Anomaly Detection on Bitcoin, Ethereum Networks Using GPU-accelerated Machine Learning Methods

Youssef Elmougy
Georgia Institute of Technology
Georgia, USA
yelmougy3@gatech.edu

Oliver Manzi
Luleå University of Technology
Luleå, Sweden
oliman-8@ltu.se

Abstract—**Blockchain technology is continually gaining momentum, with applications expanding in sectors beyond digital assets and financial services. With the existence of a public distributed ledger, the validity of transactions and accounts on the blockchain can be easily reviewed. Nevertheless, there are malicious persons that attempt to fraud cryptocurrency holders, undermining the reliability of the blockchain. This study focuses on identifying fraudulent transactions and accounts by detecting anomalies in the Bitcoin and the Ethereum transaction networks, the two largest cryptocurrencies. By leveraging GPU-accelerated machine learning models, including Support Vector Machines, Random Forest, and Logistic Regression, we draw the metadata of over 30 million transactions on the Bitcoin network and confirmed transactions from over 500 thousand accounts on the Ethereum network. We offer insight into feature importance through sensitivity analysis, as well as train accurate models that allow for method adoption in automated fraud detection systems. The trained models achieve an accuracy and recall of 96.9% and 0.987 on the Bitcoin dataset, and 80.2% and 0.835 on the Ethereum dataset. The study of anomaly detection in the cryptocurrency blockchain done in this paper can be generalized to other blockchain networks, including health service blockchains, public sector blockchains, and financial intelligence blockchains.**

Index Terms—**Cryptocurrency; Blockchain; Bitcoin; Ethereum; Anomaly Detection; Fraud; Machine Learning; GPU-acceleration; SVM, Random Forest; Logistic Regression**

I. Introduction

Cryptocurrencies are becoming increasingly prominent, being largely utilized as an investment platform due to its implementation of a fundamentally new technology, the Blockchain. These digital currencies draw keen interest towards cryptography-enabled payments that utilize digital signatures and hash functions, and distributed transaction retention (the Blockchain). The cryptocurrency market is constantly expanding, with a current market capitalization of 2.42 trillion USD[1]. Since the first public blockchain, Bitcoin [1], other blockchains have been created such as Ethereum [2]. Advancements made possible through blockchain are expected to expand in sectors beyond financial services in the near future [3], [4].

With the immense amounts of fiat money invested and traded into cryptocurrencies, it has attracted scammers that attempt to fraud cryptocurrency holders. In 2015, Bitcoin-based scams defrauded 13,000 victims and stole over 11 million USD [5]. Conversely to traditional financial networks, there exists a distributed ledger that stores all transaction details for public reference. Once a transaction, or an account (wallet address), has been deemed fraudulent, it is shared with the public, allowing current and potential cryptocurrency investors to review the validity of the account with whom they are transacting with.

Bitcoin and Ethereum currently hold a share of 64.5%[1] of the entire cryptocurrency market capitalization. Thus, it is important to analyze the transactions and accounts on both networks in order to provide a useful basis of reference for the two largest cryptocurrencies.

This paper focuses on identifying fraudulent transactions and accounts by detecting anomalies in the Bitcoin and Ethereum transaction networks. By drawing datasets consisting of 30,294,698 transactions on the Bitcoin network and confirmed transactions from 504,942 wallets on the Ethereum network, the paper uses anomaly detection-based approaches and trains machine learning algorithms including Support Vector Machine (SVM), Random Forest, and Logistic Regression. The authors' major contributions in this paper is fourfold. First, the paper is the first to systematically analyze and accurately detect anomalies on both the Bitcoin and the Ethereum networks synchronically. Second, GPU-accelerated machine learning algorithms were developed and deployed to allow for the analysis of datasets consisting of millions of transactions. Third, thorough sensitivity analysis is done to show the relationships and correlation among model features and their importance towards anomaly detection on these networks. Lastly, as mentioned by Farren et al. [6], the study of anomaly detection in the cryptocurrency blockchain can be generalized to other blockchain networks. Thus, the investigations done in this paper can be applicable to other blockchain networks including health service blockchains, public sector blockchains, and more. Moreover, the models trained in this study can be adopted to automated fraud detection systems.

The remainder of the paper is organized as follows. Section II discusses the related works. Section III provides a background to the machine learning models that will be trained. Section IV introduces the dataset collection and preparation

[1] As of October 14^{th} 2021 on Coinmarketcap.com.

methods, feature extraction and sensitivity analysis, and the model verification techniques. Section V presents the results of the machine learning models on each dataset, and evaluates their effectiveness, and Section VI concludes the paper.

II. Related Work

Previous studies have been conducted to identify anomalous transactions in the blockchain. Pham et al. [7] studied anomaly detection in the Bitcoin blockchain from a network perspective, modelling both a user graph and a transaction graph. Each vertex in the graphs were represented by 12 features, and the laws of power degree & densification as well as local outlier factor method were applied to the graphs. Ostapowicz et al. [8] used supervised methods, such as Random Forests, SVM, and XGBoost, to detect fraudulent accounts on the Ethereum blockchain. The study was able to achive recall and precision values to design an anti-fraud rule for digital wallets or currency exchanges. Baek et al. [9] create a model for detecting transactions with discernible purpose on Binance, utilizing the EM algorithm for the Gaussian mix model. Using feature engineering, the study trained the random forest algorithm to label suspicious wallets with high precision.

There have been additional studies done regarding existing scams in the cryptocurrency field, such as the classic Pump and Dump scheme [10]. These studies have the aim of defining existing and emerging fraud schemes in the cryptocurrency field in order to allow for a more efficient resource allocation of fraud detection systems.

The study done in this paper trains machine learning models concurrently on the Bitcoin and Ethereum networks, presenting a side-by-side evaluation and comparison of the networks, rather than an evaluation of a single network. Leveraging GPU-acceleration, the datasets utilized in this study consisted of millions of transactions, covering a wider set of data points compared to previous studies. Consequently, the trained models in this study are able to reach higher accuracy and recall values than previous studies.

III. Machine Learning Models

Anomalous transaction detection is a natural binary classification problem, where the model aims to classify a transaction as either *fraudulent* or *non-fraudulent*. The machine learning models that will be trained in this study include *Support Vector Machine* (SVM), *Random Forest*, and *Logistic Regression*. In this section, we define these machine learning models as well as express their optimization mathematically.

A. Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm [11] constructs a hyperplane or a set of hyperplanes in an *n*-dimensional space, where *n* is the number of features, so as to find the optimal separating hyperplane with the maximum margin (distance between the hyperplane and the training samples closest to the hyperplane) that uniquely classifies the data and lowers the generalization error of the classifier. It is a supervised learning method that is used for classification, regression, and outlier detection that is effective in high dimensional spaces. Let's assume that (x_i, y_i) is a pair of data points in the dataset D, where x_i is a vector that contains attributes of the i^{th} data point and y_i is the class label which satisfies $y_i \in \{-1, +1\}$. The hyperplane to classify a data point x is expressed as $y = w^T x + b$. The following is the optimization problem of the SVM model:

$$\min_{w,b} \frac{1}{2}||w||^2$$

$$s.t., y_i(w^T x_i + b) \geq 1, i = 1, 2, ..., m$$

B. Random Forest

The Random Forest algorithm [12] consists of an ensemble of individual classification trees, which are trained on bootstrapped samples from the original dataset while using random subsets of features for each decision. The final decision (shown below) of Random Forest models is the average of classification predictions from all decision trees.

$$C_{RF} \leftarrow majorityVote\{C_i(x)\}_1^n$$

Where $C_i(x)$ is the predicted classification of the i^{th} random tree.

C. Logistic Regression

The Logistic Regression algorithm [13] is a single layer neural network with a binary response variable. Given our binary classification problem, Logistic Regression will assign probabilities to each row of the feature matrix D with the sample size N. The Logistic Regression minimizes the following optimization problem:

$$\min_{w,c} \frac{w^T w}{2} + C \sum_{i=1}^{N} \log(\exp(-y_i(x_i^T w + c)) + 1)$$

Where we have a set of d features, $x = (x_1, ..., x_d)$, parameter vector w, and optimal value C calculated via cross validation.

IV. Methodology

A. Dataset Collection and Preparation

This study aims to identify fraudulent transactions on both the Bitcoin and Ethereum networks. In the Bitcoin blockchain analysis, raw data was scraped from the Bitcoin transactions available on the public ledger. This public ledger contains all Bitcoin transactions from the date of inception to present time, with 678,012,452 transactions so far on the ledger[2]. Due to our ability to use GPU acceleration, our dataset consisted of 30,294,698 transactions: 30,290,045 of which were non-fraudulent transactions, and 4,653 of which were fraudulent transactions.

Ethereum offers two types of accounts (wallets): contract and externally owned. In the Ethereum blockchain analysis,

[2] As of October 14^{th} 2021 on Blockchain.com.

TABLE I
MODEL FEATURES.

Feature	Description
IN-TXS	Number of incoming transactions
OUT-TXS	Number of outgoing transactions
IN-BTC / IN-ETH	Amount (in Bitcoin/Ether) on incoming transactions
OUT-BTC / OUT-ETH	Amount (in Bitcoin/Ether) on outgoing transactions
AVG-IN	Average amount (in Bitcoin/Ether) on incoming transactions
AVG-OUT	Average amount (in Bitcoin/Ether) on outgoing transactions
TOTAL-BTC / TOTAL-ETH	Total amount (in Bitcoin/Ether) on all incoming and outgoing transactions
FRAUD	Fraud boolean classifier

raw data was scraped from the Ethereum blockchain browser Etherscan.io. To compile the dataset, confirmed transactions interacting with all 4,942 wallets[3] (both contract and externally owned) tagged in fraudulent transactions were scraped using the Etherscan API, as well as confirmed transactions from 500,000 unflagged wallets.

B. Feature Extraction: Sensitivity Analysis

In order to effectively apply machine learning models to detect anomalous transactions, feature extraction must occur. In this section, we extract the set of features in our dataset as well as evaluate the feature importance through sensitivity analysis. Table I shows the features used in both the Bitcoin and Ethereum datasets.

Exploratory analysis of the datasets assist in understanding the data and the relationship among the dimensions in-depth. This analysis will detect underlying class imbalances, which will guide our data transformation and model verification, as well as feature correlation. There exists a high imbalance of fraudulent and non-fraudulent transactions in both datasets due to the low public availability of fraudulent data on the blockchain. Moreover, the data followed a right-skewed distribution, hence a $\log(x+1)$ transformation was applied. To solve this class imbalance and the skew distribution, normalization and standardization transformations were applied, to which the features became less imbalanced and skewed. This adjustment increases the performance accuracy of the machine learning models.

We use multiple pairwise bi-variate distribution plots to assist in data distribution exploration and visualize the dimension relationships. Figure 1 shows the pair plot for the Bitcoin dataset, where there are relationship plots for each x-axis variable and y-axis variable pairs. The diagonal of the pair plot shows a histogram depicting the distribution for each variable. The blue points represent non-fraudulent transactions, whereas the orange points represent fraudulent transactions. It can be seen that the pattern of fraudulent transactions is greatly non-linear.

Based on Figure 1, it can be seen that `IN-TXS` and `OUT-TXS` feature a negative correlation with maliciousness, thus the lower the number of incoming and outgoing transactions, the higher the probability of a transaction being classified as fraudulent. It can also be seen that `IN-BTC/IN-ETH`, `OUT-BTC/OUT-ETH`, `AVG-IN`, `AVG-OUT`, and `TOTAL-BTC/TOTAL-ETH` feature a positive correlation, thus the higher the amount (in Bitcoin/Ether) on incoming and outgoing transactions as well the total amount (in Bitcoin/Ether) on all incoming and outgoing transactions, the higher the probability of a transaction being classified as fraudulent. This sensitivity analysis is valuable to further discussions as it reveals the non-linearity of the dataset, which will require complex representations to train a machine learning model.

C. Model Verification Techniques

It is important to verify the accuracy of our models in order to evaluate their effectiveness in detecting fraud transactions on the blockchain. The following metrics were used:

1) *Confusion Matrix* - this gives an overall view of classifiers. It has components: true positive (TP), false positive (FP), false negative (FN), and true negative (TN).

TABLE II
CONFUSION MATRIX.

		Actual	
		Non-Fraudulent	Fraudulent
Predicted	Non-Fraudulent	TP	FP
	Fraudulent	FN	TN

2) *Precision* - this is the fraction of positive points that were correctly classified, calculated by:

$$Precision = \frac{tp}{tp + fp}$$

3) *Recall* - this is the fraction of actual positive points that were correctly classified, calculated by:

$$Recall = \frac{tp}{tp + fn}$$

4) *Accuracy* - this measures the ratio between the correctly predicted observations and total observations, calculated by:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn}$$

5) *F1 Score* - this is the harmonic mean of precision and recall, calculated by:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

[3] As of October 14th 2021 on Etherscan.io.

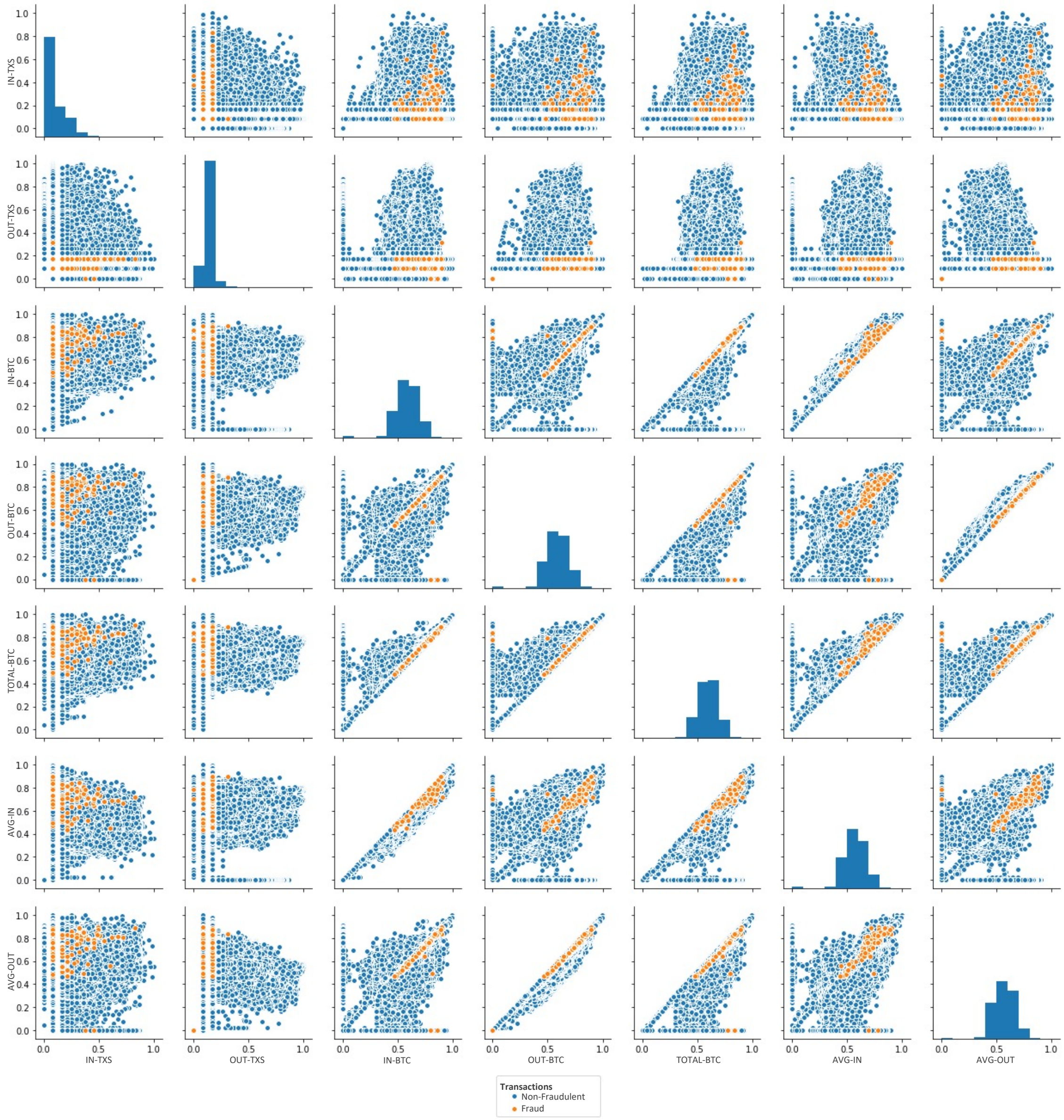


Fig. 1. Pairwise bi-variate distribution plots and distribution histograms for the Bitcoin dataset.

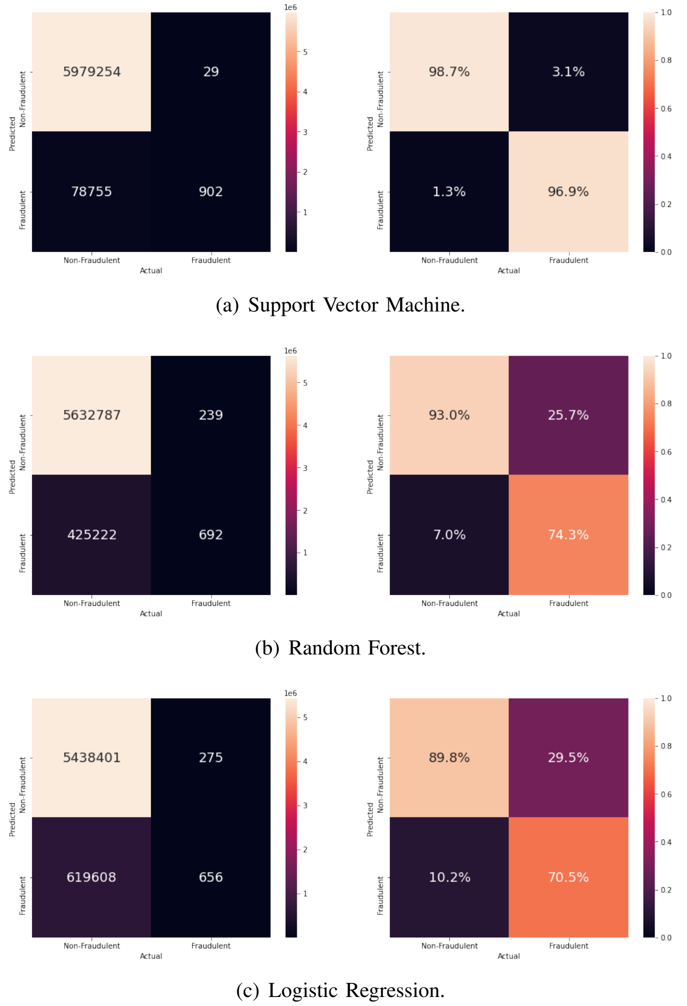
V. Results Evaluation

We used an effective 80/20 split of training and testing data. Specifically, 6,058,009 non-fraudulent and 931 fraudulent transactions were used for testing from the Bitcoin dataset, whereas 100,000 non-fraudulent and 988 fraudulent accounts were used for testing from the Ethereum dataset. Each machine learning model was trained, and the resulting Confusion Matrix, Precision, Recall, Accuracy, and F1-Score values were calculated. Figures 2 and 3 show the Confusion Matrix for each machine learning model for the Bitcoin and Ethereum datasets respectively.

Due to the high imbalance of fraudulent and non-fraudulent transactions in both datasets, it was evident that the precision statistic was inaccurate because it was dependent on, and hence became vulnerable to, the number of fraudulent data points in the datasets. In order to measure the effective accuracy of our models, we instead focused on calculating the Confusion Matrix, Recall, Accuracy, and F1-Score values for the models since they do not depend on the number of fraudulent data points. Table III provides the validation results for each model.

For the Bitcoin dataset, the SVM machine learning algorithm was the best classifier with a Recall of 0.987, Accuracy of 0.987, and F1 Score of 0.994. Moreover, as seen from the Confusion Matrix in Figure 2(a), SVM correctly classified 902 fraudulent transactions from a total of 931 fraudulent transactions, producing an accuracy of 96.9%, while accurately classifying 98.7% of non-fraudulent transactions.

For the Ethereum dataset, the Random Forest machine



(a) Support Vector Machine.

(b) Random Forest.

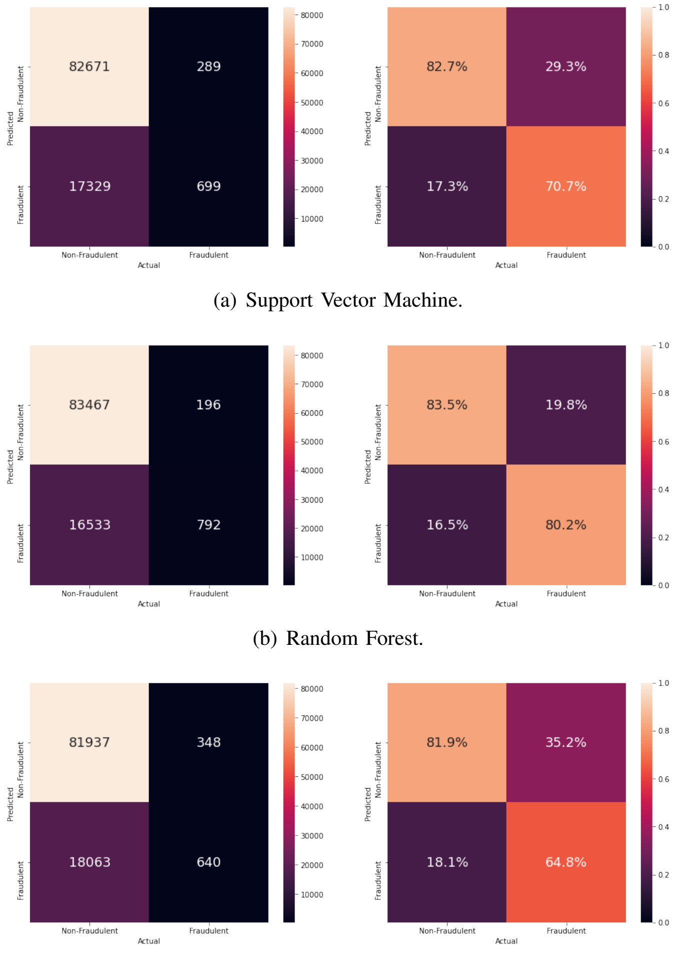
(c) Logistic Regression.

Fig. 2. Confusion Matrices for the ML models on the Bitcoin dataset.

TABLE III
VALIDATION RESULTS OF SVM, RANDOM FOREST, AND LOGISTIC REGRESSION CLASSIFIERS.

Algorithm		Recall	Accuracy	F1 Score
SVM	Bitcoin	0.987	0.987	0.994
	Ethereum	0.827	0.826	0.904
Random Forest	Bitcoin	0.930	0.930	0.964
	Ethereum	0.835	0.834	0.909
Logistic Regression	Bitcoin	0.898	0.897	0.946
	Ethereum	0.819	0.818	0.899

learning algorithm was the best classifier with a Recall of 0.835, Accuracy of 0.834, and F1 Score of 0.909. Similarly, as seen from the Confusion Matrix in Figure 3(b), the Random Forest model correctly classified 792 fraudulent transactions from a total of 988 fraudulent transactions, producing an accuracy of 80.2%, while accurately classifying 83.5% of non-fraudulent transactions. It is important to note that the Bitcoin dataset was comparably larger, thus, allowed for higher predictive accuracy from the machine learning models.



(a) Support Vector Machine.

(b) Random Forest.

(c) Logistic Regression.

Fig. 3. Confusion Matrices for the ML models on the Ethereum dataset.

VI. CONCLUSION

Cryptocurrencies are increasing in popularity as they attempt to expand in sectors beyond digital assets and financial services in the near future. Due to the anonymity and speed of transactions on the blockchain, it is naturally vulnerable to scammers attempting to fraud cryptocurrency holders. Therefore, it is important to conduct research studies in fraud detection on the blockchain.

In this paper, we focus on identifying fraudulent transactions and accounts by detecting anomalies in Bitcoin and Ethereum transaction networks, the two largest cryptocurrencies. Three GPU-accelerated machine learning classifiers were analyzed, Support Vector Machine (SVM), Random Forest, and Logistic Regression, on a dataset containing 30,294,698 transactions and transactions from 504,942 wallets for the Bitcoin and Ethereum networks respectively. The SVM algorithm obtained the best results on the Bitcoin dataset, with an accuracy of 96.9% and a Recall of 0.987. Similarly, the Random Forest algorithm obtained the best results on the Ethereum dataset, with an accuracy of 80.2% and a recall of 0.835.

The models trained in this study can be adopted to automated fraud detection systems. The study of anomaly detection in the cryptocurrency blockchain done in this paper can be generalized to other blockchain networks, such as the public sector, health service, and financial intelligence, which will prove beneficial to the continued reliability and expansion of cryptocurrencies and the blockchain.

The findings in this study create interesting opportunities for future studies. An Intrusion Detection System (IDS) is a cyber security software that detects malicious activities on systems and networks. It would be interesting to draw parallels regarding the design and implementation of IDS' with respect to the models used in this study.

References

[1] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system," *Decentralized Business Review*, p. 21260, 2008.

[2] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, no. 2014, pp. 1–32, 2014.

[3] M. S. Ahmad and S. M. Shah, "Moving beyond the crypto-currency success of blockchain: A systematic survey," *Scalable Computing: Practice and Experience*, vol. 22, no. 3, pp. 321–346, 2021.

[4] D. Wörner, T. Von Bomhard, Y.-P. Schreier, and D. Bilgeri, "The bitcoin ecosystem: Disruption beyond financial services?" 2016.

[5] M. Vasek and T. Moore, "There's no free lunch, even using bitcoin: Tracking the popularity and profits of virtual currency scams," in *International conference on financial cryptography and data security*. Springer, 2015, pp. 44–61.

[6] D. Farren, T. Pham, and M. Alban-Hidalgo, "Low latency anomaly detection and bayesian network prediction of anomaly likelihood," *arXiv preprint arXiv:1611.03898*, 2016.

[7] T. Pham and S. Lee, "Anomaly detection in bitcoin network using unsupervised learning methods," *arXiv preprint arXiv:1611.03941*, 2016.

[8] M. Ostapowicz and K. Żbikowski, "Detecting fraudulent accounts on blockchain: A supervised approach," in *International Conference on Web Information Systems Engineering*. Springer, 2020, pp. 18–31.

[9] H. Baek, J. Oh, C. Y. Kim, and K. Lee, "A model for detecting cryptocurrency transactions with discernible purpose," in *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, 2019, pp. 713–717.

[10] J. Kamps and B. Kleinberg, "To the moon: defining and detecting cryptocurrency pump-and-dumps," *Crime Science*, vol. 7, no. 1, pp. 1–18, 2018.

[11] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.