Joseph Autorino

CSN 190

Professor Edwin Reed Sanchez

10/15/25

4.2 Abstract Draft with Implementation Focus

Explainable ML for malware detection on endpoints detects achieve high accuracy but lack explainability for analysts, explainable models reduce triage time and false positives.

Train 3 interpretable models XGBoost, SHAP, rule-based learners and attention based sequence model on a dataset of 200k labeled telemetry events. Evaluate on 40k holdout set and measure F1, false positive rate and average analyst triage time reduction.

Deliverable models SHAP explainers, analyst UX prototypes, reproducible evaluation code.

This project builds and evaluates interpretable ML detectors for endpoint telemetry trains models on 200,000 labeled events and tests on a 40,000 event holdout and measure whether explanation tools SHAP, attention visualization, rule extraction reduce analyst triage time by >30% while keeping detection performance high. Malware detection increasingly relies on opaque ML models that burden analysts with uninterpretable alerts

and high false positive costs. Delivering accurate, fast and human readable explanations improves triage speed, reduces analyst workload and increases operational trust making ML detectors practical at scale for SOCs and small IR teams.