ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 9, Issue 3, April 2024

Analyzing Ensemble Learning Techniques in Fraud Detection of Crypto-Currency

[1] A.Jeeva(M.E), [2]Abhiyanshu Choudhary, [3]R. Kabilesh, [4]M. Keerthivasan, [5]S. Kowshikkumar
[1] Assistant Professor, Department of Artificial Intelligence and Data Science,Gnanamani College of Technology, Namakkal, Tamilnadu, India.
jeeva@gct.org.in
[2] [3] [4] [5] Student, Department of Artificial Intelligence and Data Science,Gnanamani College of Technology, Namakkal,Tamilnadu, India.
cabhiyanshu@gmail.com, kabileshramamoorthy2002@gmail.com, keerthi74mkp@gmail.com, skowshikkumar0@gmail.com

Abstract: ***Cryptocurrency ecosystems face increasing challenges associated with fraudulent activities, necessitating sophisticated solutions for timely detection and prevention. The heuristic and mark based approaches were the underpinning of before location strategies, yet unfortunately, these techniques were lacking to investigate the whole intricacy of peculiarity identification. With this contextual understanding, isn't it true that prevention (avoiding sending crypto to scammers) is a better option than cure (retrieving sent crypto). The proposed system integrates multiple machine learning models into an ensemble, harnessing the collective intelligence of diverse algorithms for improved accuracy and robustness in identifying fraudulent patterns. Experimental evaluations, conducted on historical data and simulated scenarios, demonstrate the effectiveness of the proposed ensemble learning framework in bolstering the security and trustworthiness of crypto currency transactions. This is why the aim of this project is to develop an application whereby users can check whether an address is a potential fraudulent one before making the irreversible decision of sending crypto over to that address. This will be done by developing machine learning models (Ensemble learning) which use common attributes of crypto addresses to make calculated decisions on whether an address might be potentially fraudulent.***

Keywords—Ensemble, Machine Learning, Machine Learning Algorithms, crypto currency transaction.

I. INTRODUCTION

For a very long time, attempts to identify dishonest financial dealings have been studied. Fraudulent transactions discourage potential Bitcoin investors and others from putting their faith in blockchain technology. They also hurt the economy. Most of the time, fraud is caused by something about the transaction's nature or the people involved. To ensure that the community and the network's integrity are not compromised, members of a blockchain network strive to quickly identify fraudulent transactions. A decentralized and distributed ledger that records transactions safely and publicly is called a blockchain. Each SNI in chain contains a series of transactions that have been confirmed and accepted by the network. Without network consensus, a block cannot be modified or deleted once it has been added to the chain. [8]

The following are some unlawful activities connected to Bitcoin and other cryptocurrencies taking place in smart cities. Money laundering: criminals can move illegal funds undetectably across borders using Bitcoin. Dark web transactions: Bitcoin is used to pay for criminal operations including selling of guns or drugs on the dark web because of its anonymity. Ransomware payments: hackers and online criminals utilize Bitcoin to pay for ransomware attacks, in which they demand money in return for access to the victim's computer or data.

The motivation for this work comes from the fact that the existing methods were insufficient to explore the entire complexity of anomaly detection. ML can play an essential role in anomaly detection, as it can learn from historical data and detect new unseen attacks. In this paper, we address the limitations of the existing techniques and present a various ensemble model by combining multiple ML techniques: Random Forest (RF), Decision Tree (DT), Naive Bayes (NB), and K-nearest Neighbors (KNN), etc.

II. LITERATURE SURVEY

H. H. Sun Yin et al. [12] proposed supervised machine learning-based anomaly and criminal activity detection in a Bitcoin-based ecosystem. The proposed solution is based on a dataset of 395 million transactions with 957 unique clusters. The study also compares the performance of Random Forests (RF), Decision Trees (DT), K-Nearest Neighbour (KNN), Gradient Boosting, Ada Boosting, and Bagging classifiers for performance detection of anomalies in Bitcoin-based systems. The Gradient Boosting is the most accurate out of the seven tested models, with 80.83% accuracy. S. Sayadi [14] proposed a solution to detect fraudulent transactions in cryptocurrency and proposed a technique to determine anomalies in electronic

transactions of Bitcoin by machine learning, with high accuracy with k-mean and SVM. B. Chen et al. [15] proposed machine learning-assisted solutions to detect Bitcoin theft transactions. The performance of five machine learning models (KNN, SVM, RF, AdaBoost, and MLP) is evaluated to identify the theft transactions in Bitcoin. The study results reveal that the RF performs best, with an F1-value of 95.9%. Kasera [16] proposed an artificial intelligence-based approach for identifying fraud in cryptocurrency. This study focuses on how artificial intelligence provides us with an empirical framework to identify such frauds to ensure more security in the cryptosphere. Researchers have been focusing on determining an efficient fraud and security threat detection model that overcomes the drawbacks of existing methods. The Light Gradient Boosting Model (LGBM) algorithm is proposed for detecting fraudulent transactions performed in the Ethereum blockchain [17], [18]. The authors first in the names of the dataset columns can be found in the first row of the screen, and the values of the dataset can be found in the rows that follow. In the dataset, there is a column called FLAG, and its values range from 0 to 1, with 0 representing a normal transaction and 1 representing a fraudulent one.

III. METHODOLOGY

A user's participation in a Blockchain transaction does not guarantee that he will not commit fraud, which can hurt the economy of any country, even though Blockchain is thought to be secure against attacks due to its proof of work and transaction validation using hash code. To evaluate these tools' efficacy, the author of this work employs a wide range of machine learning techniques, including Define the Rotation ForestClassifier, Ensemble of classifiers based on diverse classifiers (ECDC), CatBoost ensemble learning algorithms, Isolation Forest Ensemble Technique, Voting Classifiers Ensemble Technique. To compose this article, all client and exchange data was assembled from the Blockchain extortion exchange dataset. The dataset was then processed to normalize values, replace missing values with central tendency, and remove all non-numerical data. The following are different ensemble learning algorithms:

A) Rotation Forest Classifier,
B) Ensemble of classifiers based on diverse classifiers (ECDC),
C) CatBoost ensemble learning algorithms,
D) Isolation Forest Ensemble Technique,
E) Voting Classifiers Ensemble Technique.

A) Rotation Forest Classifier:

Rotation Forest is an ensemble learning algorithm that extends the traditional Random Forest algorithm by incorporating Principal Component Analysis (PCA) into the training of individual decision trees. The goal is to enhance diversity among base learners and improve generalization performance. Here's how Rotation Forest works are Feature Subset Selection, Principal Component Analysis (PCA), Transformation, Train Decision Tree, Repeat, Ensemble Aggregation.

Formulas:

Let Xsi be the data matrix containing only the features in the subset Si. The means can be numerically addressed as follows:

1. Covariance Matrix: Csi = (XtsiXsi)/n-1
2. Eigenvalues and Eigenvectors: Csivj = λjvj
3. Select First k Eigenvectors: Pi = [v1, v2, …, vk]
4. Transformation: Xrotsi = XsiPi

Where Xrotsi is the transformed feature set.

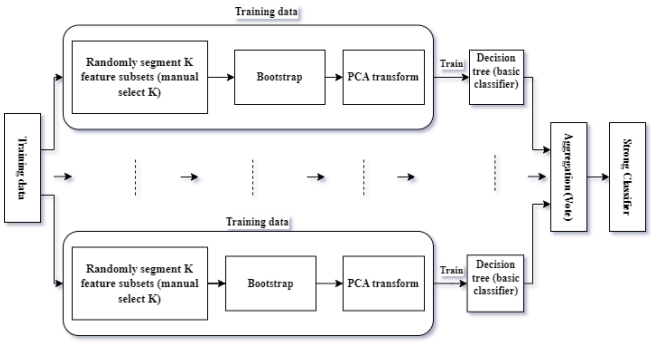

Fig 1: Rotation Forest Classifier

B) Ensemble of classifiers based on diverse classifiers (ECDC):

Ensemble learning involves combining the predictions of multiple individual models (classifiers or regressors) to create a stronger, more robust model. The key idea is that by aggregating the opinions of diverse models, the ensemble can often achieve better performance than any individual model alone. In the context of classification, an ensemble of diverse classifiers can be particularly effective. Here's an explanation of the concept of an ensemble of diverse classifiers.
Diverse Classifiers: Diversity is a crucial aspect of ensemble learning. The individual classifiers in the ensemble should be different from each other in terms of their learning algorithms, input features, or training data. Diversity helps ensure that errors made by one classifier are compensated by correct predictions from others, leading to a more balanced and accurate overall prediction.
Ensemble Methods: Bagging (Bootstrap Aggregating), Boosting, Stacking.
Voting or Averaging: Ensemble methods typically combine predictions through voting or averaging. In classification, voting can be either soft or hard. Hard Voting: The final prediction is based on a simple majority vote among the individual classifiers. Soft Voting: Each classifier assigns a probability to each class, and the final prediction is based on the weighted average of these probabilities.
Benefits of Diversity: Diverse classifiers are less likely to make the same errors on unseen data, which helps improve the generalization performance of the ensemble. Diversity ensures that the ensemble is not overly sensitive to the idiosyncrasies of a particular model or subset of the data.

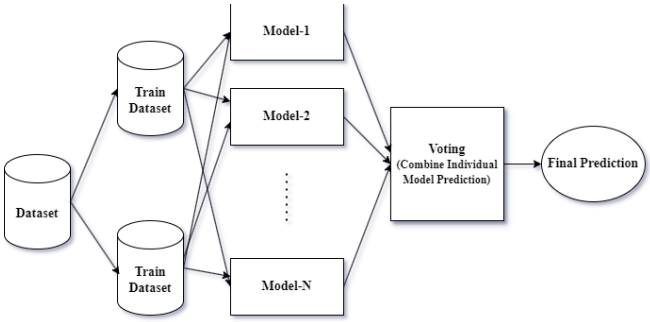

Fig 2: Ensemble of classifiers based on diverse classifiers ECDC)

C) CatBoost ensemble learning algorithms:

CatBoost is a gradient-boosting algorithm designed for categorical feature support. It is an ensemble learning method that builds an additive model in a forward stage-wise manner.
The formula for the final prediction in CatBoost can be expressed as follows:
Prediction Formula:

1. $F(x)=\text{offset}+\sum_{t=1}^{T} \eta \cdot f_t(x)$

Here:

2. $F(x)$ is the last expectation for input x.

offset is an optional constant term that can be added to the final prediction.

3. T is the quantity of trees in the group.

η is the learning rate, a hyper parameter that controls the contribution of each tree to the final prediction.

4. $f_t(x)$ is the prediction of the t-th tree for input x.

Tree Prediction: The prediction of each tree in CatBoost is obtained by traversing the tree structure. For a given input x, the prediction $f_t(x)$ is obtained by following the decision rules in the t-th tree.
Objective function: CatBoost minimizes a specific objective function during training. The objective function consists of two parts: the loss function and a regularization term.

5. Objective=Loss+Regularization

The loss function measures the difference between the predicted values and the actual target values. The regularization term helps prevent overfitting.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 9, Issue 3, April 2024

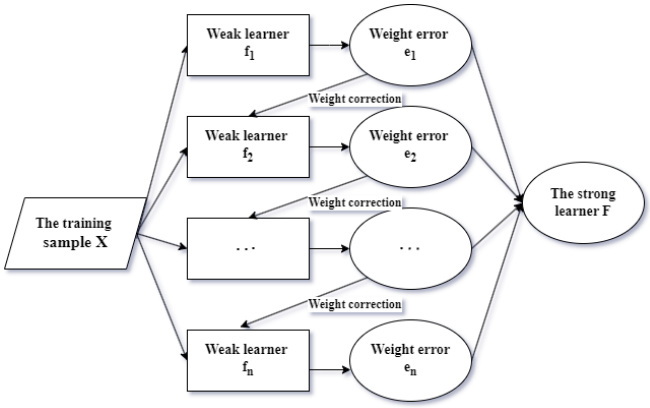

Fig 3: CatBoost ensemble learning algorithms

A) Isolation Forest Ensemble Technique:

The Isolation Forest algorithm is an unsupervised ensemble technique used for anomaly detection. It isolates anomalies by randomly selecting a feature and then randomly selecting a split value between the minimum and maximum values of that feature. The process is repeated recursively until the anomalies are isolated. The formula for the isolation score, which measures how easily a data point can be isolated, is key to understanding how Isolation Forest works.

Isolation Score Formula:

For each data point x, the isolation score s(x,n) is calculated as follows:

1. $S(x, n) = 2^{-E(h(x))/c(n)}$

Here:

2. h(x) is the path length of data point x in the tree.

E(h(x)) is the expected path length for a given h(x) in a random sample of the same size.

3. c(n) is a normalization term defined as $2 \cdot \ln(n-1) - 2(n-1)/n$.

Anomaly Score: The anomaly score A (x, n) is calculated by normalizing the isolation score:

4. $A(x, n) = s(x, n)/ c(n)$

Decision Function: A threshold is defined, and data points with anomaly scores above the threshold are considered anomalies. Decision Function: if A (x, n)>threshold, then x is an anomaly. The Isolation Forest algorithm builds an ensemble of isolation trees, each constructed using a random subset of the data. The final anomaly score for a data point is the average anomaly score across all trees.

Ensemble Score:

6. ensemble $(x, n) = \sum_{i=1}^{n} Ai (x, n)/ n$

Here:

i. N is the quantity of trees in the troupe.

Ai (x, n) is the anomaly score for data point x in the i-th tree.

ii. Data points with higher ensemble anomaly scores are more likely to be anomalies.

The Isolation Forest algorithm efficiently isolates anomalies by leveraging the fact that anomalies are typically fewer and have different characteristics than normal instances in a dataset. The recursive partitioning of the feature space allows for quick isolation of anomalies.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 9, Issue 3, April 2024

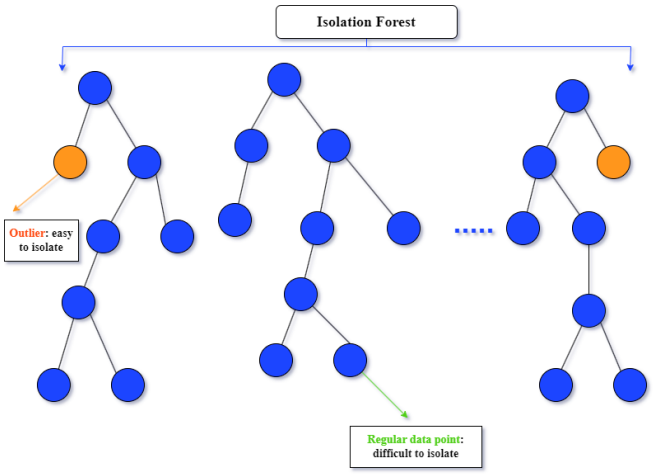

Fig 4: Isolation Forest Ensemble Technique

E) Voting Classifiers Ensemble Technique:

A Democratic Classifier is a group procedure that joins the expectations of numerous singular models to make a last forecast. It can operate in two modes: hard voting and soft voting. A voting classifier is an ensemble technique in machine learning where multiple models are combined to make predictions. Each model gets a vote, and the final prediction is determined by the majority vote (for classification tasks) or by averaging (for regression tasks). There are various sorts of casting a ballot classifier:

Hard Voting: In hard voting, each model in the ensemble predicts the class label, and the majority class label is chosen as the final prediction.
Soft Voting: In soft voting, each model predicts the probability of each class, and the average probabilities across all models are calculated. The class with the most elevated typical likelihood is picked as the last forecast.
Voting classifiers can be composed of various types of base models, such as decision trees, support vector machines, logistic regression, etc. The idea is that by combining the predictions of multiple models, the ensemble can achieve better performance than any individual model. It's a form of wisdom of the crowd approach in machine learning.

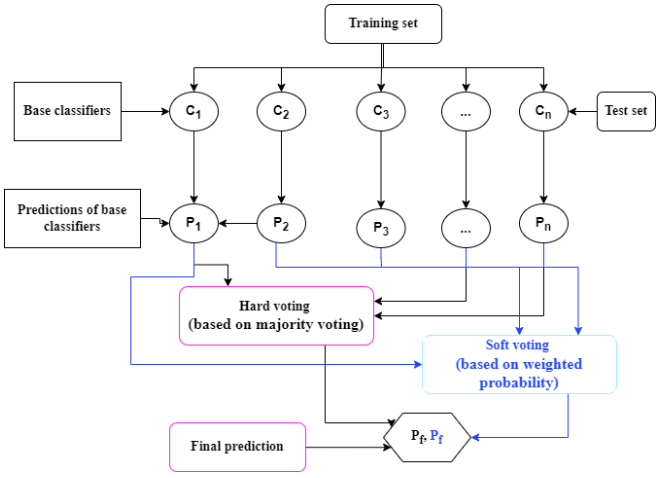

Fig 5: Voting Classifiers Ensemble Technique

IV. WORK FLOW

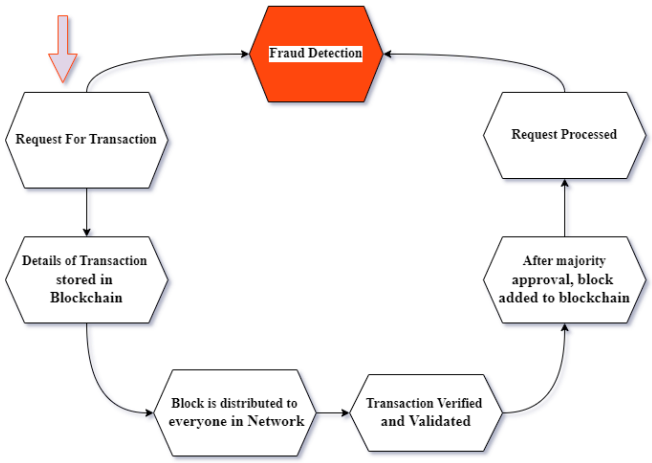

Fig 6: Workflow Diagram

V. RESULTS AND DISCUSSION

In the figure 8 given below, the accuracy of different ensemble learning techniques is evaluated.

s.no	Ensemble algorithms	Accuracy (%)
1.	Rotation Forest Classifier	96.80
2.	Ensemble of classifiers based on diverse classifiers (ECDC)	97.15
3.	CatBoost ensemble learning algorithms	94.26
4.	Isolation Forest Ensemble Technique	75.77
5.	Voting Classifiers Ensemble Technique	65.41

Therefore, the Ensemble of classifiers based on diverse classifiers (ECDC) has achieved the highest accuracy of 97.15.

ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 9, Issue 3, April 2024

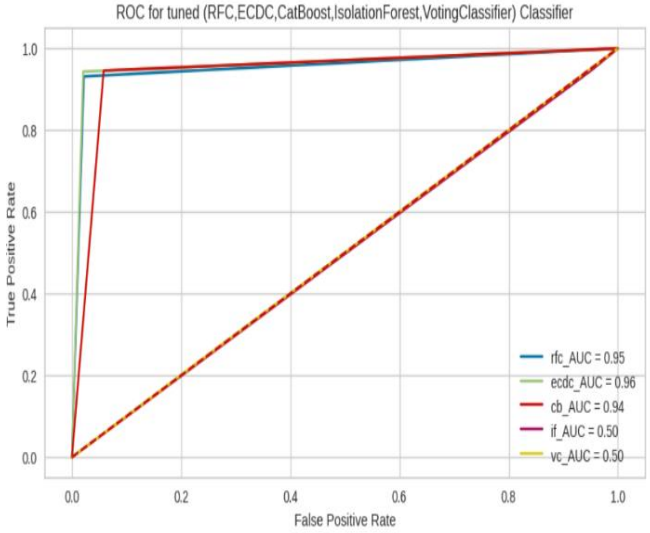

Fig 7: ROC Curve of Best Model

VI. CONCLUSION

A technique for recognizing suspicious blockchain transactions has been presented using machine learning. This method looked at the Rotation Forest Classifier, Ensemble of classifiers based on diverse classifiers (ECDC), CatBoost ensemble learning algorithms, Isolation Forest Ensemble Technique, and Voting Classifiers Ensemble Technique. A comprehensive evaluation and comparison of all available methods is carried out using accuracy. this work could be expanded to include a comparison of various ensemble learning and other supervised algorithms. Majority of these crypto scams are either investment-related, romance manipulations, or someone acting as a business and government imposter, and their end goal is usually to get the victims to send crypto over to them through the blockchain, which being decentralized, means that victims can never recover any crypto they've sent.

VII. REFERENCES

[1] S. Giribabu, V. Sriharsha. "Analyzing various machine learning algorithms for bockchain based fraud detection". May(2022)
[2] Patrick M., Vukosi, Bhesipho. "A Multifaceted approch to bitcoin fraud detection global and local outliers." (2016)
[3] Patrick M., Vukosi, Bheki Twala. "Unsupervised learning for robust bitcoin fraud detection." (2016)
[4] Pranav, Yann, Romaric, Kunjal, Sunil, Dhiren. "Detecting illicit entities in bitcoin using supervised learning of ensemble decision trees." (27 November 2020)
[5] Farrugia S, Ellul J, Azzopardi G. Detection of illicit accounts over the Ethereum blockchain. Expert Systems with Applications
[6] Qasim Umer, Jian-Weili, Rab Nawaj. "Ensemble deep learning based prediction of fraudulent cryptocurrency transactions." (2023)
[7] Vishvesh Pathak. "Ensemble learning based social engineering fraud detection module for cryptocurrency transactions." (2023)
[8] Noor, Nadeem Javaid. "A new framework for fraud detection in bitcoin transactions through ensemble stacking model in smart cities" (2016)
[9] Stefansson, Hilmar. "Detecting potential money laundering addresses in the bitcoin blockchainusing unsupervised machine learning." (2022)
[10] Binjie Chen, Fushan, Chunxiang Gu. "Bitcoin theft detection based on supervised ML algorithms." (2021)
[11] Han Wei Lun, Austin Loh, Jotham Wong. "Data crypto fraud detection." (2023)
[12] H. H. Sun Yin, K. Langenheldt, M. Harlev, R. R. Mukkamala, and R. Vatrapu, "Regulating cryptocurrencies: a supervised machine learning approach to de-anonymizing the bitcoin blockchain," Journal of Management Information Systems, vol. 36, no. 1, pp. 37–73, 2019.

[13] "Coinbase - Buy Sell Bitcoin, Ethereum, and more with trust," accessed on: 22-Jul.-2022. [Online]. Available: https://www.coinbase.com/
[14] S. Sayadi, S. B. Rejeb, and Z. Choukair, "Anomaly detection model over blockchain electronic transactions," in 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC). IEEE, 2019, pp. 895–900.
[15] B. Chen, F. Wei, and C. Gu, "Bitcoin theft detection based on supervised machine learning algorithms," Security and Communication Networks, vol. 2021, 2021.
[16] A. Kasera, "Cryptocurrency frauds," International Journal of Engineering and Advanced Technology (IJEAT), vol. 9, no. 6, pp. 261–268, Aug. 2020. [Online]. Available: https://doi.org/10.35940/ijeat.F1391.089620
[17] R. M. Aziz, M. F. Baluch, S. Patel, and A. H. Ganie, "LGBM: a machine learning approach for Ethereum fraud detection," Int. J. Inf. Technol., vol. 14, no. 7, pp. 3321–3331, 2022, doi: 10.1007/s41870- 022-00864-6.
[18] L. Pahuja and A. Kamal, "EnLEFD-DM: Ensemble Learning based Ethereum Fraud Detection using CRISP-DM framework," Expert Syst., pp. 1–18, 2023, doi: 10.1111/exsy.133