

AI for on-chain fraud and scam detection

Author: Eligio Haddad¹

*Affiliation: CIRAME Research Center, Business School
Holy Spirit University of Kaslik, P.O Box 446, Jounieh, Lebanon*

Abstract:

The rapid expansion of decentralized finance (DeFi) has introduced unprecedented opportunities for financial inclusion, yet the irreversible and permissionless nature of blockchain networks has simultaneously made them a primary target for sophisticated fraud, including rug pulls, phishing drains, and layering schemes. Traditional manual compliance and rule-based systems are increasingly inadequate against the speed of on-chain attackers. This study proposes an advanced AI-driven detection framework that utilizes temporal graph modeling and behavioral data fusion to identify malicious activities in real-time. Using a substantial dataset of over 1.25 million unique addresses and 18.4 million transactions across the Ethereum, BSC, and Solana blockchains, the research evaluates the efficacy of Graph Neural Networks (GNNs) and unsupervised autoencoders. The methodology emphasizes a "ground truth" labeling process and chronological data splitting to prevent data leakage. Results indicate that the hybrid system—incorporating on-chain transaction traces, smart contract bytecode, and off-chain social signals—achieves an F1-score of 0.81 with a remarkably low false-positive rate of 0.008. Furthermore, the study demonstrates operational efficiency with an inference time of 18ms, making it viable for high-throughput networks. The findings suggest that while AI significantly enhances blockchain security, ongoing challenges in "explainability" and adversarial obfuscation require a "human-in-the-loop" approach to maintain long-term ecosystem trust.

Keywords:

Blockchain Security, Fraud Detection, Graph Neural Networks (GNN), Smart Contracts, Decentralized Finance (DeFi), Machine Learning, Real-time Monitoring, Rug Pulls, Temporal Analysis, Anomaly Detection.

Introduction: "The What & Why"

The rise of blockchain and cryptocurrencies has fundamentally shifted how we think about digital finance by introducing decentralized, transparent, and permissionless systems. These innovations offer a real chance for better financial inclusion and the use of programmable assets. However, the nature of these networks, combined with their global reach and the fact that you can't reverse a transaction, has unfortunately made them a primary target for fraud and money laundering. We are seeing an increase in high-profile rug pulls, phishing drains, and complex layering schemes that exploit both technical bugs and human psychology. These activities don't just cause massive financial losses; they erode the trust needed for the industry to grow and often trigger reactionary regulations that could unintentionally stifle genuine progress. Standard compliance tools that rely on manual reviews, if-then rules simply cannot keep up with the speed and evolving tactics of on-chain attackers. This is where AI becomes essential. AI allows us to analyse big amount of data to spot weird behaviour and map out relationships between entities in real-time. By blending transaction graphs with behavioural data, we can detect fraud more accurately and respond faster.

1. Categories the most common on chain scams and pinpoint the specific transaction fingerprints that define each one.
2. Assemble a comprehensive dataset that links on chain transaction graphs with smart contract metadata and external signals for training.
3. Evaluate different graph based temporal models to see which best represents how wallets and contracts behave over time.
4. Build and stress-test detection pipelines, focusing specifically on their accuracy and speed.
5. Analyse how well these models hold up against adversarial tactics, such as when scammers try to mask their transactions or use new obfuscation techniques.
6. Review the practical side of deployment, specifically looking at how much computing power is needed for real time monitoring.

Literature Review: "The story so far "

Research into using AI to flag fraud scams on public blockchains has grown really fast over the last few years. This surge is largely a respond to the rise in illicit activity and the fact that chains like Ethereum, BSC, and Solana offer a goldmine of publicly accessible transaction data. In the early days, most work focused on supervised classifiers. These models looked at basic stats to try and separate malicious actors from legitimate users.

More recently, the field has moved toward graph-based approaches. Instead of just looking at flat data points, these methods treat addresses and transactions as part of a giant web. By using Graph Neural Networks researchers can map out relational patterns and community structures that simple models miss.

Timing has also turned out to be a critical factor. In this Data-Driven cutlture (Azoury et al., 2024) Researchers are now using sequence models and GNN to catch sudden transaction spikes. At the same time, there is a whole other branch of research focusing on the logic of smart contracts. By analyzing source code, experts can identify vulnerable or malicious contracts before they cause damage.

We are also seeing the rise of hybrid system that mix on-chain data with off-chain intelligence. These combined models usually have much better luck catching scams early while keeping false alarms to minimum. A major run that still exists for everyone is the lack of high-quality, labelled data. Most public datasets are small and usually only cover very famous hacks, which makes it hard for an AI to learn how to spot a brand-new type of scam.

As result, many are turning to unsupervised anomaly detection to find novel attacks, though these systems still struggle with higher error rates. On top of that, scammers are getting smarter and are now trying to hide their patterns to trick AI filters. This is pushing the industry toward more robust. Looking ahead, the real challenge is making these systems fast enough to work in real time on high speed blockchains while ensuring the results are clear enough for a human to actually use.

Methodology: "The how we did it "

We designed this study around a transparent, reproducible workflow to ensure that other researchers can verify and replicate our results in the future. This allowed us to build a comprehensive dataset containing transaction traces, token transfer events, smart contract bytecode, and detailed account metadata for the study period.

Establishing a reliable ground truth for our labels was a top priority. We combined automated matching against known malicious addresses with a hard manual verification process. To minimize errors, we only assigned a positive label if a threat was confirmed by at least two independent sources. For our negative samples, we selected long-lived accounts that had no history of reports. We also implemented quality control measures, including random audits of the data to maintain consistency.

The preprocessing stage was extensive. We normalized all timestamps to UTC, eliminated duplicate traces to get a full view of contract interactions. We then aggregated features into per-address and per-contract windows and generated semantic embeddings for contract logic using transformer-based encoders. For the graph construction, we developed directed multi-edge transaction graphs that included temporal attributes. We also applied specific heuristics for bridges and decentralized exchanges to clearly distinguish between standard users and infrastructure like mixers or exchanges.

When it came to training, we used strictly chronological splits instead of random shuffling. This is a critical step in time-series data to prevent data leakage, ensuring the model never learns from future events it wouldn't have access to in a real-world scenario. Finally, we addressed ethical and legal concerns by strictly excluding any personally identifiable off-chain information and adhering to all platform terms of service for data scraping. By documenting every step of our feature computation and model training, we aim to provide a clear roadmap for independent researchers to evaluate or build upon our work.

Results: "The facts "

Our final dataset was substantial, covering more than 1,250,000 unique addresses and a total of approximately 18,450,372 transactions across the Ethereum, BSC and Solana blockchains. Within this pool, we successfully identified and labelled 12,430 malicious addresses, which accounted for 48,760 specific transactions. To add depth to the on-chain data, we also integrated

8,900 records containing off-chain social signals, providing a more holistic view of how these scams operate outside of the ledger itself.

In terms of model performance, the unsupervised autoencoder showed promising results as a first line of defense. It managed to flag 54% of known malicious entities without any prior training on those specific cases.

The descriptive analysis also revealed some striking patterns in how these scammers behave. We found that the median lifespan of a malicious address is only 6 days, highlighting the hit-and-run nature of these attacks. For rug-pull contracts specifically, the median initial outflow was a staggering 94.3% of total funds, usually occurring shortly after the contract's creation. We also noted that 41% of malicious contracts utilized proxy or forwarder patterns to obfuscate their true logic and source of funds.

From an operational standpoint, the system proved to be highly efficient and capable of real-time monitoring. Feature extraction took just 0.45 seconds per block, and the inference time for a single address was a lightning-fast 18ms. This allowed our pipeline to maintain a throughput of 2000 addresses per second, ensuring it could keep up with the high-velocity traffic seen on modern high-throughput blockchains.

Discussion and conclusion: "The so what ? "

Our findings clearly demonstrate that the most effective way to secure a blockchain is to look at the big picture. Models that combine transaction history, the evolving structure of the wallet graph, smart-contract code logic, and external off-chain signals consistently outperform more narrow approaches. Specifically, our hybrid system reached a recall of 0.78 and an F1-score of 0.81, while maintaining a very low false positive rate of 0.008. These results directly answer our research question: fusing multiple types of signals with temporal graph modelling provides a significant leap in accuracy and operational precision compared to traditional, single-signal baselines.

However, no system is perfect, and this study highlights several ongoing challenges. We still have to contend with noisy labels and a natural imbalance in the data, where legitimate transactions far outnumber scams. There are also hurdles regarding how well these models generalize across different chains and their vulnerability to obfuscation the clever ways scammers try to hide their tracks.

Looking ahead, the research community should prioritize several key areas:

- Data quality: developing larger, high-quality and temporally labelled benchmarks that cover multiple chains.
- Adaptability: creating models that use continual learning and domain adaptation to keep up with scammers who change their tactics overnight.
- Explainability: moving beyond black box AI to provide scoring and alerts that actually make sense to a human investigator.
- Scale: designing streaming implementations that can handle the massive throughput of modern chains without sacrificing real-time responsiveness.

In conclusion, our study strongly supports the deployment of fused, temporal-graph systems as a more reliable way to detect on-chain fraud. While these tools offer a powerful defence, they aren't a set-and-forget solution. They require ongoing dataset curation, rigorous robustness testing, and most importantly, human in the loop verification to stay ahead of the curve.

References

- Azoury, N., Subrahmanyam, S., & Sarkis, N. (2024). The Influence of a Data-Driven Culture on Product Development and Organizational Success through the Use of Business Analytics. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 15(2), 123-134.