Machine Learning Approaches for Bitcoin Address Classification: A Comprehensive Review and Dataset

Abstract

The recent crises and bubble bursts in the crypto market, such as the Terra incident and the collapse of FTX, have raised concerns among government entities and investors. To increase transparency and monitor on-chain activities, reliable methods are needed to identify illegal transactions. Machine learning techniques can serve as valuable tools for classifying addresses and detecting illicit on-chain activities. This paper focuses on Bitcoin, the most popular cryptocurrency, and aims to provide a comprehensive review of the feasibility of using machine learning classification methods for labeling Bitcoin addresses. Additionally, we present a publicly available dataset for address labeling by scraping data from Blockchain.com. The dataset includes 15,355 labeled Bitcoin addresses, as well as labels for Tokens, ETH, and BCH addresses. These resources are made available on GitHub (https://github.com/Jasontth/crypto_label) to facilitate research and enable the training of classification models in this domain.

Introduction

In May 2022, TerraClassicUSD (USTC), one of the largest stablecoins, lost its peg to the US dollar, resulting in a loss of \$500 billion, which accounted for approximately a 20% devaluation in the DeFi ecosystem (Jalan & Matkovskyy, 2023). The collapse of Terra also triggered a 'crypto winter', causing a sharp decline in cryptocurrency prices and an extended period of low prices (Chohan, 2022). Bitcoin, in particular, experienced a significant drop from around \$68,000 to below \$20,000, representing a 70% decline (Jalan & Matkovskyy, 2023).

Worse still, just a few months after the Terra incident, FTX, the third-largest centralized exchange (CEX), collapsed in November (Fu et al., 2023). Scholars have even described the FTX collapse as the "Pozi Game," as it was unable to sustain its governance and resulted in the disappearance of \$32 billion overnight. On the same day, the prices of major cryptocurrencies declined by more than 30% (Fu et al., 2023). This incident caused significant financial losses for investors and hedge funds. Additionally, the US Department of Justice and other federal agencies launched investigations into this event (Shumba, 2022).

Although cryptocurrencies like Bitcoin are designed to be decentralized and address issues in traditional finance, such as systematic risks and third-party intervention (Yuneline, 2019), the evolving crypto market has witnessed a series of market failures and crises, giving rise to new financial problems (Arner et al., 2023). In order to maintain the functionality of cryptocurrencies and protect investors, regulations are needed. The "Crypto Winter" of 2022-2023 has sounded the alarm

for government and financial regulators to find ways to address the decentralized nature of cryptocurrencies (Arner et al., 2023).

In this paper, we will conduct a comprehensive review to explore various models for classifying Bitcoin addresses based on their on-chain transaction activities using machine learning techniques. These machine learning classification models can serve as powerful tools to detect and deter illegal transactions by identifying addresses that are labeled and involved in illicit activities (Lee et al., 2020).

Research-background

Scholars have employed supervised machine learning methods and conducted experiments using various features extracted from Bitcoin on-chain activities (Lee et al., 2020). However, to fully comprehend the classification model, it is crucial to first grasp the underlying data structure of Bitcoin and the process of extracting these features.

The structure of a Bitcoin transaction is based on the concept of unspent transaction outputs (UTXOs) (Huang et al., 2021). Each transaction in Bitcoin's blockchain is a public entry, and transactions form a chain where one transaction spends the outputs of the previous transaction. In Figure 1, the inputs of the processed Bitcoin transaction are mapped from the previous UTXOs transaction.

Figure 1. Bitcoin Transaction Data Structure

A Bitcoin transaction typically includes various components, such as version, inputs, outputs, and lock time (Huang et al., 2021). However, for the purpose of understanding transactions and their

application in classification models, we can simplify the data structure as depicted in Figure 1. Some important features in a transaction include the transaction time (in UNIX Timestamp), the block height and position in the block, the gas fee, the transaction ID (txid), and the lists of inputs and outputs. It is worth noting that Bitcoin transactions are many-to-many, where multiple inputs can meet multiple outputs (Eck et al., 2021). This is quite different from traditional transactions, as it is not a simple one-way transfer. A suitable methodology for understanding Bitcoin transactions is to consider that coins are inputted into a melting pot and then redistributed to the output addresses and gas fees.

From this data structure, we can further process the transactions and extract features for model training (Lee et al., 2020). Here are few of the most important features that can be extracted:

Total Bitcoin amount (transmit/receive): The Bitcoin amount feature involves filtering transactions related to the selected addresses and calculating the net amount of Bitcoin transmitted or received by each address. It is computed by subtracting the total spent UTXOs from the total received UTXOs. The mathematical equation for the Bitcoin amount feature is: Σ (output value) - Σ (input value). This can also be understood as the balance of the bitcoin address.

Total transmit amount: This feature is computed by summing up the spent UTXOs for each address. This calculation is performed by grouping the transactions based on the inputs list, and summing the input values for each occurrence of the address. Mathematically, it can be expressed as Σ (input value).

Total receive amount: This feature is computed by summing up the received UTXOs for each address. This calculation is performed by grouping the transactions based on the outputs list, and summing the output values for each occurrence of the address. Mathematically, it can be expressed as Σ (output value).

Total transmit/received count: This transmit count feature is computed by counting the occurrences of each address appearing in the input list of transactions. Similarly the received count feature is computed by counting the occurrences of each address appearing in the output list of transactions.

Average Transaction Interval: The average transaction interval is computed by taking the sum of the Unix time differences between each transaction and dividing it by the total number of transactions which can be expressed by $\Sigma(T_{i+1} - T_i) / (n - 1)$.

Transaction Lifetime: The Transaction Lifetime of an address is defined as the total length of time from when it was first recorded as a receiving address to the last time it transmitted or received a transaction. It can be calculated by subtracting the Unix time of the first recorded instance as a

receiving address from the Unix time of the last transmission or receipt. This can be expressed by T_{Last} - T_{Initial} .

Total Lifetime: The total lifetime of an address is defined as the total length of time between its first recorded instance as a receiving address and the current time. It can be computed by subtracting the Unix time of the first recorded instance as a receiving address from the current Unix time. This can be expressed by T_{Current} - T_{Initial} .

Label: This is the target label for the dataset, which involves classifying the addresses. The labeling approach used in papers like Febrero–Bande et al. (2023) and Lee et al. (2020) involves categorizing the data using labels obtained from the website https://www.walletexplorer.com/. These categories include Exchanges, Mining Pools, Services, Gambling, and Old/Historic.

In reality, the features of Bitcoin can encompass a wider range. The previously mentioned features highlight some key aspects of Bitcoin, as its transaction network is publicly accessible and recorded on a ledger. Transaction data can be extracted and analyzed in various ways. Additionally, there are features that are closely associated with illegal transactions, such as high transaction gas fees to speed up the inclusion in blocks, the presence of multiple identical outputs within a single transaction which might be indicative of money laundering or the utilization of coin mixing services (Lee et al., 2020). Another paper, such as the one by Lin et al. (2019), incorporates numerous higher-order moments to provide a comprehensive summary of the transaction history.

Model Analysis

Kanemura et al. (2019) proposed a new classification model that utilized a voting-based method to classify the label of a user. The approach involved classifying all the addresses associated with a the same owner and using majority voting to determine if the owner could be classified as being related to darknet market activities. It is important to note that the paper's findings were based on computer simulations rather than real-life on-chain data. The researchers extracted 73 features and used them to train the classifiers. The voting-based method achieved an F1 score of 0.8. Furthermore, the model successfully identified that addresses involved in darknet activities were more inclined to pay higher gas fees compared to normal bitcoin participants.

Lin et al. (2019) introduced new features for Bitcoin address classification. They applied basic statistics, extra statistics, and transaction moments. The training features captured the frequency of transactions, total amount, distribution of transactions, and interval of transactions, combined with 64 features. Furthermore, the research applied Light Gradient Boosting Machine (LightGBM), which is a decision tree-based model, to identify the most important features in the model (Figure 2). Moreover, the authors proved that the proposed new features, such as the frequency of transactions, the mean value of the number of outputs, the mean value of the number of inputs, the number of received transactions, and the moment of transaction interval distribution, dominated as the most important features. The results also demonstrated that both LightGBM and Neural Network performed well,

with LightGBM achieving the highest accuracy as the best address-based classification model, yielding an F1 score of 0.87.

Figure 2. Top 20 Important Features

(a) The top 10 features.		(b) The 11^{th} to 20^{th} features.	
Feature Name	Feature Type	Feature Name	Feature Type
f_{TX}	Basic Stats	$f_{received}$	Basic Stat
$\bar{N}_{ m outputs}$	Basic Stats	$n_{ m spent}$	Extra Stat
$\bar{N}_{ ext{inputs}}$	Basic Stats	n_{TX}	Extra Stat
$n_{ m received}$	Extra Stats	$m_{2, { m received}}$	Moment
$m_{1, \text{interval}}$	Moments	BTC_{spent}	Extra Stat
$\sigma_{ m balance_btc}$	Extra Stats	$\mu_{ m balance_usd}$	Extra Stat
lifetime	Extra Stats	$m_{1,\mathrm{total}}$	Moment
$r_{\rm payback}$	Basic Stats	$m_{2,\mathrm{total}}$	Moment
$m_{1, m received}$	Moments	$BTC_{ m received}$	Extra Stat
$\mu_{\mathrm{balance_btc}}$	Extra Stats	$f_{\text{received}}(10^2)$	Basic Stat

Lee et al. (2020) conducted a similar study employing a similar methodology. They designed a workflow that involved historic transaction retrieval, bitcoin address feature extraction, model training, and testing (Figure 3). The transaction records were retrieved and categorized based on exchange, mining pool, mixer, gambling, and darknet labels provided by WalletExplore. The number of features used for model training was increased to 80. The research demonstrated that Random Forest, another decision tree-based classifier that utilizes the bagging method, achieved an accuracy of 84%. The model also exhibited high accuracy in classifying dark market activities, including the Silk Road, which was one of the largest dark markets.

Mining Silk Road Transaction Collection Address & Feature Extraction Transmission Address vs Recipient Address 80 Features Labeling Exchange (0) Mining Pool (1) Mixer(2) Gambling (3)Silk Road (4) ML model training Artificial Random Neural Network Decision / Testing

Figure 3. Classification methodology

Recommendation

This paper identified several limitations, namely the difficulty in retrieving historical transactions and limited labeled addresses for training.

The public and researchers lack an open-source public NoSQL or SQL database for bitcoin transaction data. Extract, Transform, and Load (ETL) processes are needed to process and further analyze the blockchain (Galici et al., 2020). One of the popular Blockchain ETL libraries is https://github.com/blockchain-etl; however, it lacks maintenance as there are version changes in bitcoin. Researchers and blockchain experts in this area should have standard tools to review and adapt to data structural changes when there are version updates inside bitcoin. Currently, researchers can only rely on third-party tools such as Blockchain.com, BTC.com, and Blockchair services. However, their databases are not publicly available, and the API prices can be costly, which is a burden for researchers and data scientists. Therefore, a public database that stores the extracted data in a conventional database is needed to make data analysis and querying more efficient.

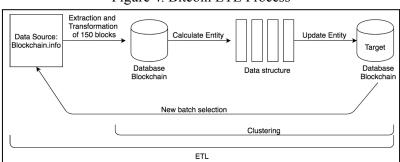


Figure 4. Bitcoin ETL Process

This paper proposes a more concise and detailed ETL (Extract, Transform, Load) process, as shown in Figure 4. Instead of extracting resources from third-party sources, the proposal suggests extracting data directly from the public ledger. The collection process should follow these steps: first, set up a full Bitcoin node to synchronize with the Bitcoin blockchain. Extraction can then be performed and transformed in batches. Tools will be required to map the input of the transaction and ensure complete transaction data.

Furthermore, this paper introduces a method to achieve a real-time database. We continuously extract historic transactions in batches until the sum of processed blocks and the number of batches is greater than the current block height. At that point, we reduce the batch size to one. Since the interval between new mined blocks is currently around 10 minutes, we can set the batch retrieval interval to somewhere between 1 minute and 5 minutes to save computational cost if needed or implement constant recalling. The pseudo code is as follow:

Figure 4. Real-Time Database Pseudo Code

```
# Initialize count of processed blocks

processed_blocks = 0

# Set up database connection

database.connect()

# Loop until the desired condition is met

while True:

# Check if the condition for decreasing batch size is met

if (processed_blocks + batch_size) > bitcoin_node.current_block_height:

# Decrease the batch size to 1

batch_size = 1

# Set the starting point for block retrieval to the last block height

bitcoin_node.set_starting_point(processed_blocks)

# Extract batch of transactions from blockchain

data = bitcoin_node.extract_data(batch_size)

# Transform extracted data

transformed_data = transform_data(data)

# Store transformed data in database

database.load(transformed_data)

# Increment the count of processed blocks by batch size

processed_blocks += batch_size

# Set batch retrieval interval to a value between 1 and 5 minutes

wait(batch_interval)
```

Apart from the Bitcoin ETL and database recommendation, this paper realizes the limitations and problems of the paper by Lee et al. (2020). The labeled data was extracted and provided by WalletExplorer, which is no longer maintained and outdated since 2016. The labeled data is also of poor quality as most addresses are no longer in use and have empty balances. Therefore, to address the need for newly labeled and constantly maintained data, we extracted and scraped the updated data from Blockchain.com. The appendix provides further details on this process. The scraped data is publicly uploaded to https://github.com/Jasontth/crypto_label. The dataset includes 15,355 BTC, 196 ETH, 1,205 ETH Contracts, 39 BCH, and 987 labeled tokens. We believe that these data will be useful and reliable sources for researchers and data scientists in training their classification models.

Furthermore, labeling addresses requires effort from government and state agencies. Methods such as dusting attacks can be used to deanonymize abnormal addresses (Loporchio et al., 2023). Government or law enforcement entities can send small amounts of BTC (dust) to suspicious accounts to identify and label the addresses. Governments should also collaborate with exchanges to label and halt transactions involving wallets associated with money laundering or coin mixing services. Moreover, the labeled addresses and computed features dataset can be made open source for data scientists to explore different supervised machine learning techniques and propose new features to improve address classification techniques.

Currently, the government relies on public-private partnerships and consulting companies like Chainalysis for monitoring illegal activities on blockchain (CCN, 2023). However, this approach raises concerns about privacy intrusion and disruption to the Bitcoin ecosystem. The classification model can help mitigate these risks by enabling surveillance and investigation only on selected or labeled addresses, without requiring the disclosure of the true identities of all Bitcoin users. We hope that governments and policymakers will consider adopting such technology to ensure the safety of blockchain users with minimal intervention.

Conclusion

This paper presents a comprehensive review of various research studies on bitcoin address classification and explores different machine learning techniques for this purpose. The research demonstrates that address classification models are feasible and can achieve high accuracy. The paper also discusses the data structure of Bitcoin transactions and the extraction of features in the machine learning process. Additionally, the paper proposes a method for maintaining an Extract, Transform, and Load (ETL) structured database for bitcoin transaction analysis. Furthermore, a public dataset of labeled addresses is provided for machine learning purposes.

The main objective of this paper is to contribute to the ongoing efforts aimed at increasing transparency and monitoring on-chain activities in the cryptocurrency domain. By gathering research and community efforts in creating open-source databases and machine learning datasets for training, the paper aims to facilitate the development of valuable tools for classifying addresses and detecting illicit on-chain activities. Ultimately, these efforts help ensure the functionality of cryptocurrencies and protect investors from financial losses.

References

- Arner, D. W., Zetzsche, D. A., Buckley, R. P., & Kirkwood, J. (2023). The Financialization of Crypto: Lessons from FTX and the Crypto Winter of 2022-2023. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4372516
- CCN. (2023, November 1). Government surveillance of your crypto has never been so high | *CCN.com.* CCN.com.
 - https://www.ccn.com/news/government-surveillance-of-crypto-never-been-higher/
- Chohan, U. W. (2022). Crypto Winters. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4142885
- Eck, D., Torek, A., Cutchin, S., & Dagher, G. G. (2021). Diffusion: Analysis of Many-to-Many Transactions in Bitcoin. *2021 IEEE International Conference on Blockchain (Blockchain)*. https://doi.org/10.1109/blockchain53845.2021.00061

- Febrero–Bande, M., González-Manteiga, W., Prallon, B., & Saporito, Y. F. (2023). Functional classification of bitcoin addresses. *Computational Statistics & Data Analysis*, *181*, 107687. https://doi.org/10.1016/j.csda.2022.107687
- Fu, S., Qin, W., Yu, J., & Chen, S. (2023). FTX collapse: a Ponzi story. In *Lecture Notes in Computer Science* (pp. 208–215). https://doi.org/10.1007/978-3-031-48806-1 14
- Galici, R., Ordile, L., Marchesi, M., Pinna, A., & Tonelli, R. (2020). Applying the ETL Process to blockchain Data. Prospect and findings. *Information*, 11(4), 204. https://doi.org/10.3390/info11040204
- Huangt, Y., Qiu, W., & Wut, C. (2021). Comparison and Analysis of Data and Transaction Structure of Bitcoin and Ethereum. 2021 2nd International Conference on Computer Communication and Network Security (CCNS). https://doi.org/10.1109/ccns53852.2021.00029
- Jalan, A., & Matkovskyy, R. (2023). Systemic risks in the cryptocurrency market: Evidence from the FTX collapse. *Finance Research Letters*, *53*, 103670. https://doi.org/10.1016/j.frl.2023.103670
- Kanemura, K., Toyoda, K., & Ohtsuki, T. (2019). Identification of Darknet Markets' Bitcoin Addresses by Voting Per-address Classification Results. 2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). https://doi.org/10.1109/bloc.2019.8751391
- Lee, C., Maharjan, S., Ko, K., Woo, J., & Hong, J. W. (2020). Machine Learning based Bitcoin address Classification. In *Communications in computer and information science* (pp. 517–531). https://doi.org/10.1007/978-981-15-9213-3 40
- Lin, Y., Wu, P., Hsu, C., Tu, I., & Liao, S. (2019). An Evaluation of Bitcoin Address Classification based on Transaction History Summarization. *2019 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*. https://doi.org/10.1109/bloc.2019.8751410
- Loporchio, M., Bernasconi, A., Di Francesco Maesa, D., & Ricci, L. (2023). Is Bitcoin gathering dust? An analysis of low-amount Bitcoin transactions. *Applied Network Science*, 8(1). https://doi.org/10.1007/s41109-023-00557-4
- Shumba, C. (2022, December 5). US Justice Department Wants FTX Fraud Allegations to Be Investigated. *CoinDesk*. Retrieved December 8, 2023, from https://www.coindesk.com/policy/2022/12/02/us-justice-department-wants-ftx-fraud-allegations-to-be-investigated/
- Yuneline, M. H. (2019). Analysis of cryptocurrency's characteristics in four perspectives. *Journal of Asian Business and Economic Studies*, *26*(2), 206–219. https://doi.org/10.1108/jabes-12-2018-0107

■ blockchain.com/explorer/_next/static/chunks/7674-56eed0017e032434.js

1. Open a web browser and navigate to the URL:

https://www.blockchain.com/explorer/_next/static/chunks/7674-56eed0017e032434.js

2. Once the page loads, you will see the JavaScript file content

"use strict";(self.webpackChunk_N_E=self.webpackChunk_N_E||[]).push([[7674],{3772}], n=i(85893),t=i(53918),c=i(67294),s=i(27134),b=i(11163),d=i(36297),f=i(14900); func {value:0bject.freeze(a)}}))}function h(){var e=r(["\n display: flex;\n align-it 7px;\n width: 100%;\n\n ",";\n"]); return h=function(){return e},e}function o(){ auto;\n height: max-content;\n\n @media only screen and (min-width: 0px) and (min-width: 640px) and (min-height: 0px) {\n min-width: 728px;\n min-height: 46845f4-0"})(h(),(function(e){var a=e.hiddenBelow;t=(unction a&&"@media (max-width: (unction w(e){var a=e.zone,i=e.hiddenBelow;t=(unction w(e){unction w(e){unction w(e)},r=t[0],h=t[1]} (a),M=(0,c.useRef)(null),U=(unction w(e),v=ret{unction w(e),r=t[0],h=t[1]} (a),M=(unction w(e),t=t[0],h=t[1]} (a),M=(unction w(e),t=t[0],h=t[1]}

Search for the function with the identifier "99486: function(e)" which is the dictionary of token address

```
twelvefold minter":return Me;default:return!a&&J}}, 99486:function(e)
{e.exports=JSON.parse('{"0x006BeA43Baa3f7A6f765F14f10A1a1b08334EF45":
{"hash":"0x006BeA43Baa3f7A6f765F14f10A1a1b08334EF45", "name":"Stox", "token":"STX"}, "0x03271124;
c5CB9f7C33fd":
{"hash":"0x0377112423F3A68efdF1fcF402F6c5CB9f7C33fd", "name":"PieDA0BTC++", "token":"BTC"}, "0x047aF4CF447DE572eF828";
{"hash":"0x04Fa0d235C4abf4BcF4787aF4CF447DE572eF828", "name":"UMAVotingTokenv1", "token":"UMA"},
7f7cfb6E246680c53927DD30";
{"hash":"0x08d967bb0134F2d07f7cfb6E246680c53927DD30", "name":"MATHToken", "token":"MATH"}, "0x0488Ee10d9d922659cB7";
{"hash":"0x0A913beaD80F321E7Ac35285Ee10d9d922659cB7", "name":"DOSNetworkToken", "token":"DOS"},'
9521C384F1D2123D1f195e6":
{"hash":"0x0Ae055097C6d159879521C384F1D2123D1f195e6", "name":"STAKE", "token":"STAKE"}, "0x88dF58b0fBc02cf3778a0":
```

4. Search for the function with the identifier " 28104: function(e)" which is the dictionary of BTC/ETH/BCH address

5. Search for the function with the identifier "99486: function(e)" which is the dictionary of token address

```
{"hash":"qzrpa9hut3jaydnsxelzff3dws88nrvghcm8zsu4xe","miner":"Miner","name":"BCH M1"}}}')}, 36492:function(e)
{e.exports=JSON.parse('{"0xb4e16d0168e52d35cacd2c6185b44281ec28c9dc":{"token":"uni","name":"Uniswap V2 -
USDC/WETH","hash":"0xb4e16d0168e52d35cacd2c6185b44281ec28c9dc"},"0x3139ffc91b99aa94da8a2dc13f1fc36f9bdc98ee"
{"token":"uni","name":"Uniswap V2 -
PAX/USDC","hash":"0x3139ffc91b99aa94da8a2dc13f1fc36f9bdc98ee"},"0x12ede161c702d1494612d19f05992f43aa6a26fb":
{"token":"uni","name":"Uniswap V2 -
CHAI/WETH","hash":"0x12ede161c702d1494612d19f05992f43aa6a26fb"},"0xa478c2975ab1ea89e8196811f51a7b7ade33eb11"
{"token":"uni","name":"Uniswap V2 -
DAI/WETH","hash":"0xa478c2975ab1ea89e8196811f51a7b7ade33eb11"},"0x07f068ca326a469fc1d87d85d448990c8cba7df9":
{"token":"uni","name":"Uniswap V2 -
REN/USDC","hash":"0x07f068ca326a469fc1d87d85d448990c8cba7df9"},"0xae461ca67b15dc8dc81ce7615e0320da1a9ab8d5":
{"token":"uni","name":"Uniswap V2 -
DAI/USDC","hash":"0xae461ca67b15dc8dc81ce7615e0320da1a9ab8d5"},"0xce407cd7b95b39d3b4d53065e711e713dd5c5999":
{"token":"uni","name":"Uniswap V2 -
```