Misleading Metrics: On Evaluating ML for Malware with Confidence

Roberto Jordaney, Zhi Wang, Davide Papini, Ilia Nouretdinov, Kumar Sharad, and Lorenzo Cavallaro

s2lab.isg.rhul.ac.uk – s2lab.isg.rhul.ac.uk/projects/ce/

Research Problem

- Machine learning (ML) promising when applied to the malware domain.
- In-lab experiments (CV, leave-one-out) show performance metrics, but:
  - Malware is an evolving target.
  - Unclear how ML would really perform once deployed.
- Real-world deployment hard to assess.
  - New, previously unseen labels.
  - Changes in the data distribution.

We need statistical (quality) metrics

Conformal Evaluator: ML Evaluation with Confidence

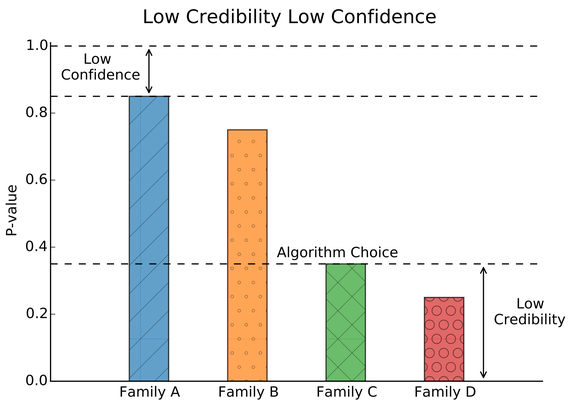
Build statistical metrics (RQ1) to highlight:

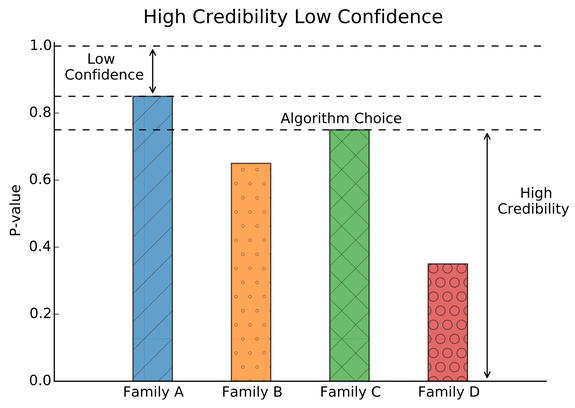
- ***Statistical data distribution according to the algorithm*** (RQ2.1)
  - Shows whether classes overlap.
  - Provides insights during ML design (RQ3).
- ***Statistical confidence on the ML algorithm choices*** (RQ2.2)
  - High confidence puts trust on the ML process (RQ3).
  - Provides thresholds to detect and contain ML decay during deployment (RQ4).

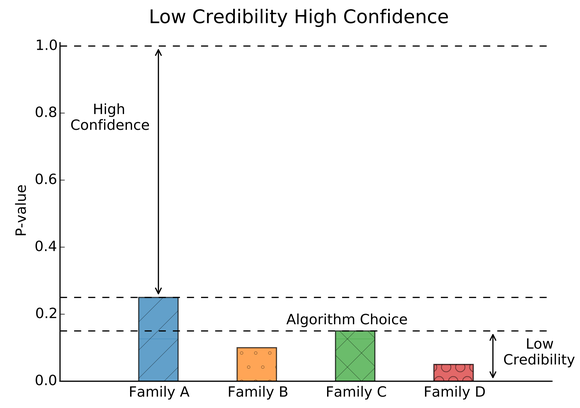
RQ1: Statistical Metrics

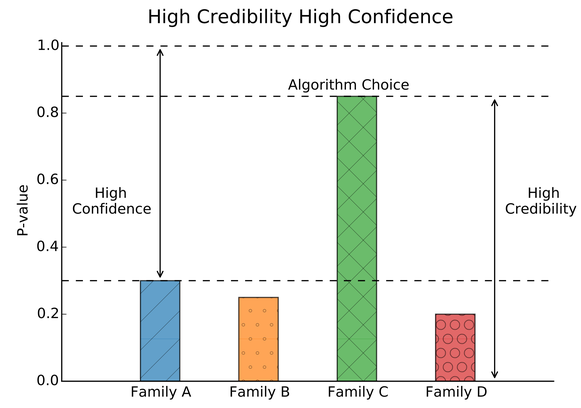
Algorithm credibility – Captures the conformance of the algorithm's prediction for a given data point.

Algorithm confidence – Measure the distinguishability of a sample's prediction among the other families.

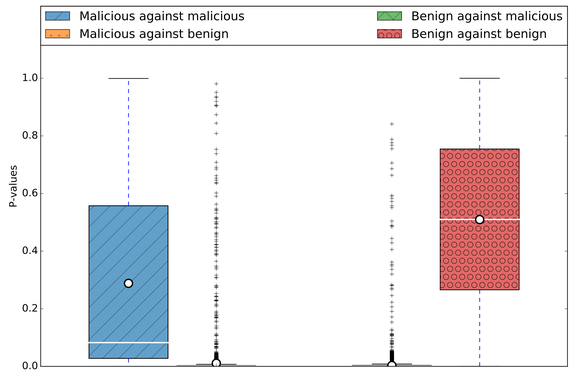




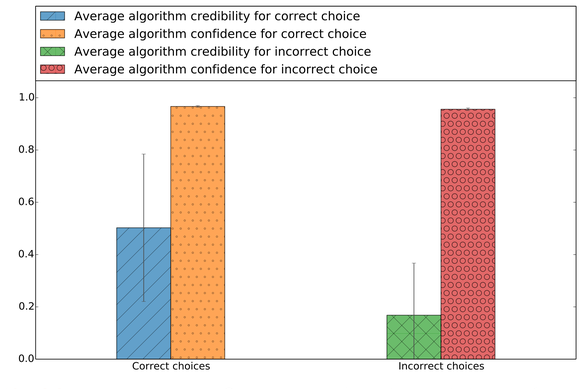




RQ2, RQ3: Statistical Metrics Insights (Design)

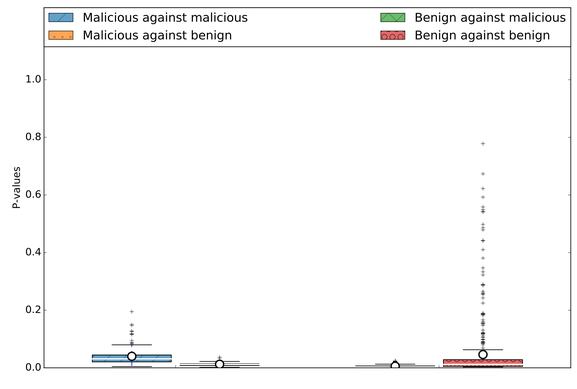


Alpha assessment – Shows the statistical distribution of data according to the algorithm under evaluation.



Decision assessment – Shows the quality with statistical confidence of correct and incorrect choices of the algorithm under evaluation.

RQ4: Statistical Indicators of ML Decay (Deployment)



Testing in real-world scenarios – Statistical metrics provide indications on ML decays when no labels are available.

This research has been supported by the UK research grants EPSRC EP/K033344/1, EP/K006266/1, and EP/L022710/1.