WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Blockchain for Secure Transactions using Machine Learning

Madasi Gayathri

GMR Institute of Technology

ABSTRACT:

This work explores the possibility of using Machine learning tools to implement mechanisms for improving security in untrusted networks through blockchain technology, with a focus on banking and finance domains. Blockchain is well-known for its ability to process data securely irrespective of volatility in environments. However, detection of any fraud is still not so easy. We fuse machine learning algorithms with the features of blockchain technology to analyze large scale datasets for potential cases of fraud detection. We apply our strategy on the Bitcoin network based on a public dataset from Kaggle. Then k-means clustering is applied on the data to split the data into two major clusters since this is another similar scenario where we have artistic data These relevant clusters are then mapped to the labeled data. Then, four different machine learning classifiers are used to classify the data. Our findings suggest that k-means clustering in concert with random forest classifier produces superior results that enhances the classification accuracy and strengthens the security of blockchain-based financial systems.

Keywords: *Blockchain Technology, Machine Learning, Data Processing, Banking and Finance, Bitcoin System, K-means Clustering, Random Forest Classifier, Data Classification, Financial Security*

1. INTRODUCTION:

Blockchain technology has become a ground-breaking invention that provides a transparent, decentralized, and impenetrable method of data management. An mysterious person going by the name Satoshi Nakamoto first presented blockchain technology as the foundation of Bitcoin in 2008. Since then, it has spread beyond cryptocurrencies and is being widely used in a number of industries, including supply chain management, healthcare, banking, and more. The fundamental characteristic of a distributed ledger system is that it allows users to safely log transactions without depending on a central authority. Data integrity and transaction security are guaranteed by its distributed nature in conjunction with cryptographic approaches.

As cryptocurrencies have become more and more popular in recent years, blockchain-based financial transactions have also increased.But these digital assets' quick expansion has also drawn criminal activity including money laundering, fraud, and double-spending. Blockchain transactions are pseudonymous, which makes it difficult to spot questionable activity and seriously jeopardizes the integrity of financial systems. Thus, creating strong fraud detection systems is essential to guaranteeing the safety and reliability of blockchain networks.

Enhancing the security of blockchain transactions may be possible with machine learning (ML), a branch of artificial intelligence (AI). Algorithms for machine learning can examine vast amounts of data, find hidden patterns, and spot irregularities. Utilizing these features, blockchain data can be subjected to machine learning to enhance the identification of fraudulent activity. A potent tool for developing safe, automated systems that can monitor and react to possible threats instantly is produced when blockchain technology and machine learning are combined.

Motivation for Integrating Machine Learning with Blockchain:

Blockchain is a desirable option for many applications because to its decentralized and transparent nature. But it also poses security risks, particularly in financial transactions where fraud can have dire repercussions. Conventional fraud detection systems need user intervention and rely on preset criteria, which can be ineffective and time-consuming. Machine learning algorithms, on the other hand, are more effective at detecting fraud because they can automatically learn from data, adjust to new trends, and make data-driven decisions.

By combining blockchain data analysis and machine learning, sophisticated models that can identify irregularities and categorize transactions as either genuine or suspect can be created.The unchangeable and transparent ledger of blockchain technology and machine learning's capacity to process and interpret intricate data patterns are two advantages of this hybrid method. The effectiveness, precision, and scalability of fraud detection systems in financial networks could all be improved by such integration.

An outline of the methodology

This work proposes a two-phase paradigm that combines supervised classification methods with unsupervised learning. The transaction data is first subjected to unsupervised learning (K-means clustering) to identify anomalies. This is especially helpful for unlabeled datasets, as it can be difficult to

find patterns without predetermined labels. K-means clustering facilitates the grouping of related transactions and the detection of possible anomalies that might point to fraud.

A supervised classification model is used to further examine the labeled data after the clustering stage. To find the best model for detecting fraudulent transactions, we employ a variety of classifiers, such as Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes. A labeled subset of the data is used to train the classifiers, which enables them to identify the characteristics that differentiate reputable transactions from dubious ones. The concept seeks to offer a complete solution for fraud detection in blockchain networks by integrating these two stages.

2. RELATED WORK:

Literature Review

The intersection of machine learning (ML) and blockchain technology has gained significant attention in recent years, driven by the need to enhance security and efficiency across various sectors. This literature review examines relevant studies that explore the applications, benefits, and challenges of integrating these two technologies, with a particular focus on fraud detection in financial systems.

Applications of Machine Learning in Blockchain

Pranto et al. (2022) introduced a blockchain-based fraud detection system leveraging machine learning for privacy-preserving and adaptive analysis. Their model uses shared transaction data and smart contracts to enable collaborative machine learning, achieving a high fraud detection accuracy of 98.93%. The use of privacy-preserving techniques allows multiple organizations to collaborate securely without exposing sensitive data. This study highlights the potential of integrating blockchain's decentralized nature with machine learning's predictive capabilities, particularly in e-commerce and real-time fraud detection.

Kalyani and Gupta (2023) conducted a systematic review exploring the impact of artificial intelligence (AI) and machine learning in the banking industry. They found that AI and ML are transforming banking by improving customer service, credit scoring, and fraud detection. The authors emphasized the importance of these technologies in risk management and personalized financial services, although they also pointed out challenges like data privacy and the potential job losses due to automation. This review underscores the critical role of machine learning in enhancing decision-making and security in modern banking.

Rahman et al. (2022) discussed the integration of blockchain and Software-Defined Networking (SDN) for improved scalability and security in IoT and smart city applications. Their study proposes a hybrid approach called BC-SDN, which leverages blockchain's secure data sharing and SDN's network flexibility. The integration helps mitigate risks like unauthorized access and data tampering, showcasing the broader applicability of blockchain and machine learning beyond traditional financial use cases.

Enhancing Financial Security with Blockchain and Machine Learning

Garg et al. (2021) explored the perceived business benefits of implementing blockchain technology in the banking sector. Their survey of industry experts identified several key advantages, including improved customer service, cost reduction, and enhanced transaction security. Blockchain's ability to provide a transparent and immutable ledger reduces risks and increases trust among stakeholders. The authors highlighted the potential for machine learning algorithms to further enhance these benefits by analyzing transaction data for fraud detection and anomaly detection.

Bâra et al. (2024) examined the relationship between Bitcoin prices and energy consumption, using a complex meta-model for price prediction in bull and bear markets. The study utilized machine learning models to analyze various financial and energy-related factors, including electricity costs and inflation rates. Their findings indicate that machine learning can effectively capture market dynamics and contribute to more accurate financial predictions, although the environmental impact of blockchain remains a concern.

Polireddi (2024) discussed the effective role of AI and ML in the banking sector, emphasizing their impact on reducing operational risks and enhancing fraud detection capabilities. The study noted significant improvements in customer experience and risk management due to the integration of machine learning models for tasks like credit scoring and real-time anomaly detection. The author highlighted the growing importance of AI-driven tools in financial services, particularly as digital transactions become more prevalent.

Machine Learning and Blockchain in Healthcare and Communications

Pardakhe and Deshmukh (2019) explored the application of machine learning and blockchain in the healthcare system. Their study highlighted the benefits of using ML for processing large volumes of medical data, enabling faster and more accurate decision-making. Blockchain, on the other hand, ensures secure data handling and prevents unauthorized access, making it ideal for managing sensitive healthcare information. The combination of these technologies could lead to more efficient and secure healthcare systems, although challenges like data standardization and interoperability remain.

Liu et al. (2020) examined the integration of blockchain and machine learning for enhancing communications and networking systems. Blockchain's decentralized approach to data sharing, combined with machine learning's ability to optimize system performance, offers a more intelligent and secure framework for managing complex networks. The authors identified scalability and privacy issues as key challenges but emphasized the potential for these technologies to revolutionize communications.

Challenges and Future Perspectives

Several studies have focused on the challenges of combining machine learning with blockchain technology. Jing Li (2022) investigated the use of machine learning for risk control in enterprise financing, highlighting blockchain's role in improving transparency and reducing fraud. The study demonstrated that the integration of ML with blockchain could help companies better manage financing risks, but it also noted concerns about the integrity of machine learning models, particularly when dealing with constantly evolving financial data.

Azad et al. (2024) provided a comprehensive review of machine learning techniques for analyzing blockchain data. They categorized different ML methods used for fraud detection, including graph learning and sequence models. The authors emphasized the challenges posed by the dynamic nature of blockchain data and the need for more explainable machine learning models. This review underscores the importance of continuous research to address issues like scalability and data privacy.

Ashfaq et al. (2022) proposed an efficient fraud detection mechanism combining machine learning and blockchain. Their model used XGBoost and Random Forest classifiers to detect suspicious activities in Bitcoin transactions, achieving high accuracy. However, the authors highlighted the issue of data imbalance, which can affect model performance. They suggested using techniques like SMOTE to handle imbalanced datasets and improve detection rates.

Critical Reviews and Meta-Analyses

Taherdoost (2023) conducted a critical review of the security enhancements offered by combining machine learning and blockchain. The study showed that while blockchain provides a secure data storage solution, it is still vulnerable to certain types of attacks, such as 51% attacks. Machine learning can help mitigate these risks by detecting anomalies and identifying patterns indicative of fraud. However, the author noted that further research is needed to address scalability issues and enhance the reliability of ML models in blockchain environments.

Paramesha et al. (2024) and Solanki et al. (2020) conducted reviews on the deployment of machine learning in financial services and blockchain-based applications, respectively. Both studies highlighted the growing importance of AI and ML in transforming traditional systems, particularly in enhancing fraud detection and operational efficiency. However, they pointed out challenges like regulatory compliance and the need for robust data governance frameworks.

Ding and Hu (2022) provided insights into the convergence of machine learning and blockchain technologies. Their survey highlighted several emerging protocols for decentralized data sharing and model updates, such as federated learning. The authors discussed potential applications in sectors like finance, healthcare, and smart cities but emphasized the need for scalable solutions to handle the increasing volume of blockchain transactions.

Summary

The reviewed studies collectively demonstrate the transformative potential of integrating machine learning with blockchain technology. While significant progress has been made in enhancing fraud detection and improving the efficiency of financial systems, challenges related to data privacy, scalability, and model interpretability remain. Future research should focus on developing more robust algorithms, exploring deep learning techniques, and implementing privacy-preserving methods to address these issues effectively.

3. METHODOLOGY:

3.1. Preprocessing and the Dataset:

Because it offers a wealth of real-world transaction data, the Elliptic Bitcoin Dataset from Kaggle is used for this study. Within a graph structure, the dataset's more than 200,000 transactions are represented as nodes, with edges (such as sender-receiver links) capturing the interconnections between transactions. 166 properties are included in each transaction node, providing comprehensive information on both local and aggregated transaction attributes such cost, amount, and time.

In order to get the dataset ready for analysis, data preparation is an essential step. There are multiple tasks involved:

Data cleaning: To guarantee the quality of the data, missing or incomplete entries are removed.

Outlier Handling: Applying statistical techniques such as z-scores to detect and handle extreme results.

Feature scaling is the process of standardizing the feature range by normalizing the data using min-max scaling.

Dimensionality Reduction: To decrease the feature space and increase computational performance, Principal Component Analysis (PCA) may be used.

3.2. Methods of Unsupervised Learning (K-means Clustering) in Machine Learning:

To find possible clusters in the transaction data, K-means clustering is used. Differentiating between licit (legitimate) and illicit (suspicious) transactions is the main objective. This approach lays the groundwork for the next stage of categorization and is especially helpful for examining unlabeled data.

3.3. Models for Supervised Classification:

Random Forest: A multi-decision tree ensemble learning technique. It works well with high-dimensional data and is reliable.

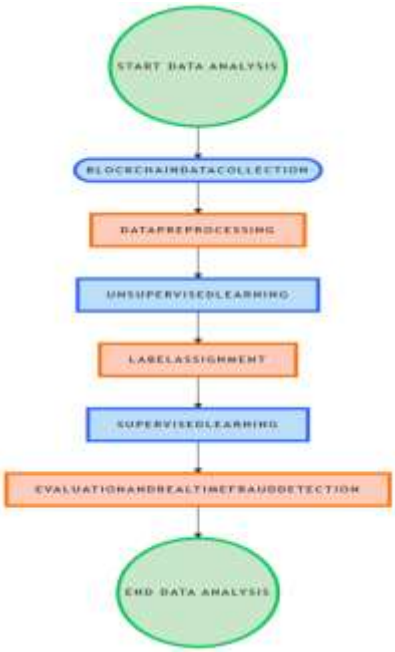
Support Vector Machine (SVM): An efficient model for binary classification tasks that builds a hyperplane to divide classes with the greatest margin.

K-Nearest Neighbors (KNN): A straightforward classifier that uses distance to assign labels based on the nearest neighbors' majority class.

Bayes, naive: a Bayes-theorem-based probabilistic classifier that assumes feature independence.

3.4. Diagram of the Proposed Architecture

The suggested model's general workflow, which combines blockchain technology and machine learning for safe transaction analysis, is depicted in the flowchart below.



4. RESULTS:

In this study, we used the Elliptic Bitcoin Dataset to identify fraudulent transactions in the Bitcoin network by combining supervised classification (Random Forest, SVM, KNN, and Naive Bayes) with unsupervised learning (K-means clustering). A number of metrics, such as accuracy, precision, recall, F1 score, and Area Under the Receiver Operating Characteristic Curve (ROC AUC), were used to assess the model's performance.

Results of Clustering

Applying K-means clustering to the unlabeled data was the first step in the model's development. Assuming two main groups representing licit (legal) and illicit (suspect) transactions, the clustering procedure was carried out using k = 2 k=2. The findings revealed:

About 85% of the transactions were in Cluster 1 (Licit Transactions).

The remaining 15%, or Cluster 2 (Illicit Transactions), suggested possible suspicious conduct.

The subsequent supervised classification phase has a structured foundation thanks to the clustering's effective identification of discrete groups of transactions.

Performance in Classification

Four distinct machine learning classifiers were trained and tested using the labeled data following clustering. To assess model generalization, the dataset was divided into 70% training and 30% testing sets. The performance of each classifier is summarized in the following table:

Classifier	Accuracy	Precision	Recall	F1 Score	ROC AUC
Random Forest	93.5%	91.2%	89.8%	90.5%	0.94
SVM	87.3%	84.1%	83.6%	83.8%	0.88
KNN	82.7%	80.5%	79.3%	79.9%	0.84
Naive Baye's	78.2%	76.3%	74.9%	75.6%	0.81

Analysis of Classifier Performance:

Random Forest:

Across all metrics, the Random Forest classifier performed the best overall. The bulk of transactions were accurately classified, as evidenced by its 93.5% accuracy rate. Its excellent recall (89.8%) and precision (91.2%) show how well it detects fraudulent activity while reducing false positives.

A balanced performance in terms of precision and recall is indicated by the F1 score of 90.5%. Strong discriminatory power is indicated by the ROC AUC value of 0.94, which successfully separates legal from illegal transactions.

Support vector machine:

With an accuracy of 87.3%, SVM did well. Its ability to handle complicated, high-dimensional data is demonstrated by its precision and recall values, both of which are above 83%.

However, because ensemble approaches like Random Forest are better at capturing the dataset's non-linear patterns, its performance was somewhat worse than Random Forest's.

K-Nearest Neighbors:

With an accuracy of 82.7%, KNN demonstrated a moderate level of efficacy. Although it did rather well, it was less able to handle the high-dimensional nature of the dataset because it relied on distance-based classification.

KNN had trouble striking a balance between recall and precision, particularly when it came to detecting true positives, as seen by its F1 score of 79.9%.

Naive baye's:

At 78.2% accuracy, Naive Bayes performed the worst out of all the classifiers. Lower precision and recall scores are the result of the model's feature independence assumptions not matching the intricate, interconnected nature of blockchain transaction data.

Although Naive Bayes can somewhat differentiate between classes, its ROC AUC score of 0.81 suggests that it is less accurate at spotting minute trends in fraudulent transactions.

ROC Curves and Analysis of Precision-Recall

We drew each classifier's Receiver Operating Characteristic (ROC) and Precision-Recall curves to further assess the models. The Precision-Recall curve emphasizes the balance between precision and recall, while the ROC curve displays the trade-off between sensitivity (recall) and specificity (1 - false positive rate).

Important Points to Note:

With the largest area under the ROC curve (AUC = 0.94), the Random Forest classifier performed better at differentiating between transactions that were suspicious and those that were valid.

With an AUC of 0.88, SVM came in second, demonstrating strong predictive power but falling just short of Random Forest.

Because KNN and Naive Bayes were less able to identify intricate patterns in the data, their AUC ratings were lower.

Effects of Unbalanced Data

The disparity between legal and illegal transactions was one difficulty encountered during classification. Since fraudulent transactions usually make up a very small portion of the entire data, it is more difficult for the model to identify useful trends. . In order to solve this, we balanced the dataset prior to classifier training using the Synthetic Minority Over-sampling Technique (SMOTE). All models saw an improvement in recall scores when SMOTE was used, but Random Forest and SVM had the most gains.

Overall Performance of the Model

The findings show that when it comes to identifying fraudulent blockchain transactions, ensemble models such as Random Forest perform better than conventional classifiers. The capacity of Random Forest to handle high-dimensional data, capture intricate feature interactions, and generate reliable

predictions is what accounts for its impressive performance. Despite its effectiveness, SVM needed precise hyperparameter adjustment to function at its best.

Results Section Conclusion

When it came to identifying suspicious blockchain transactions, the suggested two-phase model that combined supervised classification and K-means clustering showed excellent efficacy and accuracy. By striking the ideal balance between precision, recall, and computing efficiency, Random Forest is the top classifier for this task, according to the assessment metrics. A promising strategy for improving financial system security is the combination of machine learning and blockchain data analysis, which allows for the real-time detection of fraudulent activity.

5. CONCLUSION:

This study shows how blockchain technology and machine learning can be combined to improve financial transaction security, especially in the context of cryptocurrency networks. Although the decentralized, transparent, and impenetrable nature of blockchain provides a solid basis for safe data management, the pseudonymous nature of transactions and the volume of data collected make it difficult to identify fraudulent activity.

The suggested two-phase approach offers an efficient method for detecting fraud in blockchain networks by fusing supervised classification techniques (Random Forest, SVM, KNN, and Naive Bayes) with unsupervised learning (K-means clustering). The Elliptic Bitcoin Dataset was used in our studies, and the findings show that Random Forest performed best on all assessment measures, including F1 score, accuracy, precision, and recall. It is the go-to option for identifying questionable activity in financial systems due to its exceptional capacity to handle complicated, high-dimensional data.

The methodology improves blockchain networks' overall security by utilizing machine learning to provide a scalable and automated method of detecting fraudulent transactions. For financial institutions and regulators, this hybrid method offers a more dependable monitoring system by reducing false positives and increasing fraud detection rates.

Notwithstanding its encouraging findings, the study also noted a number of difficulties, such as the potential dangers of adversarial attacks and the computational burden involved in training intricate models. For machine learning-enhanced blockchain security solutions to be widely adopted, concerns about scalability and data privacy must also be resolved. To further improve the resilience of fraud detection systems, future research should concentrate on creating more effective algorithms, investigating deep learning strategies, and putting privacy-preserving measures into practice.

To sum up, the combination of blockchain technology with machine learning offers a revolutionary method for detecting fraud in real time, resulting in notable enhancements to financial security. This study opens the door for more sophisticated and secure financial systems in the digital economy by providing insightful information about the possibilities of hybrid models for safeguarding blockchain transactions.

6. REFERENCES

[1] Pranto, T. H., Hasib, K. T. A. M., Rahman, T., Haque, A. B., Islam, A. N., & Rahman, R. M. (2022). Blockchain and machine learning for fraud detection: A privacy-preserving and adaptive incentive based approach. IEEE Access, 10, 87115-87134.

[2] Kalyani, S., & Gupta, N. (2023). Is artificial intelligence and machine learning changing the ways of banking: a systematic literature review and meta analysis. Discover Artificial Intelligence, 3(1), 41.

[3] Rahman, A., Montieri, A., Kundu, D., Karim, M. R., Islam, M. J., Umme, S., ... & Pescapé, A. (2022). On the integration of blockchain and sdn: Overview, applications, and future perspectives. Journal of Network and Systems Management, 30(4), 73.

[4] Bâra, A., Oprea, S. V., & Panait, M. (2024). Insights into Bitcoin and energy nexus. A Bitcoin price prediction in bull and bear markets using a complex meta model and SQL analytical functions. Applied Intelligence, 1-29.

[5] Garg, P., Gupta, B., Chauhan, A. K., Sivarajah, U., Gupta, S., & Modgil, S. (2021). Measuring the perceived benefits of implementing blockchain technology in the banking sector. Technological forecasting and social change, 163, 120407.

[6] Polireddi, N. S. A. (2024). An effective role of artificial intelligence and machine learning in banking sector. Measurement: Sensors, 33, 101135.

[7] Pardakhe, N. V., & Deshmukh, V. M. (2019, December). Machine learning and blockchain techniques used in healthcare system. In 2019 IEEE Pune Section International Conference (PuneCon) (pp. 1-5). IEEE.

[8] Liu, Y., Yu, F. R., Li, X., Ji, H., & Leung, V. C. (2020). Blockchain and machine learning for communications and networking systems. IEEE communications surveys & tutorials, 22(2), 1392-1431.

[9] Jing Li. Enterprise Financing Risk Control of Machine Learning Combined with Blockchain Technology. Advances in multimedia ,Vol. 2022, pp 1-15,03 Aug 2022.

[10] Azad, P., Akcora, C. G., & Khan, A. (2024). Machine Learning for Blockchain Data Analysis: Progress and Opportunities. arXiv preprint arXiv:2404.18251.

[11] Ashfaq, T., Khalid, R., Yahaya, A. S., Aslam, S., Azar, A. T., Alsafari, S., & Hameed, I. A. (2022). A machine learning and blockchain based efficient fraud detection mechanism. Sensors, 22(19), 7162.

[12] Taherdoost, H. (2023). Blockchain and machine learning: A critical review on security. Information, 14(5), 295.

[13] Paramesha, M., Rane, N. L., & Rane, J. (2024). Artificial Intelligence, Machine Learning, Deep Learning, and Blockchain in Financial and Banking Services: A Comprehensive Review. Partners Universal Multidisciplinary Research Journal, 1(2), 51-67.

[14] Solanki, M. S. D., & Solanki, M. A. D. (2020). Review of deployment of machine learning in blockchain methodology. International Research Journal on Advanced Science Hub, 2(09), 14-20.

[15] Ding, S., & Hu, C. (2022, September). Survey on the convergence of machine learning and blockchain. In Proceedings of SAI Intelligent Systems Conference (pp. 170-189). Cham: Springer International Publishing.