International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 09 | September - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

EFFECTIVE FRAUD DETECTION IN BLOCKCHAIN USING XGBOOST WITH RANDOM FOREST

Sandeep Rao[1], Prof. Brajesh Patel[2,] Prof. Anshul Khurana[3]

[1]Roll(0205CS21MT05), M. Tech Scholar, M. Tech (CTA), Department of Computer Science & Engineering,

Shri Ram Institute of Technology, Jabalpur, MP.

[2]Head of Department, Department of Computer Science & Engineering, Shri Ram Institute of Technology, Jabalpur, MP.

[3]Head of P.G Department, Department of Computer Science & Engineering, Shri Ram Institute of Technology, Jabalpur, MP.

Abstract

To solve the issue of fraud and abnormalities in the Bitcoin network, we offer a model. Here, we classify transactions based on integrated and fraudulent transaction patterns using machine learning methods like XGboost and Random Forest (RF). Then, future incoming transactions are predicted using the trained dataset. To identify fraudulent transactions, the authors use blockchain technology with machine learning algorithms.

The proposed model performs a security analysis of the proposed smart contract to show the system's robustness and determines the precision and AUC of the models to gauge accuracy. To guard against assaults and vulnerabilities on the suggested system, the authors additionally suggest an attacker model.

Overall, as financial technology advances, the proposed approach seeks to offer a more secure method for spotting fraud in the Bitcoin network. The authors assert that their suggested system's use of machine learning and blockchain technology may successfully identify and stop fraudulent transactions.

Introduction:

Bitcoin, the first decentralized digital currency, was introduced by Satoshi Nakamoto in 2008 as a breakthrough digital payment mechanism. Bitcoin is protected by encryption and a proof-of-work consensus process, and it can be sent from user to user without the use of a middleman organization. Other digital currencies have been created since its launch, including Ethereum, Ripple, and Litecoin, and the market value of cryptocurrencies has increased dramatically. Digital currencies built on blockchain technology have given users new options for transactions and financial infrastructure.

Companies must utilize secure authentication techniques like multi-factor authentication, biometric authentication, and encryption to safeguard digital transactions from such attacks. In order to access a service or system, users must give extra layers of authentication, such as a password, PIN, or fingerprint. Users must present physical identification in order to employ biometric authentication, such as a face or iris scan. Data is encrypted and secured during transmission across networks and storage on databases thanks to encryption.

Every year, billions of dollars are lost worldwide as a result of fraud. Advanced technologies like machine learning, AI, biometrics, and analytics can be used to address this. These technologies can assist in real-

time fraud detection and prevention, lowering the likelihood of loss and fraud. Additionally, banks and other financial institutions can collaborate with outside companies to leverage their analytics and data to spot and stop fraudulent activity.

Literature review

M. Ostapowicz and others. Consensus approaches look at the accuracy of the transactions that were carried out. However, the transfers cannot be efficiently identified. As a result, employing blockchain as a fraud detection technique does not completely address the problem. As a result, novel methods, such as ML techniques, are employed to address the shortcomings in the current systems. Additionally, it is possible to spot possibly fraudulent actions using AI-based solutions like machine learning and deep learning.

To identify fraudulent transactions, Azar A.T. et al offered a variety of supervised machine learning algorithms, including logistic regression, decision trees, support vector machines (SVM), artificial neural networks (ANN), and naive Bayes classifiers. These machine learning techniques are compared by the authors, who also take into account their recall, accuracy, and other metrics. The findings indicate that the SVM method is the most effective at detecting fraud and has the highest accuracy. Finally, they argue that additional study is required to investigate the possibilities of alternative machine learning techniques and to raise the precision of the existing approaches.

To recognize fraudulent businesses, P. K. R. et al. offered a variety of Supervised Machine Learning approaches. Additionally, they evaluated the test dataset using XGBoost and Random Forest classifiers; the accuracy of the suggested solutions was over 96%. The writers also discussed the limitations and offered ideas for additional research directions.

According to Byun Y.C. et al., XGboost was able to anticipate the experiment's outcome with greater accuracy than other machine learning methods. This study came to the conclusion that XGboost might be a helpful tool for forecasting results in various trials.

According to Guillen M. et al., the XGboost model has an F1 score of 0.8, an accuracy of 90%, and a precision of 86%. The model was also successful in identifying the dataset's key properties, according to the authors. They came to the conclusion that XGboost is a useful technique for predicting an individual's driving performance.

Using Li J et al Utilizing data mining techniques, fraud can be found by looking at transaction history, account information, and other relevant data. Algorithms for data mining can find ominous trends and reveal hidden connections between transactions. Data mining can assist in identifying these patterns and detecting fraud to help stop it from happening again.

Banks are concerned that sharing client data could expose them to hazards such possible hackers having access to the data, according to Reid F et al. utilized confidential property for banks. Banks are also concerned that disclosing consumer information may result in legal problems, such as potential lawsuits or other forms of liability. Banks are reluctant to share client information with researchers as a result, which might make it challenging to investigate the banking sector.

The approach suggested by Sun X et al. included multiple phases, including preprocessing of the data, feature extraction, classification, and postprocessing. Data cleansing, normalization, and feature selection were all parts of the data preprocessing stage. Feature extraction and selection were stages of the feature extraction process. The stage of classification involved classifying the data using various techniques. In the postprocessing stage, the findings were assessed, and a general conclusion was given.

The model's outcomes demonstrated its accuracy in identifying tilted delivery and non-uniform expenses.

The method presented by Saia R et al. is based on the idea of self-supervised learning and uses the transactional data's intrinsic structure to find innovations in the data that might be signs of fraud. A training phase and a detection phase make up the suggested algorithm. A self-supervised learning model is trained utilizing the transactional data during the training phase. This model is used to determine the data's typical behaviour and produce typical profiles. The created profiles are utilized in the detection phase to find fraud and spot outliers. The suggested approach outperforms the existing algorithms in terms of accuracy and precision, according to testing on a real-world dataset.

Vila M.A. et al. suggested a brand-new technique for locating suspected fraud tendencies utilizing a cutting-edge algorithm for mining ambiguous association rules. The program was evaluated using a sizable dataset of credit card transactions, and it showed promise in spotting possible fraud trends. The findings demonstrated that the suggested algorithm outperformed conventional association rule mining techniques in terms of precision, recall, and F-measure. The suggested method also offered a more thorough study of the data, allowing the discovery of intricate fraud patterns. The proposed algorithm is a useful tool for identifying fraud in credit card transactions, according to the authors' findings.

A dataset of credit card transactions with both labelled and unlabelled data was used to train the suggested model by Abdulai J.D. et al. The model had a 97.4% accuracy rate. The model had a 96.3% accuracy rate in identifying the fraudulent transactions. This model can be used to spot fraudulent activity and aid in minimizing the damages brought on by such activity.

The dataset from a bank was used by P.J. et al. to train the model. The precision and accuracy of the model's predictions were used to assess it. The findings demonstrated the model's accuracy and precision in identifying fraudulent transactions. These findings show how well the SVM model works to identify fraudulent credit card transactions.

By using the Bayesian learning technique, Kundu A. et al. were able to recognize the patterns of frauds and produce a probabilistic model to categorize them. A preset set of rules that can be used to spot frauds is created using the rule-based learning technique. The results from the other two procedures are then combined using the Dempster-Shafer theory, which aids in lowering the number of false alarms. The study's findings demonstrated that combining these three methods greatly increased the accuracy of fraud detection.

To identify potentially fraudulent transactions, G.F.B et al. use a variety of supervised learning techniques, including Random Forest and SVM. To further categorise the transactions in the dataset, the authors combined a number of data mining techniques, including logistic regression and decision trees. The authors have created a model that can recognize fraudulent transactions from unknown datasets using a supervised learning method. Both clean and noisy datasets can use this technique to detect fraudulent transactions. The model can also spot fraudulent transactions in both recent and historical records. The model can give outstanding results and does a great job of identifying fraudulent transactions.

In centralized-based IoT-driven smart cities, Tripathi et al. identified the problems with trust, privacy, security, and verifiability. The authors suggested a dependable privacy-preserving secure framework (TP2SF) to overcome these problems. This framework attempts to guarantee the security, reliability, and verifiability of all data and information transferred in a smart city. The TP2SF

framework is built on distributed ledger and blockchain technology, which enables safe and dependable data sharing and management. The framework also makes use of intelligent agents and cryptographic methods to protect the privacy of user data and information. The authors also ran simulations to gauge how well their suggested framework worked. The outcomes demonstrated that the TP2SF architecture is capable of offering a secure and dependable environment for data transmission.

The system is safe and private because to the two-layered privacy modules that Zack v. et al. proposed. The system is protected against different attacks, including DDoS attacks, using the enhanced proof of work (ePoW) technology. User data is safeguarded and malicious users are kept out of the system using the principle component analysis (PCA) technique. The system's trustworthiness module is a critical component because it is essential to maintaining the system's security and privacy. The trustworthiness module also contributes to ensuring the system's dependability and credibility.

Y. Zhao et al. suggested a thorough strategy for protecting privacy. To guarantee the confidentiality of the data, they combined asymmetric, symmetric, and homomorphic encryption approaches. These methods provide safe data transport, calculation, and storage. The considerable computing power and implementation time needed for these systems, meanwhile, may be a drawback

Moustafa N. et al., Proposed In order to examine the data recorded in the blockchain and spot fraudulent activity, deep learning algorithms are deployed. The algorithms can also be employed to find weaknesses in the system and suggest fixes. Deep learning algorithms can also be used to spot patterns and correlations in data that indicate suspicious or malicious activity. Then, with this knowledge, any threats can be immediately addressed. In conclusion, data privacy and security can be enhanced by using blockchain technology and deep learning algorithms to efficiently detect cyber threats and incursion attempts.

The local anomaly detection models and the global anomaly detection model make up the two components of the Turnbull B et al. proposed MLO system. The global anomaly detection model evaluates the results of the local anomaly detection models to determine if the activities are normal or anomalous, whereas the local anomaly detection models are used to identify aberrant activities in each of the cloud nodes. The technology was successfully tested by the authors in a real-world cloud setting. As a result, they came to the conclusion that the MLO system is a useful tool for identifying internal and external attacks on cloud-based systems.

Using a deep learning-based system, Keshk M. et al. suggested an enhanced privacy-preserving anomaly detection technique. The pre-processing module and the anomaly detection module are the two components on which this system is built. The features are extracted from the raw data by the pre-processing module using a neural network. Recurrent neural networks (RNNs) are used by the anomaly detection module to find anomalies in the data. A real-world dataset is used to test the suggested system, and the results are compared to those of other approaches. The outcomes demonstrate that the suggested approach performs better in terms of accuracy and efficiency than the current methods.

Methodology:

A machine learning model is used to categorize transactions as malicious or genuine in the system model, which consists of a distributed ledger (Block chain) that maintains the transactions. The machine learning model is created using the data from the blockchain and is trained to recognize anomalies. Following that, the transactions are examined for fraud and anomaly detection using the machine learning model. Systems for finding anomalies in a

system can aid in locating fraud, security risks, process anomalies, and other problems. The dataset utilized for the model includes information on each transaction, including the sender, receiver, amount, time, etc.

Additionally, the information is classified as either good or bad transactions. The classifiers are then used to determine whether a new transaction is trustworthy or malicious after the model has been trained on this dataset. Precision, recall, f1-score, and confusion matrix are some accuracy metrics that are used to assess the model's correctness.

Machine learning techniques are used to train the model, which uses labelled data to teach the system how to classify data. A training set and a testing set are created from the data, with the training set being used to train the model and the testing set being used to assess the model's performance and accuracy.

SMOTE Data Balancing

In machine learning, data imbalance is a prevalent issue that can result in classifiers that are biased towards the majority class and have low predictive accuracy for the minority class. Machine learning algorithms typically perform less accurately when the data is unbalanced because they are unable to identify the underlying trends and linkages. It is determined as the ratio between the number of instances of one class and the number of instances of the other class and serves as a gauge of the dataset's imbalance.

The dataset is more unbalanced the greater the ratio. SMOTE creates fresh synthetic samples in between similar instances by first choosing them. This helps to lessen the bias toward the dominant class and guarantees that the synthetic data is as accurate as possible. By increasing the amount of minority class samples and balancing the dataset, this strategy helps to increase the accuracy of a classification model.

SMOTE creates fictitious new data points to oversample the minority class. The data points for the majority or minority classes remain unchanged. By interpolating between already existing minority class data points, it creates new data points. To achieve this, it generates points along the line segments that connect instances of the minority class that are randomly chosen. It chooses minority class instances at random to interpolate between. It makes no specific selections. It just selects neighbouring minority class instances at random. SMOTE's primary objective is to combine more minority class samples in order to balance the data.

Identification of False Transactions

The hazards of fraud, scams, hacking, and other unwanted cyber activities are rising as more corporate transactions, customer contacts, and sensitive data are conducted online. The static rules and models developed by human fraud analysts and specialists have traditionally been the foundation of traditional fraud detection systems. Machine learning techniques are used by more recent fraud detection systems to find patterns and anomalies in vast datasets. These computers outperform manual approaches in terms of speed and accuracy when it comes to detecting possibly fraudulent activities. They can also be easily modified to take into account shifting fraud trends and brand-new fraud kinds. Machine learning algorithms can also be used to produce alerts for additional research.

Statistical methods are used in anomaly detection approaches to find anomalies in the data and identify outliers. Algorithms for anomaly detection can be used to find possibly fraudulent activities and spot shady transactions. Data mining techniques can also be used to examine transaction patterns and find any odd or suspicious trends.

XGBoost

In XGboost, several sequential decision trees are used in a boosting strategy. By producing each new tree to enhance the prior trees, it seeks to lower the overall error. Each new tree determines how to split the data for a better categorization by using the updated residual errors from the previous trees. High predictive performance can be attained by XGboost thanks to this boosting method. The model put forth here uses an XGboost classifier to separate honest from dishonest transactions. A blockchain smart contract is linked with the XGboost model.

It assesses fresh incoming transactions and decides whether to allow them or mark them as harmful. The method improves the security and fraud detection capabilities for transactions in the blockchain network by integrating XGboost and blockchain smart contracts. To filter valid and fraudulent transactions at scale, XGboost offers a precise and reliable categorization technique.

Random Forest

It can be used to build several decision trees using various training data subsamples. As a result, the algorithm may concentrate on the minority class in each of the subsamples, allowing it to identify the minority class more precisely. Additionally, the algorithm can provide additional weight to the decision trees that correctly identify the minority class by giving each one a weight. This aids in lowering the model's overall bias and improving its accuracy. A method known as under-sampling, which involves choosing a portion of the majority class at random that is the same size as the minority class, can be used if we want to use the 0.001:0.999 ratio. This could improve and balance the dataset.

Proposed Model using Blockchain with Machine Learning

Blockchain technologies aren't always immune to fraud. Blockchains can still be subject to hostile assaults and fraudulent operations, despite the fact that they offer some helpful security and privacy features. The proposed approach tries to capitalize on the advantages of both technologies to tackle a specific problem in a novel way by integrating blockchain and machine learning. A more efficient and reliable solution might be offered by combining machine learning analytics with blockchain security and data transparency. The model is designed to find transaction data trends that can be utilized to spot fraudulent behaviour.

It is presumable that the pattern of Ethereum transactions and the pattern of bitcoin transactions recorded in the bitcoin transaction database are comparable. By "pattern," it is most likely referring to the properties, structure, and format of the transactions. For instance, similar information like sender and receiver addresses, transaction amount, transaction ID, etc., may be present in the transactions.

The machine learning model is trained and enhanced using each new Ethereum transaction as input. The new Ethereum transaction's pattern is examined and contrasted with the pattern of valid Bitcoin transactions. The Ethereum transaction is regarded as valid if the transaction patterns match. If not, it is marked as possibly harmful. By employing double-spending attacks on Ethereum, the system is put to the test to see how well it can identify malicious transaction

REFERENCES

- Staudemeyer R.C., Voyiatzis A.G., Moldovan G., Suppan S.R., Lioumpas A., Calvo D. *Human-Computer Interaction and Cybersecurity Handbook.* CRC Press; Boca Raton, FL, USA: 2018. Smart cities under attack.
- Podgorelec B., Turkanović M., Karakatič S. A machine learning-based method for automated blockchain transaction signing including personalized anomaly detection. *Sensors.* 2020;**20**:147.
- Nakamoto S. Bitcoin: A Peer-to-Peer Electronic Cash System. 2008.
- Farrugia S., Ellul J., Azzopardi G. Detection of illicit accounts over the Ethereum blockchain. *Expert Syst. Appl.* 2020;**150**:113318.
- Ostapowicz M., Żbikowski K. Detecting fraudulent accounts on blockchain: A supervised approach; Proceedings of the International Conference on Web Information Systems Engineering; Hong Kong, China. 19–22 January 2020; Cham, Switzerland: Springer; 2020. pp. 18–31
- Aziz A.S.A., Hassanien A.E., Azar A.T., Hanafy S.E. Genetic Algorithm with Different Feature Selection Techniques for Anomaly Detectors Generation; Proceedings of the 2013 Federated Conference on Computer Science and Information Systems (FedCSIS); Kraków, Poland. 8–11 September 2013
- Hassanien A.E., Tolba M., Azar A.T. *Communications in Computer and Information Science.* Volume 488. Springer; Berlin/Heidelberg, Germany: 2014. Advanced Machine Learning Technologies and Applications: Second International Conference, AMLTA 2014, Cairo, Egypt, 28–30 November 2014.
- Khan H., Asghar M.U., Asghar M.Z., Srivastava G., Maddikunta P.K.R., Gadekallu T.R. Fake review classification using supervised machine learning; Proceedings of the International Conference on Pattern Recognition; Virtual Event. 10–15 January 2021; Cham, Switzerland: Springer; 2021. pp. 269–288.
- Shahbazi Z., Hazra D.P., Park S., Byun Y.C. Toward Improving the Prediction Accuracy of Product Recommendation System Using Extreme Gradient Boosting and Encoding Approaches. *Symmetry.* 2020;**12**:1566.
- Pesantez-Narvaez J., Guillen M., Alcañiz M. Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. *Risks.* 2019;**7**:70.
- Li J., Gu C., Wei F., Chen X. A Survey on Blockchain Anomaly Detection Using Data Mining Techniques; Proceedings of the International Conference on Blockchain and Trustworthy Systems; Guangzhou, China. 7–8 December 2019; Singapore: Springer; 2019.
- Reid F., Harrigan M. *Security and Privacy in Social Networks.* Springer; New York, NY, USA: 2013. An analysis of anonymity in the bitcoin system; pp. 197–223
- Ngai E.W.T., Hu Y., Wong Y.H., Chen Y., Sun X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decis. Support Syst.* 2011;**50**:559–569.
- Saia R., Carta S. Evaluating Credit Card Transactions in the Frequency Domain for a Proactive Fraud Detection Approach; Proceedings of the 14th International Conference on Security and Cryptography (SECRYPT 2017); Madrid, Spain. 26–28 July 2017; pp. 335–342
- Sánchez D., Vila M.A., Cerda L., Serrano J.M. Association rules applied to credit card fraud detection. *Expert Syst. Appl.* 2009;**36**:3630–3640.
- Gyamfi N.K., Abdulai J.D. Bank fraud detection using support vector machine; Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON); Vancouver, BC, Canada. 1–3 November 2018; pp. 37–41

- Panigrahi S., Kundu A., Sural S., Majumdar A.K. Credit card fraud detection: A fusion approach using Dempster–Shafer theory and Bayesian learning. *Inf. Fusion.* 2009;**10**:354–363.
- Shi F.B., Sun X.Q., Gao J.H., Xu L., Shen H.W., Cheng X.Q. Anomaly detection in Bitcoin market via price return analysis. *PLoS ONE.* 2019;**14**:e0218341.
- Kumar P., Gupta G.P., Tripathi R. TP2SF: A Trustworthy Privacy-Preserving Secured Framework for sustainable smart cities by leveraging blockchain and machine learning. *J. Syst. Archit.* 2021;**115**:101954.
- Zhao Y., Tarus S.K., Yang L.T., Sun J., Ge Y., Wang J. Privacy-preserving clustering for big data in cyber-physical-social systems: Survey and perspectives. *Inf. Sci.* 2020;**515**:132–155
- Alkadi O., Moustafa N., Turnbull B., Choo K.K.R. A deep blockchain framework-enabled collaborative intrusion detection for protecting IoT and cloud networks. *IEEE Internet Things J.* 2020;**8**:9463–9472.
- AlKadi O., Moustafa N., Turnbull B., Choo K.K.R. Mixture localization-based outliers models for securing data migration in cloud centers. *IEEE Access.* 2019;**7**:114607–114618.
- Keshk M., Sitnikova E., Moustafa N., Hu J., Khalil I. An integrated framework for privacy-preserving based anomaly detection for cyber-physical systems. *IEEE Trans. Sustain. Comput.* 2019;**6**:66–79.
- Kurakin A., Goodfellow I., Bengio S. Adversarial machine learning at scale. *arXiv.* 20161611.01236
- Biggio B., Roli F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognit.* 2018;**84**:317–331.
- Xuan S., Liu G., Li Z., Zheng L., Wang S., Jiang C. Random forest for credit card fraud detection; Proceedings of the 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC); Zhuhai, China. 27–29 March 2018; pp. 1–6
- Liu C., Chan Y., Alam Kazmi S.H., Fu H. Financial fraud detection model: Based on random forest. *Int. J. Econ. Financ.* 2015;**7**:178–188.
- Apruzzese G., Andreolini M., Colajanni M., Marchetti M. Hardening random forest cyber detectors against adversarial attacks. *IEEE Trans. Emerg. Top. Comput. Intell.* 2020;**4**:427–439. 29. Primartha R., Tama B.A. Anomaly detection using random forest: A performance revisited; Proceedings of the 2017 International Conference on Data and Software Engineering (ICoDSE); Palembang, Indonesia. 1–2 November 2017; pp. 1–6.
- Laskov P. Practical evasion of a learning-based classifier: A case study; Proceedings of the 2014 IEEE Symposium on Security and Privacy; San Jose, CA, USA. 18–21 May 2014; pp. 197–211.
- Pham T., Lee S. Anomaly detection in bitcoin network using unsupervised learning methods. *arXiv.* 20161611.03941
- Martin K., Rahouti M., Ayyash M., Alsmadi I. Anomaly detection in blockchain using network representation and machine learning. *Secur. Priv.* 2022;**5**:e192.
- Pinzón C., Rocha C. Double-spend attack models with time advantange for bitcoin. *Electron. Notes Theor. Comput. Sci.* 2016;**329**:79–103.
- Bitcoin Network Transactional Metadata. [(accessed on 12 September 2022)].
- Shafiq O. *Master's Thesis.* Tampere University; Tampere, Finland: 2019. Anomaly Detection in Blockchain.
- Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 2002;**16**:321–357.
- Sadaf K., Sultana J. Intrusion detection based on autoencoder and isolation Forest in fog computing. *IEEE Access.* 2020;**8**:167059–167068.

➢ Eyal I., Sirer E.G. Majority is not enough: Bitcoin Mining is vulnerable; Proceedings of the International Conference on Financial Cryptography and Data Security; Christ Church, Barbados. 3–7 March 2014; Berlin/Heidelberg, Germany: Springer; 2014. pp. 436–454.

➢ Landa R., Griffin D., Clegg R.G., Mykoniati E., Rio M. A Sybilproof indirect reciprocity mechanism for peer-to-peer networks; Proceedings of the IEEE INFOCOM 2009, Rio De Janeiro; Brazil. 24 April 2009; pp. 343–351

➢ Luu L., Chu D.-H., Olickel H., Saxena P., Hobor A. Making smart contracts smarter; Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; Vienna, Austria. 24–28 October 2016.

➢ Nizamuddin N., Hasan H., Salah K., Iqbal R. Blockchain-based framework for protecting author royalty of digital assets. *Arab. J. Sci. Eng.* 2019;**44**:3849–3866.

➢ Halo Block, Medium How To Use Oyente, a Smart Contract Security Analyzer—Solidity Tutorial. 2020.