DRMD: Deep Reinforcement Learning for Malware Detection under Concept Drift

Shae McFadden[1,2,3], Myles Foley[2], Mario D'Onghia[3], Chris Hicks[2], Vasilios Mavroudis[2], Nicola Paoletti[1], Fabio Pierazzi[3]
[1]King's College London, [2]The Alan Turing Institute, [3]University College London

Traditional Malware Detection

Thousands of new apps per day

Limited capacity for manual review

Concept Drift

ML Assumption: data is stationary

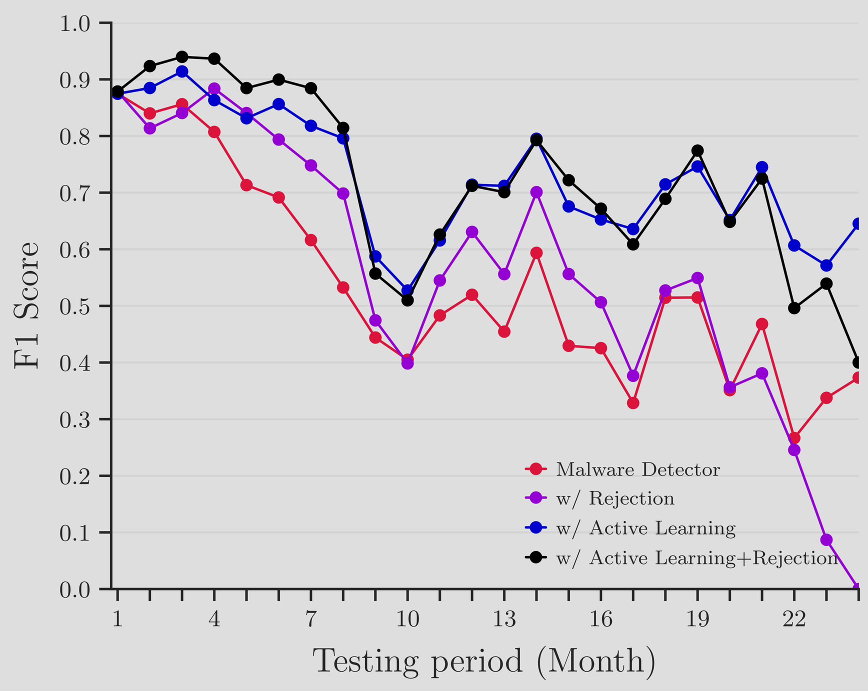
Reality: apps constantly evolve

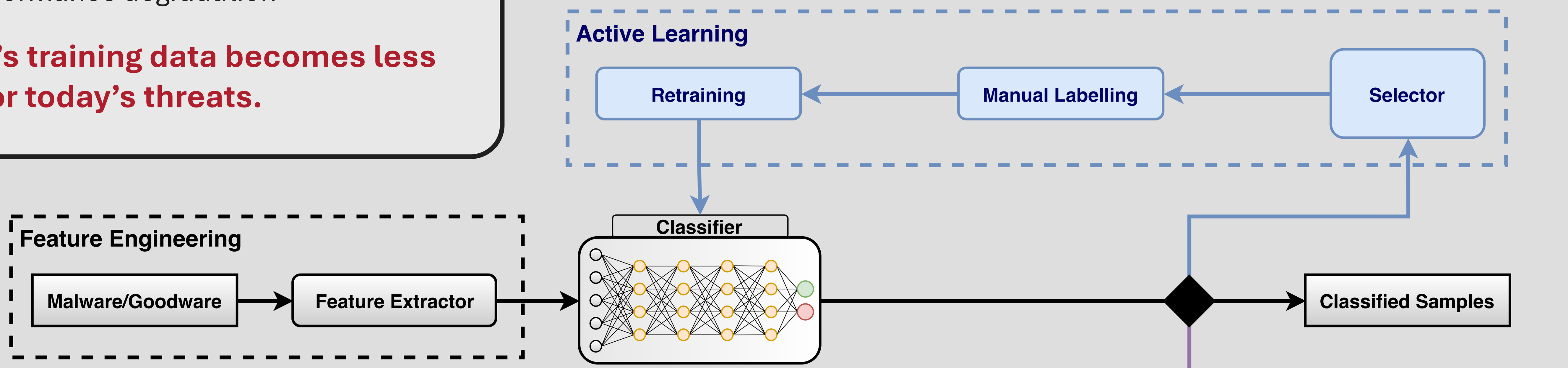
Result: Performance degradation

Yesterday's training data becomes less relevant for today's threats.

Active Learning: Adapting the Detector to Drift

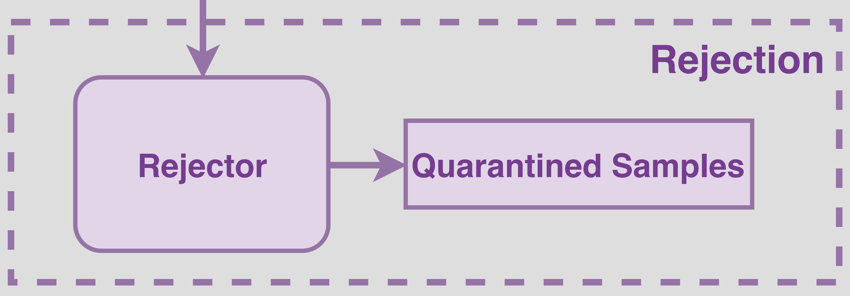
Selects an informative subset of new samples for retraining.





Rejection: Limiting the Impact of Drift

Selects samples at a high-risk of being misclassified to be quarantined.



Key Observation

Existing approaches treat active learning, rejection, and detection independently

Intuition

Treat malware detection as a **unified** decision-making problem and use deep reinforcement learning

DRMD: Deep Reinforcement Learning Malware Detection

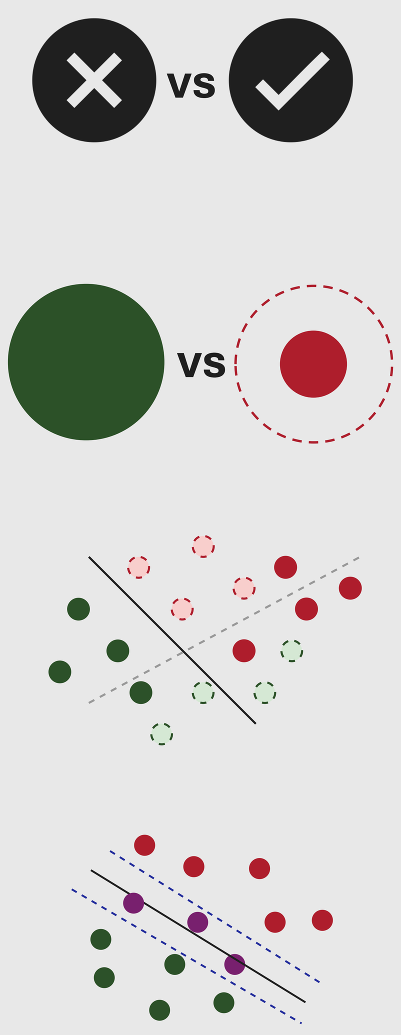
Rewards

Accuracy
Provides the foundation
+1 correct, -1 incorrect

Class Imbalance
Upscales rewards for malware based on distribution (~10%)

Temporal Robustness
Upscales rewards for samples based on temporal position

Rejection
Balances rewards for rejection relative to misclassification risk



Formulation (MD-MDP)

One-step MDP (Contextual Bandit)

Corrects spurious dependencies of prior work, ICMDP [Appl. Intell.'20]

Action Space

✓ Classify as Goodware
✗ Classify as Malware
? Reject → Active Learning

Experimental

Feature Spaces: *Drebin* (10,000D) and *Ramda* (379D)
Datasets: *Hypercube* (2021-2023) and *Transcendent* (2014-2018)
AMD Baselines: *Drebin* (SVM), *DeepDrebin* (MLP), and *Ramda* (MLP+VAE)
DRL Baselines: *ICMDP* and *DCBs*

MDP Comparison

Same CO policy architectures MD-MDP outperforms ICMDP

97%	45%	+1.94
settings	significant	ΔAUT

Classifier Comparison

Same AL and rejection budgets DRMD outperforms Baselines

90%	79%	+8.66
settings	significant	ΔAUT

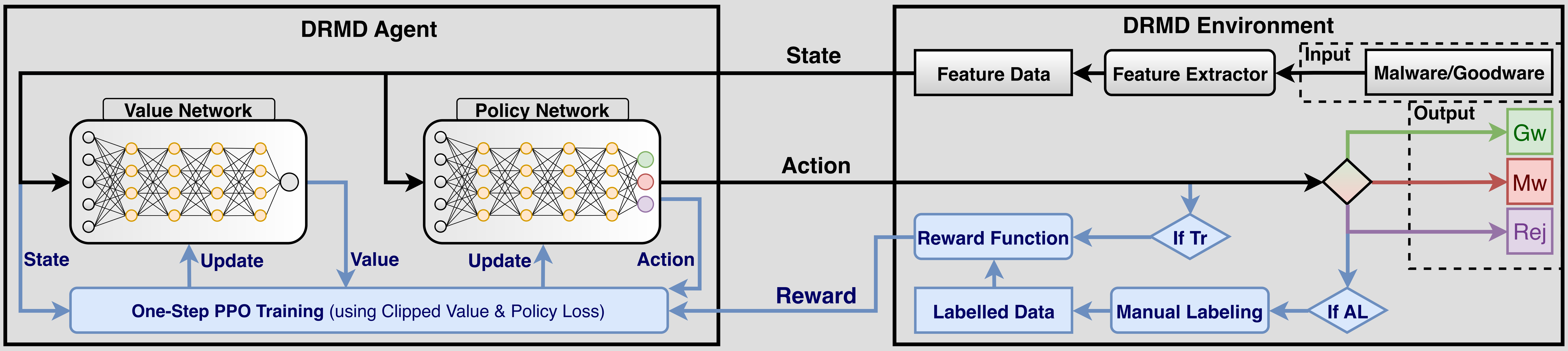
Pipeline Comparison

Same AL and rejection budgets DRMD outperforms Baselines

81%	68%	+10.90
settings	significant	ΔAUT

Takeaways

1) Adaptive decision-making, not just classification
2) One-step MDP formulation
3) Concept drift-aware DRL
4) Integration that matters
5) A starting point for future research



AAAI-26 / IAAI-26 / EAAI-26
JANUARY 20-27, 2026 | SINGAPORE

Partially Funded By
National Cyber Security Centre
a part of GCHQ
EPSRC
Engineering and Physical Sciences Research Council

Paper Code Video