
Abstract

Machine-learning malware detectors can achieve strong predictive performance, but accuracy alone does not make their decisions understandable or defensible. Explanations should reflect the detector’s reasoning, reveal decision structure, and connect evidence to meaningful cybersecurity concepts. This thesis progresses from post-hoc analysis of black-box classifiers to relational detectors whose predictions can be interpreted through their own computation.

The work extends post-hoc attribution with interaction analysis and multi-feature symbolic rules, develops Tailored Logic Explained Networks for compact and faithful explanations, and introduces PaTEx, which uses regional surrogate trees to expose paths, interactions, alternatives, and counterfactuals. It then investigates ontology-derived knowledge graphs for relational malware characterisation. The principal contribution, QuinEx, is an intrinsically decomposable Matrix Product State knowledge-graph embedding whose score separates exactly into additive channels. This enables exact channel attribution, margin-complete behaviour allocation, and model-aligned concept rules, while MITRE ATT&CK tactic coherence independently measures cybersecurity-domain alignment.

Experiments on static and dynamic representations show competitive performance with explanation capabilities absent from strong baselines. On the primary MAECO-Lite graph, QuinEx achieved ROC-AUC of 0.908 ± 0.027 and F1 of 0.839 ± 0.027 across ten seeds. Its concept rules achieved F1 of 0.832 ± 0.017 and tactic coherence of 0.813 ± 0.112 , with higher measured coherence than LIME, SHAP, and decision-tree explanations. A secondary static EMBER ontology experiment confirmed competitive performance.

Overall, the thesis distinguishes predictive performance, faithfulness, granularity, compactness, and domain alignment as jointly important properties, advancing malware explanation toward structured and computationally grounded decision support.