IEEE ICBC'24 1570978737

# Leveraging Machine Learning For Multichain DeFi Fraud Detection

***Abstract*—Smart contracts across Blockchains provide an ecosystem of decentralized finance (DeFi), with a total locked value which had exceeded 160B USD. While DeFi comes with high rewards, it also carries plenty of risks. Many financial crimes have occurred over the years making the early detection of malicious activity an issue of high priority. The proposed framework introduces an effective method for extracting a set of features from different chains, and it is evaluated over an extensive dataset with the transactions of the 23 most widely used DeFi protocols based on a novel dataset in collaboration with Covalent. Different Machine Learning methods were employed, such as a Deep Neural Network, XGBoost, and a fine-tuned Large Language Model for identifying fraud accounts interacting with DeFi and we demonstrate that the introduction of novel DeFi-related features, significantly improves the evaluation results.**

***Index Terms*—Machine Learning, Fraud Detection, Multichain, Ethereum, Decentralized Finance, Blockchain.**

## I. Introduction

The emergence of Blockchain as the accounting method for Bitcoin [1] was followed by the introduction of more Blockchains [2] with potential for different applications beyond cryptocurrency such as IoT [3], identity management [4], copyrights management [5], and many more. Smart contracts' implementation and execution allowed further the construction of sophisticated on-chain financial systems, namely DeFi [6], where users can interact with lending pools [7], [8], Automated Market Maker exchanges (AMMs), stablecoins, derivatives, and asset management platforms [9].

Several works have combined ML and Blockchain [10] focusing on both Bitcoin and Ethereum platforms, addressing problems such as Ponzi Scheme detection [11], address clustering [12], [13], [14], cryptocurrency price prediction [15], phishing detection [16], illicit accounts detection [17] and many more. Existing works on Fraud Detection in Bitcoin and Ethereum mainly focus on proposing Machine Learning methods, utilizing labeled datasets for evaluation, and extracting features from the transaction history. In this direction, we extend existing approaches by taking into consideration the interactions of entities with DeFi protocols and proposing novel features that improve the classification results.

**Overall paper contributions:**

1) **Datasets:** we collected DeFi transactions across 23 protocols and 12 chains and created three datasets: i) a very comprehensive dataset thanks to a collaboration with Covalent [18] of all the DeFi-related transactions; ii) compiled all the publicly available datasets associated with illegitimate activities such as phish, hack, heist, and scam from the literature and through online resources; iii) identified all the accounts who participated in illegitimate activities and interacted with the major DeFi protocols. As a result, we are the first, to the best of our knowledge, to utilize a labeled dataset for DeFi malicious accounts detection.
2) **Features Extraction:** a set of features was extracted based on the Ethereum transactions and behavior of the entities. Additionally, we introduce a set of novel DeFi-related features, the extraction and exploitation of which improved the evaluation results.
3) **Machine Learning Framework:** we train different ML methods to detect the malicious entities and evaluate their performance using widely used metrics, while we tackle the data imbalance issue, which is quite common in this field, by employing oversampling techniques.

The rest of this paper is structured as follows: In Section 2, we present the related work that has been done in the field, which formed the basis of our proposed solution. Section 3 contains the methodology followed, including the data collection process and the feature extraction, while in Section 4 we present the experimental results of the machine learning algorithms. Finally, Section 5 concludes with brief remarks on potential avenues for future research.

## II. Related Work

In the recent years, several research works leveraged machine learning methods to address the problem of Fraud Detection in the Bitcoin and Ethereum Blockchain platforms [17]. Poursafaei et al. [19] extracted a set of features from the Ethereum blockchain data to represent the transactional behavior of entities. In their solution, they applied different ML classification algorithms (such as LR, SVM, RF) to identify malicious entities in the Ethereum blockchain network. Farrugia et al. [20] proposed a ML classification method based on XGBoost to detect illicit accounts and published their respective dataset including accounts that were flagged by the Ethereum community for illicit behavior. In a similar direction, Weber et al. [21] utilized GCNs over a labelled dataset of ilicit Bitcoin transactions. Other research works focused on Decentralized applications (DApps) and DeFi. An analysis of the behavioral characteristics of the users of DApps was performed by [22] applying unsupervised Self-Organizing Map based classification of addresses to distinguish blockchain investors and players. Darlin et al. [23] proposed Ethereum address grouping and classification algorithms to calculate the percentage of fund flows into DeFi lending platforms that can be attributed to debt created elsewhere in the system,

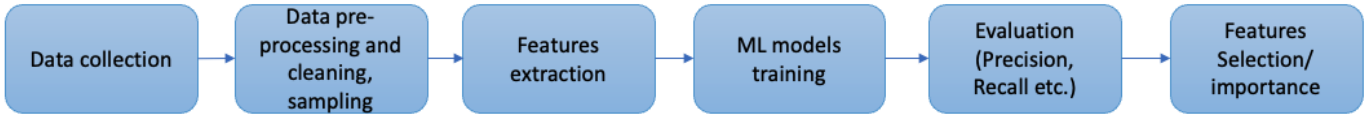


Fig. 1: Overview of the proposed solution - Machine Learning Framework pipeline

analyzing data from five major DeFi protocols. Recent studies provided useful insights after analyzing Bitcoin/Ethereum and DeFi functions and data combined with external sources and datasets, such as the public sentiment from Twitter [24], or high-frequency volatility and price data [25].

Li et al. [26] used a set of 3559 malicious labeled Ethereum addresses, analyzed transactions and applied a network clustering analysis to identify whether additional Ethereum addresses should be marked as malicious. Trozze et al. [27] used open-source investigative tools to study the DeFi frauds and money laundering, focusing on tokens, smart contracts, rug pulls and how tools can be used to extract evidence of scams on Ethereum. Wang et al. [28] proposed a system called DeFiScanner, trained on a set of 50910 DeFi transactions, extracting 37 features about functions and transactions. They adapted word2vec and then used seven models to detect attacks. DeFiScanner only deals with DeFi attacks on the transaction layer and focuses on multilayer attacks on DeFi protocols. Their proposed extracted features are for each transaction such as the number of transfer functions and withdraw functions.

In our study, we differentiate from the previous approaches since: i) we gathered a large dataset with more than 54,000,000 DeFi transactions of unique addresses interacting with major DeFi protocols from May 2019 until March 2023; ii) use a large labeled set of addresses compiling all available public sources; iii) we extract features about addresses and more specifically 414 DeFi-related features and 9 transactional features in the classification process; iv) we focus in identifying a target set of 81 addresses labeled as DeFi malicious; v) we propose a Machine Learning Framework using SVM, Random Forest, Logistic Regression, XGBoost, a Deep Neural Network and LLMs as shown in Figure 1. Each of the components of this framework, is presented in detail in the following section.

## III. Methodology

### A. Data Collection

Data used for this study originate from Ethereum as well as publicly available datasets published by existing research studies and complementary publicly available web resources. Python and NodeJS, scripts were used for data retrieval and analysis. For initial experiments, we filtered for the relevant data from the Google BigQuery's public Ethereum dataset [36]. As described, the main data on the DeFi events were provided by Covalent [18]. We established an Ethereum full archive node, featuring an 8-core Intel i7-11700 CPU and 10TB SSD, to facilitate the research and data analysis processes. By utilizing the node, we have unlimited access to all the transactions that took place since the Genesis block and we were able to search for transactions including specific actions that characterize the applications and protocols of DeFi.

TABLE I: Summary of datasets and sources available with annotated data of malicious addresses

| Author/Source | Description and Reference |
|---|---|
| CryptoScamDB | Open-source dataset tracks malicious URLs and their associated addresses [29] |
| Etherscan labels | A list of tagging, and categorizing of addresses and tokens listed on Etherscan [30] |
| Hall et al. | A labeled dataset with 5212 addresses |
| Tether Blacklisted Addresses | A list of Ethereum addresses [31] |
| MyEtherWallet blacklist | A list of Ethereum addresses [32] |
| Ferrugia et al. | A dataset of 2179 Ethereum accounts [20] |
| Xblockpro Ethereum Datasets | Various labeled datasets [33] |
| Al-E'mari et al | Dataset available at [34] [35] |

We collected transactions occurring between blocks 5771740 (May 5, 2019 1:10:38 PM UTC) and 36843000 (March 12, 2023 4:46:11 PM UTC), inclusive, for 23 protocols including Aave, Compound, Curve, Lido, and Yearn. Each protocol has upgraded the initial versions of their smart contracts, and in some cases multiple versions of the protocol may exist at one time. Therefore, we collected data from the set of smart contracts associated to the different protocol versions.

*1) DeFi transactions dataset:* We identified the smart contracts belonging to DeFi protocols. Following the data collection process described above, in this querying process for DeFi activities, we focused on addresses interacting with these DeFi protocols and the respective transactions. To this direction, we gathered all the transaction records through these contracts assembling them into the "DeFi transactions dataset", gathering more than 54,000,000 transactions for 550,000 unique entities (addresses).

*2) Covalent dataset on decoded DeFi events:* Covalent [18] is a new blockchain company that offers one of the most complete dataset on DeFi protocols through APIs to their clients. They have indexed more than 100B transactions across 60 blockchains and 200k smart contracts. On top, they clean and normalize the data, to make it more accessible for users. Covalent has graciously offered us to use their dataset for the research presented in this paper. The underlying dataset presented is hence one of the most advanced in the industry.

*3) Labeled Dataset:* We compiled all the publicly available labeled datasets associated with illegitimate activities trough publicly accessed online resources and the litterature such as (i) trying to imitate other contract addresses providing tokens,

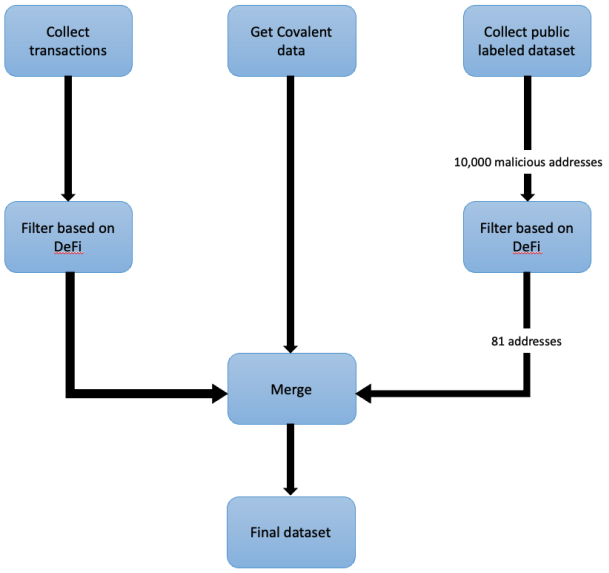


Fig. 2: Process of creating the labeled dataset for Ethereum DeFi fraud detection.

TABLE II: Extracted features from the transactional behavior of accounts

| Feature Count | Details |
|---|---|
| 1 | Total number of transactions |
| 1 | Share of total submitted transactions to mempool |
| 2 | std, and Max - Min (i.e., age) on block height (i.e., age) |
| 1 | Number of transactions divided by age of wallet |
| 3 | Mean, max, and std on gas costs per transaction |
| Total number of transactional features: 8 | |

(ii) scam lotteries, (iii) fake initial coin offerings (ICO), (iv) imitating other users, (v) Ponzi schemes, (vi) phishing, (vii) mirroring websites, viii) hack and ix) heist.

In the Table I, the online sources and available datasets are summarized and it consists of more than 10,000 addresses. Followingly, we identified all the accounts from the previous labeled dataset, which interacted with the major DeFi protocols from the "DeFi Transactions Dataset", creating a labeled dataset with addresses. While we retrieved more than 10,000 malicious addresses from the literature and previous works, only 81 addresses of them have also interacted with DeFi protocols. Figure 2 describes part of this process.

### B. Data Pre-processing

We are tackling a classification problem with a dataset, which is imbalanced. From the set of good addresses, we select 10K addresses, while the malicious class consists of 81 addresses. Several research works have processed imbalanced datasets with blockchain data [37], applying techniques such as SMOTE [19]. To this direction, we apply SMOTE, increasing the size of the smaller class in the training part of the pipeline. We also applied Data Cleaning in certain cases, where we handled missing numerical values with the respective median.

TABLE III: Introducing DeFi-related features

| Feature Count | Details |
|---|---|
| 24 | Sum, mean, and std on 8 events: Add liquidity, remove liquidity, borrow, deposit, liquidation, repay, swap, withdraw |
| 10 | Min, max, std, mean, and median on 2 types of fees: Protocol fees, and gas fees related to the DeFi protocol interaction |
| 92 | Number of transactions, sum, mean, and standard deviation of outgoing token value in USD on 23 protocols including Aave, Compound, Yearn, Curve, Lido, Balancer, etc. |
| 44 | Number of transactions, sum, mean, and standard deviation of outgoing token value in USD on 11 chains including Arbitrum, Aurora, Avalanche, etc. |
| 144 | Aggregate all transactions per every 1000 blocks to see activity in a timeframe and take min, max, and std on 8 events |
| 100 | Number of transactions involving one of the top 99 most traded tokens, plus an aggregation of the long tail |
| Total number of DeFi features: 414 | |

### C. Feature Extraction

In Table II, the list of extracted features based on the transactional behavior of entities is presented. Table III presents the extracted features based on the interactions of entities with DeFi protocols:

- **Add/Remove Liquidity:** Central to AMMs, these actions are crucial in understanding liquidity patterns, which can be exploited in schemes such as rug pulls.
- **Borrow/Repay:** Important in lending platform, where unusual borrowing or repayment patterns might indicate exploitative practices, including flash loan attacks.
- **Liquidation Events:** Anomalies in liquidation could indicate market manipulation or predatory practices, seen commonly in DeFi markets.
- **Swaps:** Irregular swap transactions could suggest wash trading or price manipulation attempts within DeFi exchanges.

We also made sure that the features are not too correlated [38] [39], using their correlation matrix Additionally, we experimented with the normalization of features and filtering out features with zero variance, investigating the distribution of the features using boxplots.

### D. Machine Learning Algorithms

We used six different machine learning algorithms: Logistic Regression, Random Forest, Support Vector Machine (SVM), XGBoost, a Deep Neural Network (DNN), and a fine-tuned large language model (LLM). For each of the models, we apply a 5-fold cross-validation (CV), hence splitting the dataset in 80% training and 20% testing data five times, and training the model on each batch independently.

- **Logistic Regression, Random Forest and XGBoost:** Standard implementation with sklearn [40].
- **SVM:** Standard implementation with sklearn and rbf kernel.
- **DNN:** Implementation with TensorFlow [41] with mean squared error (MSE) loss function and Adam optimizer. The model consists of five layers: a dense layer with 40 nodes and ReLU activation, a 0.3 dropout layer, a dense

layer with 10 nodes and ReLU activation, a dense layer with 5 nodes and ReLU activation, and a dense output layer with 1 node and sigmoid activation. The model is then trained through 30 epochs with 128 batch size.

- **LLM:** The Curie GPT-3 model from OpenAI [42] was fine-tuned using the previously described dataset, converted to strings, and used to one-shot classify a test set. The process is described below.

*1) Detailed description of Large Language Model deployment:* First, the data was preprocessed as described above, with the addition of a further downsampling of the "Good" class to 1000 examples. The features were then normalized, and put in the same string format of {decimal}{value with 3 significant digits}, such that all feature values have the same length as string. All values were then comma separated and added to a string, and the label attached in a dictionary. Using the "openai" python package, the data is then preprocessed and split into a training and validation set, and then submitted through the OpenAI API to fine-tuning. This uses the "curie" model of the GPT-3 pre-trained LLM from OpenAI [43] over 2129 steps. Finally, the tuned model can be used on the test set to evaluate the performance, and the labels are applied based on the given probabilities by the model for the classes.

### *E. Model Verification Techniques*

To evaluate the performance of our models, we use the following widely used metrics: Precision, Recall, Accuracy, F1-score. They are defined through combinations of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). Moreover, we decided to utilize the F2-score, in order to put more attention to the minority class, by minimizing False Negatives (erroneously classifying "Malicious" samples as "Good") over minimizing False Positives.

### *F. Evaluating the performance of the classification algorithms*

In this section, the effectiveness of the preprocessing and feature extraction is evaluated, while the results before and after including the DeFi-related features are compared. It is apparent that all classifiers already worked well to classify the "Good" labels with high precision, recall, and F1-Scores. By taking into consideration the DeFi related features their performance was improved. For example, the precision of the Fine-tuned LLM increased from 0.921 to 0.949.

For the "Malicious" label, there is a significant improvemnet in the performance. It was found that Deep Neural Network, XGBoost and a fine-tuned LLM are the best classifiers for malicious DeFi addresses, with an F1-Score of 0.742 (DNN), a Precision of 0.774 (XGBoost), and a Recall of 0.827 (SVM) and F2-Score of 0.732 (DNN). It was also shown that the introduction of more than 400 DeFi related features improve the classification accuracy significantly; the same models for only transactional non-DeFi related features yields very low F1-Score.

### *G. Feature importance*

We computed and vizualized the importance of each feature of the best model. It was inferred that the most important feature for this task is DeFi-related, while other DeFi-related features have bigger impact than traditional transactional features.

*1) Transactional features:* Overall, the number of transactions, the age of the wallet, and transactions over time are the most relevant features.

*2) DeFi features:* As discussed above, the DeFi features have a strong impact on the results and classification. No feature stands out in particular, and instead many of the engineered features are used in the classifier. This suggests that the classifier is utilizing information that is not immediately apparent, unlike e.g., the transactional features. This can help in the future to detect malicious wallets before they are tagged as such.

## IV. Conclusion

In this paper, we studied how Machine Learning classification methods can be used to detect entities that interacted with DeFi protocols and also have been labeled in dedicated Ethereum annotated datasets. We gathered large datasets of transactions focusing on the major DeFi protocols. We observed that the introduction of DeFi-related features significantly improved the performance of the algorithms, especially for the minority class of the dataset. The authors would like to note that they only analyze publicly available data from existing works in this field as well as public online sources (websites) and do not take part in any labeling or identification processes.

Overall, the Deep Neural Network, followed by XGBoost and the LLM, have the highest F1- and F2-Scores. They hence perform best combining precision and recall. Random Forest has a high precision, however a low recall, whereas the SVM and Logistic Regression have a low precision but high recall.

Regarding future extensions of our system, we consider exploring more machine learning algorithms and more specifically Deep Learning algorithms. Additinal preprocessing techniques will be also integrated within the Machine Learning Framework pipeline and additional methods for oversampling and undersampling will be also considered. DeFi is rapidly evolving and new types of fraud continuously emerge. In our model and feature selection the current state of DeFi fraud detection was examined. Future research should focus on updating and expanding the feature set to keep pace with the evolving DeFi landscape.

## 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] G. Palaiokrassas, P. Skoufis *et al.*, "Combining blockchains, smart contracts, and complex sensors management platform for hyper-connected smartcities: An iot data marketplace use case," *Computers*, vol. 10, no. 10, p. 133, 2021.

[4] Y. Liu, D. He, M. S. Obaidat, N. Kumar, M. K. Khan, and K.-K. R. Choo, "Blockchain-based identity management systems: A review," *Journal of network and computer applications*, vol. 166, p. 102731, 2020.

[5] G. Palaiokrassas, A. Litke, G. Fragkos, V. Papaefthymiou, and T. Varvarigou, "Deploying blockchains for a new paradigm of media experience," in *Economics of Grids, Clouds, Systems, and Services: 15th International Conference, GECON 2018, Pisa, Italy, September 18–20, 2018, Proceedings 15*. Springer, 2019, pp. 234–242.

[6] M. Bartoletti, J. H.-y. Chiang, and A. L. Lafuente, "Sok: lending pools in decentralized finance," in *Financial Cryptography and Data Security. FC 2021 International Workshops: CoDecFin, DeFi, VOTING, and WTSC, Virtual Event, March 5, 2021, Revised Selected Papers 25*. Springer, 2021, pp. 553–578.

[7] Aave, "Aave – open source liquidity protocol," https://aave.com/, 2023, accessed on: November 27, 2023.

[8] MakerDAO, "MakerDAO | an unbiased global financial system, https://makerdao.com/," 2023. [Online]. Available: https://makerdao.com/en/

[9] K. Qin, L. Zhou, P. Gamito, P. Jovanovic, and A. Gervais, "An empirical study of DeFi liquidations: incentives, risks, and instabilities," in *Proceedings of the 21st ACM Internet Measurement Conference*, ser. IMC '21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 336–350. [Online]. Available: https://doi.org/10.1145/3487552.3487811

[10] G. Palaiokrassas, S. Bouraga, and L. Tassiulas, "Machine learning on blockchain data: A systematic mapping study," *Available at SSRN 4530479*.

[11] Y. Zhang, W. Yu, Z. Li, S. Raza, and H. Cao, "Detecting ethereum ponzi schemes based on improved lightgbm algorithm," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 2, pp. 624–637, 2021.

[12] M. Wu *et al.*, "Tutela: An open-source tool for assessing user-privacy on ethereum and tornado cash," *arXiv:2201.06811*, 2022.

[13] F. Béres, I. A. Seres, A. A. Benczúr, and M. Quintyne-Collins, "Blockchain is watching you: Profiling and deanonymizing ethereum users," in *2021 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS)*, 2021.

[14] F. Victor, "Address clustering heuristics for ethereum," in *Financial Cryptography and Data Security: 24th International Conference, FC 2020*. Springer, 2020.

[15] H. Jang and J. Lee, "An empirical study on modeling and prediction of bitcoin prices with bayesian neural networks based on blockchain information," *Ieee Access*, vol. 6, pp. 5427–5437, 2017.

[16] A. H. H. Kabla, M. Anbar, S. Manickam, and S. Karupayah, "Eth-PSD: A machine learning-based phishing scam detection approach in ethereum," *IEEE Access*, vol. 10, pp. 118 043–118 057, 2022, conference Name: IEEE Access.

[17] M. Li, "A survey on ethereum illicit detection," in *Artificial Intelligence and Security: 8th International Conference, ICAIS 2022, Qinghai, China, July 15–20, 2022, Proceedings, Part III*. Springer, 2022, pp. 222–232.

[18] Covalent-API, "Covalent - One unified API. One billion possibilities." 2023. [Online]. Available: https://www.covalenthq.com/

[19] F. Poursafaei, G. B. Hamad, and Z. Zilic, "Detecting malicious ethereum entities via application of machine learning classification," in *2020 2nd Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*. IEEE, 2020, pp. 120–127.

[20] S. Farrugia, J. Ellul, and G. Azzopardi, "Detection of illicit accounts over the ethereum blockchain," *Expert Systems with Applications*, vol. 150, p. 113318, 2020.

[21] M. Weber *et al.*, "Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics," *arXiv:1908.02591*, 2019.

[22] T. Min and W. Cai, "Portrait of decentralized application users: an overview based on large-scale ethereum data," *CCF Transactions on Pervasive Computing and Interaction*, vol. 4, no. 2, pp. 124–141, 2022.

[23] M. Darlin, G. Palaiokrassas, and L. Tassiulas, "Debt-financed collateral and stability risks in the defi ecosystem," in *2022 4th Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*. IEEE, 2022.

[24] S. Bouraga, "To fork or not to fork? bitcoin forks' success analysis using twitter data: Preliminary results," in *2022 4th Conference on Blockchain Research & Applications for Innovative Networks and Services (BRAINS)*. IEEE, 2022.

[25] A. Kyriazis, I. Ofeidis, G. Palaiokrassas, and L. Tassiulas, "Monetary policy, digital assets, and defi activity," *arXiv preprint arXiv:2302.10252*, 2023.

[26] J. Li, F. Baldimtsi *et al.*, "Measuring illicit activity in defi: The case of ethereum," in *Financial Cryptography and Data Security. FC 2021 International Workshops: CoDecFin, DeFi, VOTING, and WTSC, Virtual Event*. Springer, 2021.

[27] A. Trozze, T. Davies, and B. Kleinberg, "Of degens and defrauders: Using open-source investigative tools to investigate decentralized finance frauds and money laundering," *arXiv preprint arXiv:2303.00810*, 2023.

[28] B. Wang, X. Yuan, L. Duan, H. Ma, C. Su, and W. Wang, "Defiscanner: Spotting defi attacks exploiting logic vulnerabilities on blockchain," *IEEE Transactions on Computational Social Systems*, 2022.

[29] CryptoScamDB, "Cryptoscamdb: Tracking malicious urls," https://cryptoscamdb.org/, 2023, accessed on: November 27, 2023.

[30] Etherscan, "Etherscan: The ethereum blockchain explorer," https://etherscan.io, 2023, accessed on: November 27, 2023.

[31] CoinTelegraph, "Tether blacklists 39 ethereum addresses worth over 46 million," https://cointelegraph.com/news/tether-blacklists-39-ethereum-addresses-worth-over-46-million, 2023, accessed on: November 27, 2023.

[32] MyEtherWallet, "Ethereum addresses darklist," https://github.com/MyEtherWallet/ethereum-lists/blob/master/src/addresses/addresses-darklist.json, 2023, accessed on: November 27, 2023.

[33] XBlock, "Xblock transaction datasets," http://xblock.pro/tx, 2023, accessed on: November 27, 2023.

[34] S. Al-E'mari, M. Anbar, Y. Sanjalawe, and S. Manickam, "A labeled transactions-based dataset on the ethereum network," in *International Conference on Advances in Cyber Security*. Springer, 2020, pp. 61–79.

[35] S. Ammari, "Labeled transactions-based dataset of ethereum network," https://github.com/salam-ammari/Labeled-Transactions-based-Dataset-of-Ethereum-Network, 2023, accessed on: November 27, 2023.

[36] G. BigQuery, "BigQuery Enterprise Data Warehouse," 2023. [Online]. Available: https://cloud.google.com/bigquery

[37] I. Alarab and S. Prakoonwit, "Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques," *Data Science and Management*, vol. 5, no. 2, pp. 66–76, 2022.

[38] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, pp. 483–519, 2013.

[39] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," 2000.

[40] scikit learn, "scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation," 2023. [Online]. Available: https://scikit-learn.org/stable/

[41] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[42] OpenAI, "Fine-tuning GPT-3," https://platform.openai.com/docs/guides/fine-tuning, 2023.

[43] ——, "OpenAI GPT models," https://platform.openai.com/docs/models/whisper, 2023.