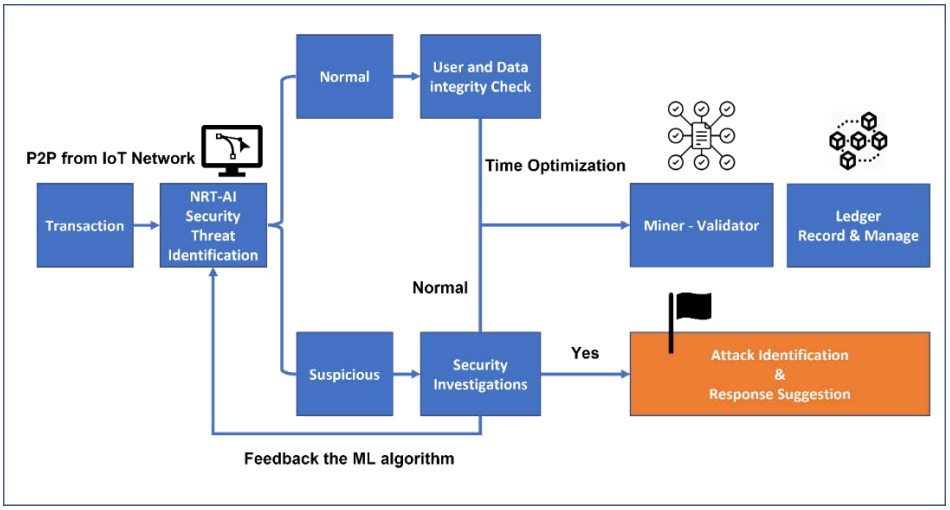

Fig. 1. System architecture for transactions anomaly prediction in IoT blockchains.

This work proposed a system model based on the design goal of a threat prediction scheme as shown in **Fig. 1**. Smart contracts on IoT blockchain networks automate digital agreements, in which the terms are coded into the contract and automatically executed when specified conditions are met during security investigations. The contracts ensure transparency and immutability, preventing after-deployment alterations to a decentralised network. Consensus protocols within the blockchain verify and validate execution, reinforcing security. With their automation and security features, smart contracts streamline complex interactions between IoT devices, providing efficient, tamper-proof management in networks ranging from IoT smart homes to large-scale industrial systems.

Transaction representation is vital in fortifying security by capturing and presenting transactional data in a structured format. The format allows more efficient identification, analysis, and detection of patterns, anomalies, and potential security threats. Techniques such as data normalisation, feature extraction, and data aggregation contribute to a comprehensive comprehension of transactional behaviour within the network. This entails analysing transactional data derived from real-world IoT blockchain networks based on their properties. Near real-time (NRT) processing is preferred over real-time processing in using Artificial Intelligence (AI) to classify transactions in a blockchain network smart contracts. In the proposed system model, the NRT-AI security threat identifications are based on the adoption of machine learning algorithms based on clustering and classifications, as explained in the following subsections.

Building on the anomaly prediction architecture in **Fig. 1**, the threat identification model is regularly updated into the smart contract, as shown in **Fig. 1**. Based on the threat prediction output, user and data integrity checks are adopted only for normal transactions. Whereas the suspicious transactions will need to undergo further procedures for threat investigation. The problem lies in the potential vulnerabilities and threats that compromise user and data integrity, leading to the manipulation, corruption, or unauthorized access of sensitive information.

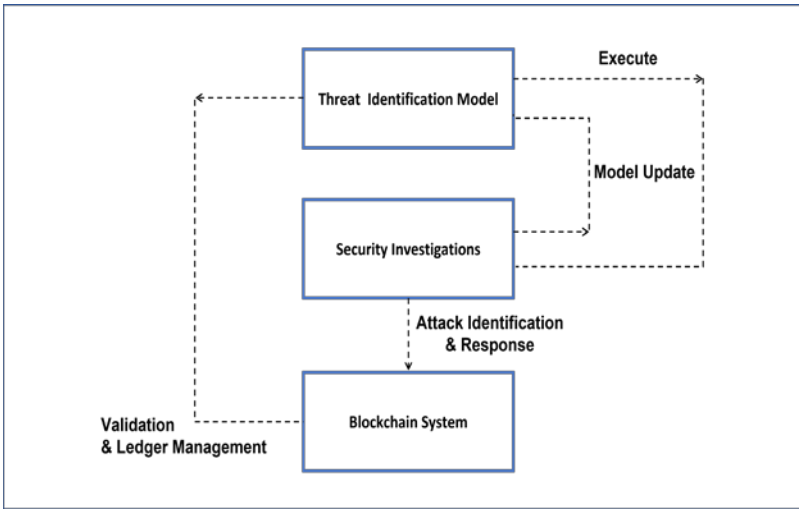

Fig. 2. Update of threat identification model in blockchain smart contracts

Two phases are needed for training and testing the security threat identification model, as shown in **Fig. 3**. Multiple dataset sources are utilised to ensure a diverse and representative sample. Phase 1 involves unlabelled and labelled datasets. Meanwhile, Phase 2 only uses the labelled dataset. Further details on each phase and dataset are described in the following subsections.

3.1 Clustering Algorithms (Phase 1)

Data clustering techniques are used to generate training data for the prediction models. By grouping transactions that exhibit similarities, labels can be assigned to the training dataset, indicating the specific cluster or group to which each transaction belonged, namely Normal or Anomaly.

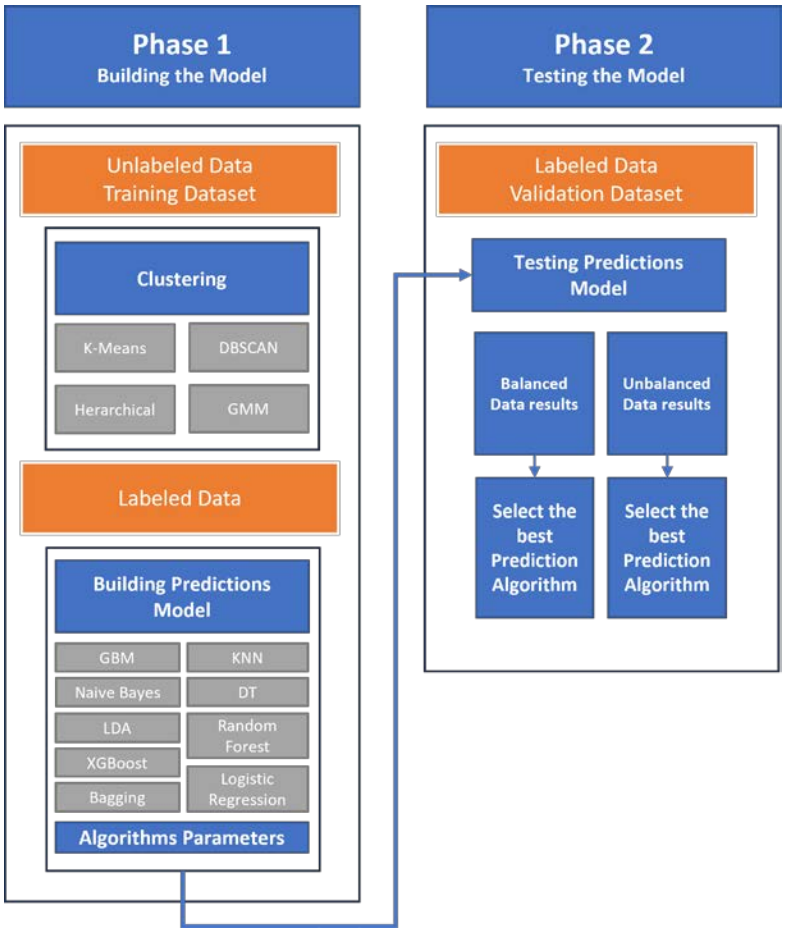

Fig. 3. Phases for Security Threat Identification Model

Several clustering algorithms were selected and trained, considering their capability to analyse transactional data, accommodate diverse datasets, handle noise and outliers, and provide valuable insights into transactional patterns within IoT blockchain networks. Specifically, k-means clustering is chosen for its simplicity and efficiency in grouping data points and hierarchical clustering for capturing the hierarchical structure of data and enabling the exploration of relationships among transactions. DBSCAN is valuable for identifying both hidden patterns and anomalies. Further descriptions of each algorithm are as follows:

1. **k-means Clustering**: k-means is an extensively used unsupervised learning algorithm that aims to partition data points into k clusters based on their similarity [33]. It iteratively assigns data points to the nearest cluster centroid and updates centroid positions until convergence.
2. **Hierarchical Clustering**: Hierarchical clustering is another unsupervised learning algorithm that constructs a hierarchy of clusters by iteratively merging or splitting clusters based on their similarity [34]. It can reveal the hierarchical structure of the data, which can effectively uncover relationships and dependencies among transactions in the IoT blockchain network.
3. **DBSCAN**: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that groups data points based on their density and connectivity [35]. It identifies core points that have enough neighbouring points

within a specified radius and expands clusters by connecting density-reachable points. DBSCAN is proficient in detecting clusters of various shapes and handling noisy data, making it suitable for analysing transactional patterns in the IoT blockchain network.

4. **GMM**: The Gaussian Mixture model (GMM) is a probabilistic model that assumes data points are generated from a mixture of Gaussian distributions [36]. It estimates the parameters of these distributions to model the underlying data distribution. GMM can be applied to identify hidden patterns and anomalies within the transactional data of the IoT blockchain network.

A total of 30,505 Ethereum blockchain transactions [37] were utilised for the ML clustering model. Data cleaning and preprocessing techniques were applied to the original dataset to ensure seamless analysis and compatibility. The Principal Component Analysis (PCA) [38] technique was employed as a preprocessing step while preserving as much information as possible to reduce dimensionality. This is achieved by transforming the original features into orthogonal components called 'principal components'. These components are ordered based on their ability to capture the maximum variance in the data, with the first component capturing the most variance, followed by subsequent components in decreasing order. For k-means clustering, PCA is usually used as a preprocessing step.

Furthermore, a data engineering technique based on feature engineering was applied to this dataset. Feature engineering is the process of creating new features (attributes) from existing data to provide additional insights and improve the performance of machine learning models or analytical tasks. This technique generated two new features: (i) 'Transaction_Frequency_To' and (ii) 'Transaction_Frequency_From'. These features are created to analyse and gain insights into both the recipient's and sender's transaction patterns and behaviours within the IoT blockchain network. **Table 2** shows the attributes of the primary dataset.

Table 2. Training Dataset attributes features description.

Data attributes	Description
Block_no	Series number in the chain of Blocks
Transaction_hash	Block hash value
Transaction_From	Sender transaction
Transaction_To	Recipients transaction
TxnFee	Captures transaction costs
Value_OUT	Transaction magnitude
Transaction_Frequency_From	Sender transaction patterns and behaviours
Transaction_Frequency_To	Recipient transaction patterns and behaviours

3.2 Anomaly Prediction (Phase 2.1)

Various classifiers have been proposed as predictive models to address security threat identification in diverse domains. For example, the Gradient Boosting algorithms ensemble learning technique iteratively combines weak learners to build a strong predictive model. It aims to minimise a loss function by sequentially g models focusing on misclassified instances. XGBoost is an optimised implementation of the Gradient Boosting algorithm with enhancements for improved speed and accuracy. It has exhibited effective security threat prediction capabilities across various applications [39] by analysing multiple features extracted from transactional data. Ensemble methods such as Random Forest and AdaBoost

also offered promising outcomes [40]. These models leverage multiple decision trees to make accurate predictions based on patterns observed in transactional data. A well-established machine learning framework such as 'scikit-learn' is employed to implement the predictive models. The predictive models will be trained using a subset of the dataset, validated using cross-validation techniques, and assessed using appropriate performance metrics such as accuracy, precision, recall, and F1 score.

Integrating prediction models into the system architecture will enable continuous real-time transactional data analysis. Appropriate security measures will be activated if any security threats or anomalies are detected, such as alert notifications, access restrictions, or transaction rejection. Through the system model, nine classifier methods stated in **Table 3** are compared. The model covers diverse approaches, including linear models, ensemble methods, distance-based algorithms, and probabilistic classifiers to capture different aspects of the data and leverage their unique strengths for accurate security threat identification and performance optimisation.

Table 3. Prediction Models Classifier Description

No.	Classifier	Description
1	Logistic Regression (LR)	Establishes a relationship between a dependent binary variable and independent variables using a logistic function.
2	Random Forest (RF)	An ensemble learning method that combines multiple decision trees to make predictions. Each tree is trained on a subset of data, and the final prediction is made by aggregating individual tree predictions.
3	K-Nearest Neighbours (KNN)	Classifies new instances based on the majority vote of their k closest neighbours in the feature space. A popular algorithm for pattern recognition and machine learning.
4	Naive Bayes	Assumes conditional independence of features given the class label, often used in text classification and other domains. Despite its simple assumptions, it has shown promising results in text classification and other domains.
5	Gradient Boosting Machines (GBM)	Generalisation of Gradient Boosting for regression and classification problems using gradient descent.
6	AdaBoost	Ensemble method that combines multiple weak classifiers to create a strong classifier by weighting misclassified samples. Each weak classifier is trained on weighted data, giving more emphasis to misclassified samples in subsequent classifiers.
7	Decision Trees (DT)	Tree-based classifiers that partition the feature space based on attribute values for interpretability and versatility. Internal nodes represent attribute tests, while leaf nodes represent class labels. DT is widely used due to their interpretability and ability to handle different data types.
8	Linear Discriminant Analysis (LDA)	Statistical method for dimensionality reduction and classification, aiming to maximise class separation by assuming Gaussian distributions and estimating class priors, means, and covariances.
9	Bagging	Ensemble technique that builds multiple models by resampling training data with replacement. Each model is trained on a different bootstrap sample, and predictions are aggregated for the result.

A secondary dataset of 9,841 blockchain transactions [41] that contains labels of both fraudulent and valid transactions is used for anomaly prediction. This provides a realistic and challenging scenario for evaluating the effectiveness of the proposed security threat model. Upon data cleaning of this secondary dataset, a data engineering technique is used to derive a new feature, known as 'TxnFee, by calculating the difference between the 'total Ether sent' and 'total Ether received' features. This new feature helps to capture the economic dimension of transactions, contributing to a more comprehensive understanding of the dataset and enhancing the effectiveness of our fraud detection and prevention models. **Table 4** shows the attributes of the secondary dataset.

Table 4. Validation Dataset attributes features description.

Data attributes	Description	Training Dataset Matching
Total_Transactions	Series number in the chain of Blocks	Block_no
Transaction_hash	Block hash value	Transaction_hash
Unique_Received_From_Addresses	Sender transaction	Transaction_From
Unique_Sent_To_Addresses	Recipient transaction	Transaction_To
Total_Ether_Received	Captures transaction costs	TxnFee
Avg_Value_Received	Transaction magnitude	Value_OUT
Total_Ether_Sent_Contracts	Sender transaction patterns and behaviours	Transaction_Frequency_From
Total_Ether_Received_Contracts	Recipient transaction patterns and behaviours	Transaction_Frequency_To

3.3 Model Testing and Implementation (Phase 2.2)

The model testing and implementation are conducted using two cases: (i) unbalanced data, referring to the original source dataset, and (ii) balanced data, which is based on the modified dataset.

The unbalanced data is based on the full original dataset, where it has been seen that the number of Cluster 0 transactions is significantly larger than Cluster 1. For a balanced data anomaly prediction analysis, a method called ‘Random Over-Sampling’ was employed. This method aims to mitigate the unbalanced dataset source by increasing the number of instances in the minority class through a random selection duplication process. As a result, classes have an equal representation, allowing the machine learning algorithms to learn more effectively from the minority class data.

4. Performance Evaluation

4.1 Clustering Results and Analysis

Initially, the Elbow method [42] is applied to determine the optimal number of clusters in a dataset. As the WCSS (within-cluster sum of squares) measures cluster tightness, the “elbow” point in **Fig. 4** shows that the optimal cluster count is two, referenced as (i) Cluster 0 and (ii) Cluster 1 in the following results. A majority of the blockchain transactions are in Cluster 0, which represents normal transaction behaviour. Meanwhile, the smaller size of Cluster 1 indicates that these transactions exhibit different characteristics from the majority of blockchain transactions, which indicates anomaly transactions.

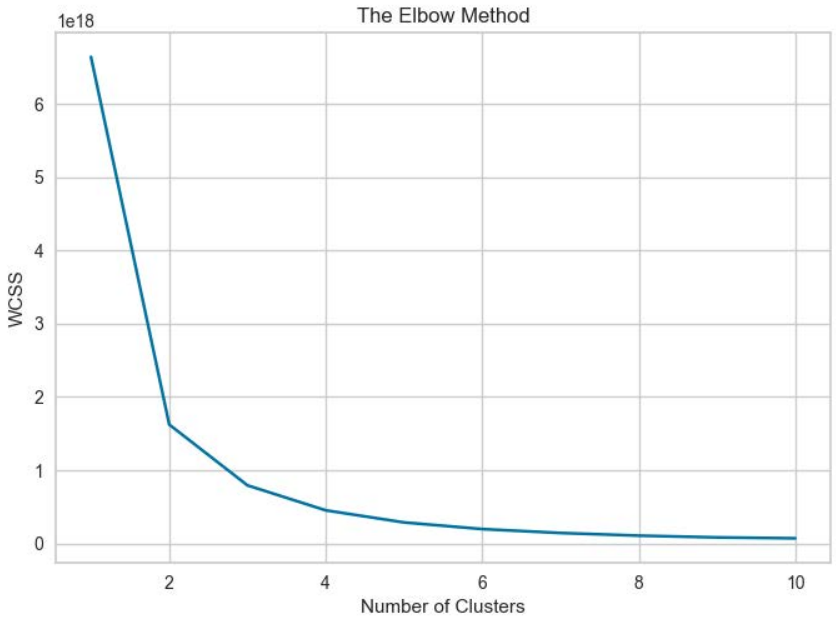

Fig. 4. Elbow method to determine cluster size

Additionally, the Silhouette Score [43] was used as a criterion to assess the quality of the clustering results using different numbers of clusters, and the Silhouette Score was computed for each configuration. To accomplish this, the 'Silhouette' considers both the average distance between samples within the same cluster (cohesion) and the average distance between samples of different clusters (separation). The Silhouette Score ranges from -1 to 1, with higher values indicating well-defined and well-separated clusters. **Fig. 5** shows the Silhouette measurement results for the selected clustering algorithms specified in Section 3. where the x-axis in a Silhouette Score graph represents Silhouette Scores for diverse cluster configurations, while the y-axis denotes the corresponding number of clusters, aiding in the identification of an optimal cluster count based on the maximal Silhouette Score.

Additionally, **Fig. 6** illustrates the Cluster 0 (Normal) and Cluster 1 (Anomaly) distribution for the four clustering algorithms. K-mean clustering achieves a Silhouette score of 0.9835 with 30,504 data points in Cluster 0 and 199 data points in Cluster 1. Hierarchical Clustering has a Silhouette score of 0.9823 with 30,312 data points in Cluster 0 and 181 data points in Cluster 1. DBSCAN clustering has a Silhouette score of 0.9462 containing 30,664 data points in Cluster 0 and 39 data points in Cluster 1. Meanwhile, GMM clustering has a score of 0.9836 with 30,514 data points in Cluster 0 and 189 data points in Cluster 1.

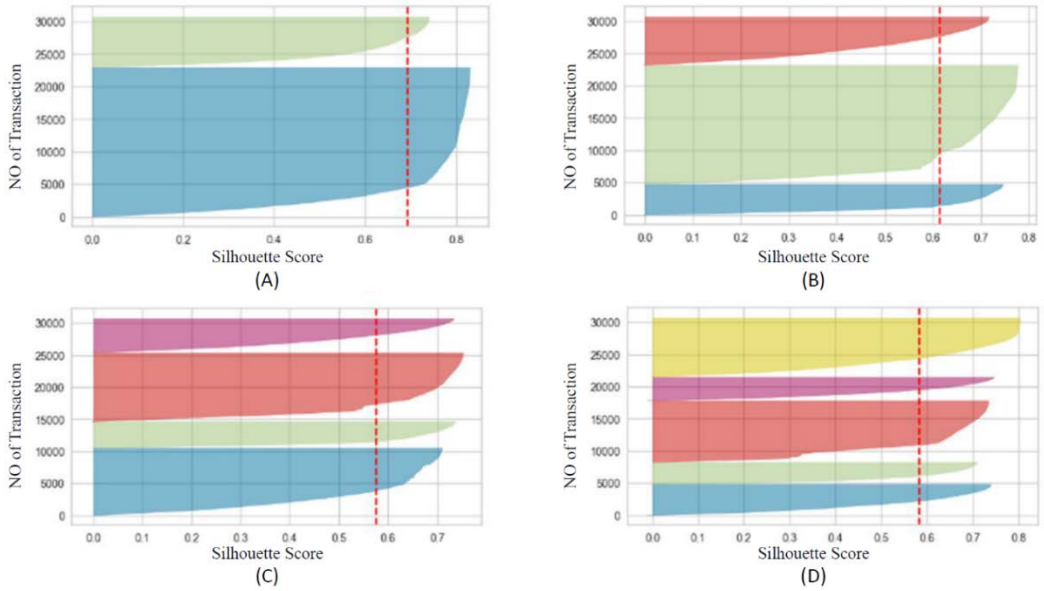

Fig. 5. Silhouette Measurement for (A) 2 clusters, (B) 3 clusters, (C) 4 clusters and (D) 5 clusters.

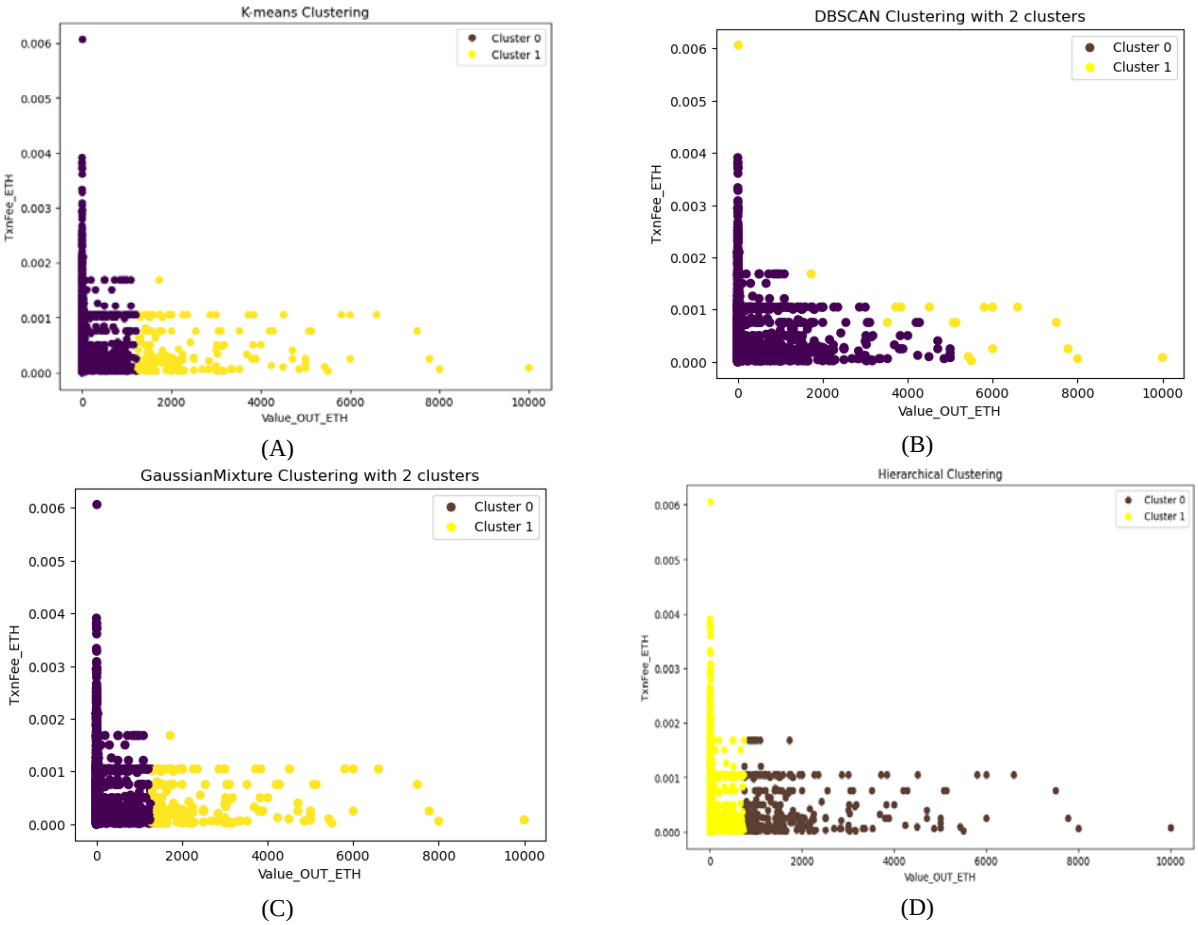

Fig. 6. Distribution of Cluster 0 (normal transactions) and Cluster 1 (anomaly transactions) for different ML clustering algorithms.

The summary results of the four clustering algorithms are presented in **Table 5**. It can be seen that all algorithms have relatively high Silhouette Scores, with GMM having the highest value. A score of 0.98 suggests that the data points within each cluster are similar to each other and well-separated from the data points in the other cluster. This indicates that the clustering algorithm successfully grouped the data points into distinct clusters based on their features or characteristics. In other words, the proposed ML-based clustering method successfully obtained a cluster that is reasonably similar to each other and well-separated from the data points in the other cluster.

Table 5. Summary of ML Clustering Algorithms results

Clustering algorithm	Normal Data Points (Cluster 0)	Suspicious Data Points (Cluster 1)	Silhouette Score
k-means	30,504	199	0.9835
Hierarchical	30,312	181	0.9823
DBSCAN	30,664	39	0.9462
Gaussian Mixture Models (GMM)	30,514	189	0.9836

Based on the clustering method, several key findings are summarised as follows:

1. **Distribution of Clusters**: The results show that all clustering algorithms were able to identify two distinct clusters labelled as Cluster 0 and Cluster 1. The distribution of data points across these clusters varied among the algorithms. Notably, in the k-means and GMM clustering methods, Cluster 0 contains a significantly larger number of data points compared to Cluster 1, indicating an imbalance in cluster sizes. In contrast, the Hierarchical and DBSCAN clustering methods show a relatively smaller difference in cluster sizes.
2. **Separation of Clusters**: The Silhouette scores indicate the separation and compactness of the clusters. Higher Silhouette scores closer to 1 suggest well-separated clusters with high similarity within each cluster. In this case, all four clustering algorithms exhibit high Silhouette scores, ranging from 0.9462 to 0.9836. This indicates that the clusters obtained by each algorithm are distinct and well-separated, reflecting the effectiveness of the algorithms in capturing the underlying patterns and structures within the dataset.
3. **Imbalance in Cluster Size**: The clustering results reveal an imbalance in cluster sizes, particularly in the k-means and GMM clustering methods, where Cluster 0 contains a significantly larger number of data points compared to Cluster 1. This suggests that the majority of transactions or data points in the dataset exhibit similarities and form a larger cluster, while a smaller cluster represents a distinct subset of transactions or data points.
4. **Identification of Malicious Transactions**: Clustering techniques can be utilised for identifying potentially malicious transactions or outlier data points. By examining the composition of the smaller clusters, such as Cluster 1 in this case, it is possible to identify transactions that deviate from the majority and may require further investigation as potential anomalies or fraudulent activities. The smaller cluster size indicates that these transactions are less common or exhibit different characteristics from the majority of transactions.
5. **Algorithm Selection**: The choice of clustering algorithm depends on the specific requirements and characteristics of the dataset. In this case, all four algorithms,

namely k-means, Hierarchical, DBSCAN, and GMM clustering, demonstrate good clustering results with high Silhouette scores.

6. **Potential Applications**: The clustering results have implications for various applications. In the context of IoT blockchain networks, the identified clusters can provide insights into the grouping of transactions based on their characteristics, aiding in anomaly prediction, fraud identification, and security threat mitigation. The results can also inform the design and optimisation of network protocols, resource allocation, and performance enhancements based on the characteristics of different clusters.

Overall, the clustering results highlight the effectiveness of the applied algorithms in identifying distinct clusters, capturing patterns within the dataset, and enabling further analysis for various applications related to security, performance, and optimisation in IoT blockchain networks.

4.2 Prediction Results and Analysis

For the anomaly prediction performance evaluation, the following common ML performance metrics were used:

- Accuracy measures the overall correctness of a classification model by calculating the ratio of correctly predicted instances to the total number of instances [44].

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \dots (1)$$

- Precision assesses the model's ability to correctly identify positive instances out of the total instances predicted as positive [45].

$$\textbf{Precision} = \frac{TP}{TP+FP} \dots (2)$$

- Recall measures the model's ability to correctly identify positive instances out of the total actual positive instances [46].

$$\textbf{Recall} = \frac{TP}{TP+FN} \dots (3)$$

- F1-measure [47], represents a harmonic mean of precision, and recall provides a balanced evaluation of the model's performance. These evaluation metrics have been widely used and studied in the field of machine learning.

$$\textbf{F1-measure} = \frac{2*(Precision*Recall)}{(Precision+Recall)} \dots (4)$$

4.2.1 Prediction from Unbalanced Data

The performance evaluation of classifiers and clustering methods for the unbalanced data, as described in Section 3.3, is shown in **Table 6** until **Table 9**, where the highest performance metrics are highlighted for each method.

For k-means and GMM clustering, the Naive Bayes classifier stood out with a higher accuracy of 0.8330, indicating better overall performance compared to the other models. Other classifiers demonstrated a lower accuracy score implying that these models struggle to effectively differentiate between positive and negative classifications. On the other hand,

hierarchical-based and DBSCAN prediction models exhibited varying performance. A mixture of LDA, LR and Naïve Bayes has the highest performance for hierarchical clustering, while KNN, Naïve Bayes and GBM for DBSCAN clustering.

Table 6. k-means clustering prediction model results for the Unbalanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.6957	0.01	0.01	0.01
2	Random Forest (RF)	0.6957	0.01	0.01	0.01
3	K-Nearest Neighbours (KNN)	0.6957	0.01	0.01	0.01
4	Naïve Bayes	0.8267	0.89	0.25	0.39
5	Gradient Boosting Machines (GBM)	0.6957	0.01	0.01	0.01
6	AdaBoost	0.6957	0.01	0.01	0.01
7	Decision Tree (DT)	0.6957	0.01	0.01	0.01
8	LDA	0.5904	0.10	0.10	0.10
9	Bagging	0.6957	0.01	0.01	0.01

Table 7. Hierarchical clustering prediction model results for the Unbalanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.3060	0.24	1.00	0.39
2	Random Forest (RF)	0.3173	0.24	0.99	0.39
3	K-Nearest Neighbours (KNN)	0.3173	0.24	0.99	0.39
4	Naïve Bayes	0.5214	0.31	0.95	0.47
5	Gradient Boosting Machines (GBM)	0.3173	0.24	0.99	0.39
6	AdaBoost	0.3173	0.24	0.99	0.39
7	Decision Tree (DT)	0.3173	0.24	0.99	0.39
8	LDA	0.7416	0.41	0.38	0.39
9	Bagging	0.3171	0.24	0.99	0.39

Table 8. DBSCAN clustering prediction model results for the Unbalanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.7443	0.09	0.02	0.03
2	Random Forest (RF)	0.7438	0.08	0.02	0.03
3	K-Nearest Neighbours (KNN)	0.7511	0.07	0.01	0.02
4	Naïve Bayes	0.6713	0.32	0.44	0.37
5	Gradient Boosting Machines (GBM)	0.2548	0.17	0.63	0.27
6	AdaBoost	0.6986	0.04	0.02	0.02
7	Decision Tree (DT)	0.7444	0.08	0.02	0.03
8	LDA	0.2573	0.17	0.62	0.27
9	Bagging	0.3171	0.24	0.99	0.39

Table 9. GMM clustering prediction model results for the Unbalanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.6962	0.01	0.01	0.01
2	Random Forest (RF)	0.6962	0.01	0.01	0.01
3	K-Nearest Neighbours (KNN)	0.6962	0.01	0.01	0.01
4	Naïve Bayes	0.8267	0.89	0.25	0.39
5	Gradient Boosting Machines (GBM)	0.7652	0.01	0.00	0.00
6	AdaBoost	0.6962	0.01	0.01	0.01
7	Decision Tree (DT)	0.6962	0.01	0.01	0.01
8	LDA	0.3696	0.12	0.29	0.17
9	Bagging	0.6962	0.01	0.01	0.01

However, for this unbalanced dataset focusing on a minority class, the **Recall** metric is prioritised because it measures the proportion of correctly identified anomalies among all the actual anomalies, ensuring effective detection of the minority class. Accuracy can be misleading due to class imbalance, and Precision may focus on false positives, potentially missing actual anomalies. F1 score offers a fair evaluation, but Recall remains most suitable for accurate prediction of anomalies [47]. Thus, the most optimal method for unbalanced data is the combination of hierarchical clustering and logical regression classifier combination, as shown in **Table 10**.

Table 10. Summary of optimal classifiers and clustering methods for Unbalanced data

Unbalanced Data	Accuracy	Precision	Recall	F1-measure
k-means clustering	0.8267 (Naïve Bayes)	0.89 (Naïve Bayes)	0.25 (Naïve Bayes)	0.39 (Naïve Bayes)
Hierarchical clustering	0.7416 (LDA)	0.41 (LDA)	1.00 (LR)	0.47 (Naïve Bayes)
DBSCAN clustering	0.7511 (KNN)	0.32 (Naïve Bayes)	0.63 (GBM)	0.37 (Naïve Bayes)
GMM clustering	0.8267 (Naïve Bayes)	0.89 (Naïve Bayes)	0.29 (LDA)	0.39 (Naïve Bayes)

4.2.2 Prediction from Balanced Data

Next, the performance evaluation of classifiers and clustering methods for the modified balanced data, as described in Section 3.3, is shown in **Table 11** until **Table 14**, where similarly, the highest performance metrics are highlighted for each method. Similar to unbalanced data, Naïve Bayes shows the best performance for k-means and GMM clustering. Meanwhile, LDA and RF give the highest performance for hierarchical clustering, while KNN and Naïve Bayes for DBSCAN clustering.

Table 11. k-means clustering prediction model results for the Balanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.6957	0.01	0.01	0.01
2	Random Forest (RF)	0.6957	0.01	0.01	0.01
3	K-Nearest Neighbours (KNN)	0.6957	0.01	0.01	0.01
4	Naïve Bayes	0.8330	1.00	0.25	0.40
5	Gradient Boosting Machines (GBM)	0.6957	0.01	0.01	0.01
6	AdaBoost	0.6957	0.01	0.01	0.01
7	Decision Tree (DT)	0.6957	0.01	0.01	0.01
8	LDA	0.5904	0.10	0.10	0.10
9	Bagging	0.6957	0.01	0.01	0.01

Table 12. Hierarchical clustering prediction model results for the Balanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.7412	0.41	0.38	0.39
2	Random Forest (RF)	0.3173	0.24	0.99	0.39
3	K-Nearest Neighbours (KNN)	0.6962	0.01	0.01	0.01
4	Naïve Bayes	0.1669	0.18	0.75	0.29
5	Gradient Boosting Machines (GBM)	0.3173	0.24	0.99	0.39
6	AdaBoost	0.3173	0.24	0.99	0.39
7	Decision Tree (DT)	0.3173	0.24	0.99	0.39
8	LDA	0.7436	0.41	0.37	0.39
9	Bagging	0.3173	0.24	0.99	0.39

Table 13. DBSCAN clustering prediction model results for the Balanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.2572	0.17	0.63	0.27
2	Random Forest (RF)	0.3469	0.16	0.48	0.25
3	K-Nearest Neighbours (KNN)	0.7489	0.07	0.01	0.02
4	Naïve Bayes	0.1669	0.18	0.75	0.29
5	Gradient Boosting Machines (GBM)	0.2548	0.17	0.63	0.27
6	AdaBoost	0.2548	0.17	0.63	0.27
7	Decision Tree (DT)	0.3032	0.18	0.63	0.28
8	LDA	0.2573	0.17	0.62	0.27
9	Bagging	0.2554	0.17	0.63	0.27

Table 14. GMM clustering prediction model results for the Balanced data

No.	Classifiers	Accuracy	Precision	Recall	F1-measure
1	Logistic Regression (LR)	0.7060	0.01	0.00	0.00
2	Random Forest (RF)	0.6962	0.01	0.01	0.01
3	K-Nearest Neighbours (KNN)	0.6962	0.01	0.01	0.01
4	Naïve Bayes	0.8330	1.00	0.25	0.40
5	Gradient Boosting Machines (GBM)	0.6962	0.01	0.01	0.01
6	AdaBoost	0.6962	0.01	0.01	0.01
7	Decision Tree (DT)	0.6962	0.01	0.01	0.01
8	LDA	0.2569	0.17	0.63	0.27
9	Bagging	0.6962	0.01	0.01	0.01

Table 15 summarises the optimal combination of classifiers and clustering methods for balanced data. It can be seen that the GMM clustering method paired with the Naïve Bayes classifier is a favourable choice, for overall ML performance metrics. In contrast to the unbalanced data, Recall is no longer a dominant performance criterion for the balanced data.

Table 15. Summary of optimal classifiers and clustering methods for Balanced data

Unbalanced Data	Accuracy	Precision	Recall	F1-measure
k-means clustering	0.8330 (Naïve Bayes)	1.00 (Naïve Bayes)	0.25 (Naïve Bayes)	0.40 (Naïve Bayes)
Hierarchical clustering	0.7436 (LDA)	0.41 (LDA)	0.99 (RF)	0.29 (RF)
DBSCAN clustering	0.7489 (KNN)	0.18 (Naïve Bayes)	0.75 (Naïve Bayes)	0.29 (Naïve Bayes)
GMM clustering	0.8330 (Naïve Bayes)	1.00 (Naïve Bayes)	0.63 (LDA)	0.40 (Naïve Bayes)

4.2.3 Overall Observations

The clustering results exhibited the successful grouping of data points into distinct clusters based on their distinctive attributes and characteristics. However, discernible disparities surfaced in terms of cluster distribution, inter-cluster separation, and cluster size imbalance. With respect to cluster distribution, both k-means and Gaussian Mixture clustering manifested seamless performances, with the majority of data points being assigned to Cluster 0. In contrast, DBSCAN clustering yielded a more balanced distribution, wherein a relatively smaller number of data points were allocated to Cluster 1. Hierarchical clustering likewise demonstrated a balanced distribution, albeit with a larger number of data points encompassed within Cluster 1. These findings imply that DBSCAN clustering and Hierarchical clustering may be better suited for scenarios necessitating balanced cluster sizes.

Moreover, based on the achieved anomaly prediction results, **Table 16** summarises the preferred methods for precision and recall in both balanced and unbalanced data scenarios.

Table 16. Summary of preferred prediction algorithms with different datasets

Data Scenario	Preferred for Precision	Preferred for Recall
Balanced Data	Naïve Bayes	Linear Discriminant Analysis
Unbalanced Data	Naïve Bayes	Logistic Regression

According to the findings, the Naïve Bayes method consistently stands out in terms of precision, regardless of the data scenario. It consistently achieves higher precision values compared to other methods. For the Recall metric, Logistic Regression (LR) offers the best performance for unbalanced data, while LDA gives the best performance for balanced data scenarios of the proposed system model. LR estimates conditional probabilities, which makes it particularly suitable for identifying rare events that are suitable for unbalanced data.

The consistent superiority of Naïve Bayes for both unbalanced and balanced data scenarios is due to its probabilistic approach and feature independence assumption. Naïve Bayes utilises Bayes' theorem to estimate the probability of a given class label based on the observed features. This method is effective even with limited training data and is resilient against irrelevant features. Therefore, it can effectively identify patterns and characteristics of fraudulent activities or abnormal behaviour, resulting in higher precision in detecting positive instances.

Based on the overall results analysis, it is recommended to utilise balanced data when building prediction models to enhance network security and performance in IoT blockchain networks. Balanced data allows the models to learn from a more representative distribution of positive and negative classes, leading to better generalisation and improved performance. Although working with balanced data may require additional efforts in data preprocessing and balancing techniques, the resulting models are more likely to provide reliable predictions and be more effective in real-world scenarios.

5. Conclusion

In this study, we proposed an innovative approach that leverages advanced techniques in data analysis to bolster security measures and optimise performance in IoT blockchain networks. The proposed model addresses the multifaceted challenges associated with security vulnerabilities and performance limitations in these networks. This work has demonstrated the effectiveness of combining advanced data analysis techniques with machine learning models to optimise the functionality and protection of IoT blockchain networks. The study has also highlighted the importance of selecting appropriate clustering algorithms and features, balancing data, and exploring alternative classification techniques to further enhance the performance of prediction models. Future research directions include investigating the scalability and efficiency of the proposed approach in larger and more complex IoT blockchain networks, addressing privacy concerns, and integrating anomaly prediction techniques in real-time.

References

[1] M. Atzori, "Blockchain technology and decentralized governance: Is the state still necessary?," *Journal of Governance and Regulation*, vol.6, no.1, pp.45-62, 2017. Article (CrossRef Link)

[2] A. Čolaković and M. Hadžialić, "Internet of Things (IoT): A review of enabling technologies, challenges, and open research issues," *Computer Networks*, vol.144, pp.17-39, 2018. Article (CrossRef Link)

[3] I. O. Adam and M. Dzang Alhassan, "Bridging the global digital divide through digital inclusion: the role of ICT access and ICT use," *Transforming Government: People, Process and Policy*, vol.15, no.4, pp.580-596, 2021. Article (CrossRef Link)

[4] N. A. M. Razali, W. N. W. Muhamad et al., "Secure blockchain-based data-sharing model and adoption among intelligence communities," *IAENG International Journal of Computer Science*, vol.48, no.1, 2021. Article(CrossRefLink)

[5] A. Botta, W. de Donato, V. Persico, and A. Pescapé, "Integration of Cloud computing and Internet of Things: A survey," *Future Generation Computer Systems*, vol.56, pp.684-700, 2016. Article (CrossRef Link)

[6] N. Nasurudeen Ahamed and R. Vignesh, "A Blockchain IoT (BIoT) Integrated into Futuristic Networking for Industry," *International Journal of Mathematical, Engineering and Management Sciences*, vol.7, no.4, pp.525-546, 2022. Article (CrossRef Link)

[7] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, "Internet of Things security and forensics: Challenges and opportunities," *Future Generation Computer Systems*, vol.78, pp.544-546, 2018. Article (CrossRef Link)

[8] S. M. R. Islam, D. Kwak, MD. H. Kabir, M. Hossain, and K. S. Kwak, "The Internet of Things for Health Care: A Comprehensive Survey," *IEEE Access*, vol.3, pp.678-708, 2015. Article (CrossRef Link)

[9] J. Yang, S. He, Y. Xu, L. Chen, and J. Ren, "A Trusted Routing Scheme Using Blockchain and Reinforcement Learning for Wireless Sensor Networks," *Sensors*, vol.19, no.4, 2019. Article (CrossRef Link)

[10] T. Veeramakali, R. Siva, B. Sivakumar, P. C. Senthil Mahesh, and N. Krishnaraj, "An intelligent Internet of Things-based secure healthcare framework using blockchain technology with an optimal deep learning model," *The Journal of Supercomputing*, vol.77, no.9, pp.9576-9596, 2021. Article (CrossRef Link)

[11] P. Urien, "Blockchain IoT (BIoT): A New Direction for Solving Internet of Things Security and Trust Issues," in *Proc. of 2018 3rd Cloudification of the Internet of Things (CIoT)*, 2018. Article (CrossRef Link)

[12] D. Saveetha and G. Maragatham, "Design of Blockchain enabled intrusion detection model for detecting security attacks using deep learning," *Pattern Recognition Letters*, vol.153, pp.24-28, 2022. Article (CrossRef Link)

[13] D. Guha Roy and S. N. Srirama, "A Blockchain-based Cyber Attack Detection Scheme for Decentralized Internet of Things using Software-Defined Network," *Software: Practice and Experience*, vol.51, no.7, pp.1540-1556, 2021. Article (CrossRef Link)

[14] W. Liang, L. Xiao, K. Zhang, M. Tang, D. He, and K. C. Li, "Data Fusion Approach for Collaborative Anomaly Intrusion Detection in Blockchain-Based Systems," *IEEE Internet of Things Journal*, vol.9, no.16, pp.14741-14751, 2022. Article (CrossRef Link)

[15] X. Li, P. Jiang, T. Chen, X. Luo, and Q. Wen, "A survey on the security of blockchain systems," *Future Generation Computer Systems*, vol.107, pp.841-853, 2020. Article (CrossRef Link)

[16] X. Xu et al., "A Taxonomy of Blockchain-Based Systems for Architecture Design," in *Proc. of 2017 IEEE International Conference on Software Architecture (ICSA)*, pp.243-252, 2017. Article (CrossRef Link)

[17] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for IoT security and privacy: The case study of a smart home," in *Proc. of 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp.618-623, 2017. Article (CrossRef Link)

[18] J. Lu, J. Shen, P. Vijayakumar, and B. B. Gupta, "Blockchain-Based Secure Data Storage Protocol for Sensors in the Industrial Internet of Things," *IEEE Transactions on Industrial Informatics*, vol.18, no.8, pp.5422-5431, 2022. Article (CrossRef Link)
[19] E. Ben-Sasson, A. Chiesa, E. Tromer, and M. Virza, "Succinct Non-Interactive Zero Knowledge for a von Neumann Architecture," in Proc. of 23rd USENIX Security Symposium, pp.781-796, Aug.2014. Article(CrossRefLink)
[20] A. Khan, "Graph Analysis of the Ethereum Blockchain Data: A Survey of Datasets, Methods, and Future Work," in *Proc. of 2022 IEEE International Conference on Blockchain (Blockchain)*, pp.250-257, 2022. Article (CrossRef Link)
[21] Y. Lu, "Blockchain and the related issues: a review of current research topics," *Journal of Management Analytics*, vol.5, no.4, pp.231-255, 2018. Article (CrossRef Link)
[22] V. Holotescu and R. Vasiu, "Challenges and Emerging Solutions for Public Blockchains," *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, vol.11, no.1, pp.58-83, 2020. Article (CrossRef Link)
[23] V. Buterin, "A Next-Generation Smart Contract and Decentralized Application Platform," *Ethereum White Paper*, pp.1-36, Jan. 2014. Article(CrossRefLink)
[24] C. Rupa, D. Midhunchakkaravarthy, M. K. Hasan, H. Alhumyani, and R. A. Saeed, "Industry 5.0: Ethereum blockchain technology based DApp smart contract," *Mathematical Biosciences and Engineering*, vol.18, no.5, pp.7010-7027, 2021. Article (CrossRef Link)
[25] F. Yang, W. Zhou, Q. Wu, R. Long, N. N. Xiong, and M. Zhou, "Delegated Proof of Stake With Downgrade: A Secure and Efficient Blockchain Consensus Algorithm With Downgrade Mechanism," *IEEE Access*, vol.7, pp.118541-118555, 2019. Article (CrossRef Link)
[26] A. R. Kairaldeen, N. F. Abdullah, and A. Abu-Samah, R. Nordin, "Data Integrity Time Optimization of a Blockchain IoT Smart Home Network Using Different Consensus and Hash Algorithms," *Wireless Communications and Mobile Computing*, vol.2021, 2021. Article (CrossRef Link)
[27] A. R. Kairaldeen, N. F. Abdullah, A. Abu-Samah, and R. Nordin, "Peer-to-Peer User Identity Verification Time Optimization in IoT Blockchain Network," *Sensors*, vol.23, no.4, 2023. Article (CrossRef Link)
[28] H. Gehani and S. Rathkanthiwar, "A study on Security of IoT based Blockchain System using Artificial Intelligence," in *Proc. of 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP)*, pp. 1-6, 2023. Article (CrossRef Link)
[29] B. Dhiyanesh, L. Shakkeera, V. Y. Sharmasth, H. Azath, S. K. Viswanathan, and V. Poonuramu, "Improved Privacy of Data Transaction in IoT-Enabled Blockchain Technology Using Privacy-Based Machine Learning Algorithms," *Revolutionizing Digital Healthcare Through Blockchain Technology Applications*, pp.187-206, 2023. Article (CrossRef Link)
[30] M. M. Moawad, M. M. Madbouly, and S. K. Guirguis, "Leveraging Blockchain and Machine Learning to Improve IoT Security for Smart Cities," in *Proc. of the 3rd International Conference on Artificial Intelligence and Computer Vision (AICV2023)*, vol.164, pp.216-228. Article (CrossRef Link)
[31] M. A. Khan et al., "A Machine Learning Approach for Blockchain-Based Smart Home Networks Security," *IEEE Network*, vol.35, no.3. pp.223-229, 2021. Article (CrossRef Link)
[32] A. Gepp, M. K. Linnenluecke, T. J. O'Neill, and T. Smith, "Big data techniques in auditing research and practice: Current trends and future opportunities," *Journal of Accounting Literature*, vol.40, no.1, pp.102-115, 2018. Article (CrossRef Link)
[33] J.A. Hartigan and M.A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol.28, no.1, pp.100-108, 1979. Article (CrossRef Link)
[34] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *WIREs Data Mining and Knowledge Discovery*, vol.2, no.1, pp.86-97, 2012. Article (CrossRef Link)
[35] K. Khan, S. U. Rehman, K. Aziz, S. Fong, S. Sarasvady, "DBSCAN: Past, present and future," in *Proc. of the 5th International Conference on the Applications of Digital Information and Web*

Technologies (ICADIWT 2014), pp.232-238, 2014. Article (CrossRef Link)
[36] T. Marwala, "Gaussian mixture models," *Handbook of Machine Learning*, vol.1, pp.245-261, 2018. Article (CrossRef Link)
[37] B. Podgorelec, "Dataset of transactions of 10 Ethereum addresses controlled by a private key, each has at least 2000 output transactions, which include a transfer of cryptocurrency, and all transactions are performed within no longer than three months period," *Zenodo*, Nov. 2019. Article (CrossRef Link)
[38] H. Abdi and L. J. Williams, "Principal component analysis," *WIREs Computational Statistics*, vol.2, no.4, pp.433-459, 2010. Article (CrossRef Link)
[39] S. S. Azmi and S. Baliga, "An overview of boosting decision tree algorithms utilizing AdaBoost and XGBoost boosting strategies," *International Research Journal of Engineering and Technology (IRJET)*, vol.7, no.5, pp.6867-6870, 2020. Article(CrossRefLink)
[40] J. C. W. Chan and D. Paelinckx, "Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sensing of Environment*, vol.112, no.6, pp.2999-3011, Jun. 2008. Article (CrossRef Link)
[41] V. Aliyev, "Ethereum Fraud Detection Dataset," 2020. [Online]. Available: https://www.kaggle.com/datasets/vagifa/ethereum-frauddetection-dataset?resource=download
[42] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *EURASIP Journal on Wireless Communications and Networking*, vol.2021, 2021. Article (CrossRef Link)
[43] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *Proc. of IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp.747-748, 2020. Article (CrossRef Link)
[44] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen, "Assessing the accuracy of prediction algorithms for classification: an overview," *Bioinformatics*, vol.16, no.5, pp.412-424, May 2000. Article (CrossRef Link)
[45] M. Kane, "The precision of measurements," *Applied Measurement in Education*, vol.9, no.4, pp.355-379, 1996. Article (CrossRef Link)
[46] B. B. Murdock, "The serial position effect of free recall," *Journal of Experimental Psychology*, vol.64, no.5, pp.482-488, 1962. Article (CrossRef Link)
[47] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, vol.3, no.10, pp.27-38, 2013. Article(CrossRefLink)

Nor Fadzilah ABDULLAH is an Associate Professor of Electrical, Electronic and Systems Engineering at Universiti Kebangsaan Malaysia (UKM). She received a M.Sc. degree from the University of Manchester (2003) and a Ph.D. degree in Electrical and Electronic Engineering from the University of Bristol (2012). Her research interests include beyond 5G networks, vehicular networks, machine learning, channel propagation modelling and information theory.

Ammar Riadh Kairaldeen ALTOTANJE received the B.Sc. degree in Computer engineering from Middle Technical University, Iraq, in 2004 and the M.Sc. degree in Computer engineering from the Cankaya University, Turkey, in 2014, and Ph.D. degree in Electrical and Electronic Engineering, from the Universiti Kebangsaan Malaysia (UKM), in 2024. His research interests include blockchain technology, consensus algorithms, machine learning, anomaly detection, Internet of Things (IoT), network security, and performance optimization.

Asma ABU-SAMAH received the B.Sc. and M.Sc. degrees in electrical, electronics, automation systems and signal processing from Université de Joseph Fourier Grenoble, France (2008 and 2010), and the Ph.D. degree in automated control and production systems from Université de Grenoble-Alpes, France (2016). She was a Postdoctoral Researcher in biomedical control systems with Universiti Tenaga Nasional (UNITEN), Malaysia, from 2017 to 2019. She is currently a Senior Lecturer with Universiti Kebangsaan Malaysia (UKM). Her research interests include the application of machine learning in wireless communications and IoT.

Rosdiadee NORDIN (Senior Member, IEEE) is a Professor of Electrical, Electronic & Systems Engineering at Sunway University, Malaysia. He graduated with a B.Eng. in Electrical, Electronic and Systems Engineering from Universiti Kebangsaan Malaysia (UKM) and Ph.D. in wireless engineering at the University of Bristol, UK (2010). His research interests include Beyond 5G wireless communications, channel modeling, aerial wireless communications, and wireless Internet of Things applications.