Abstract

Rug pull scams are among the most damaging forms of fraud in decentralised finance (DeFi), with global losses exceeding $27 billion. These scams erode trust in blockchain ecosystems, complicate regulatory oversight, and undermine investor confidence. Existing detection tools such as CRPWarner, TrapdoorAnalyser, and RPHunter show the feasibility of automated analysis, yet each suffers from critical limitations, including reliance on scarce verified source code, narrow coverage, and high computational costs. These gaps highlight the need for scalable approaches that adapt to evolving attack strategies.

This dissertation investigates whether rug pull scams can be detected and categorised automatically using transaction and bytecode data, even without source code. A framework was designed that integrates supervised classification, anomaly detection, and semi-supervised self-learning. The system fuses multiple feature modalities—transaction statistics, opcode sequences, and sequential behaviours—within a late-fusion ensemble whose weights are optimised using Optuna. Self-learning expands labelled datasets with high-confidence pseudo-labels, addressing data scarcity and enabling adaptability to new scam categories.

The framework was evaluated on CRPWarner, RPHunter, and Trapdoor datasets. On CRPWarner ground truth, the system achieved strong performance (macro F1 $\approx$ 0.91). On RPHunter, results were more moderate (macro F1 $\approx$ 0.56), reflecting severe class imbalance. On a 2% Trapdoor sample, the framework generalised effectively (macro F1 $\approx$ 0.94). The anomaly detection module achieved perfect recall, acting as a fail-safe for unseen behaviours. These findings confirm that rug pulls leave detectable blockchain patterns and that multi-source fusion improves robustness. Contributions include a reproducible framework, a pseudo-labelling mechanism for dataset expansion, and insights into runtime trade-offs for deployment.