Detecting Concept Drift in Malware Classification Models

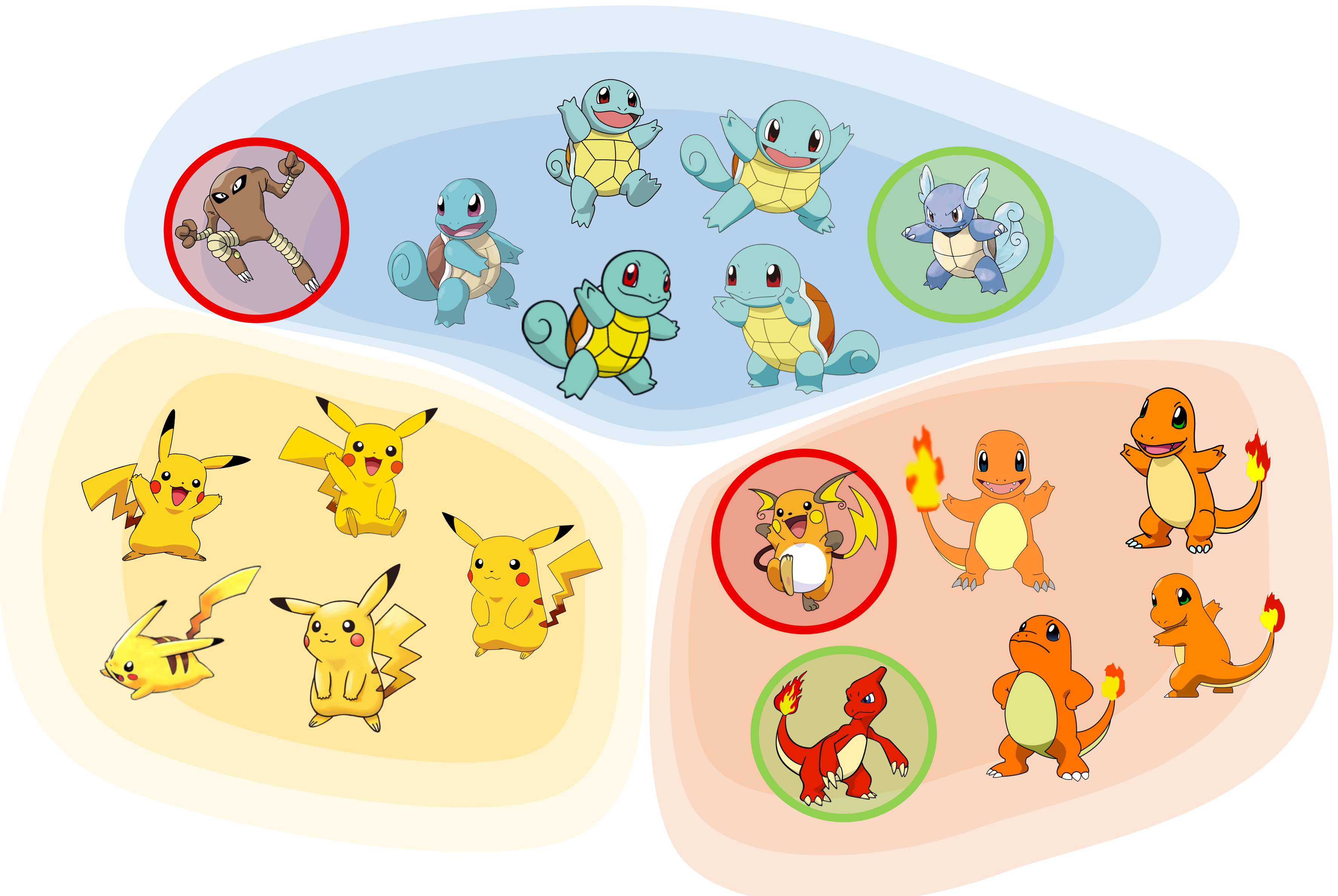
Roberto Jordaney, Feargus Pendlebury, Fabio Pierazzi, Lorenzo Cavallaro

Royal Holloway, University of London and King's College London

Research Objective

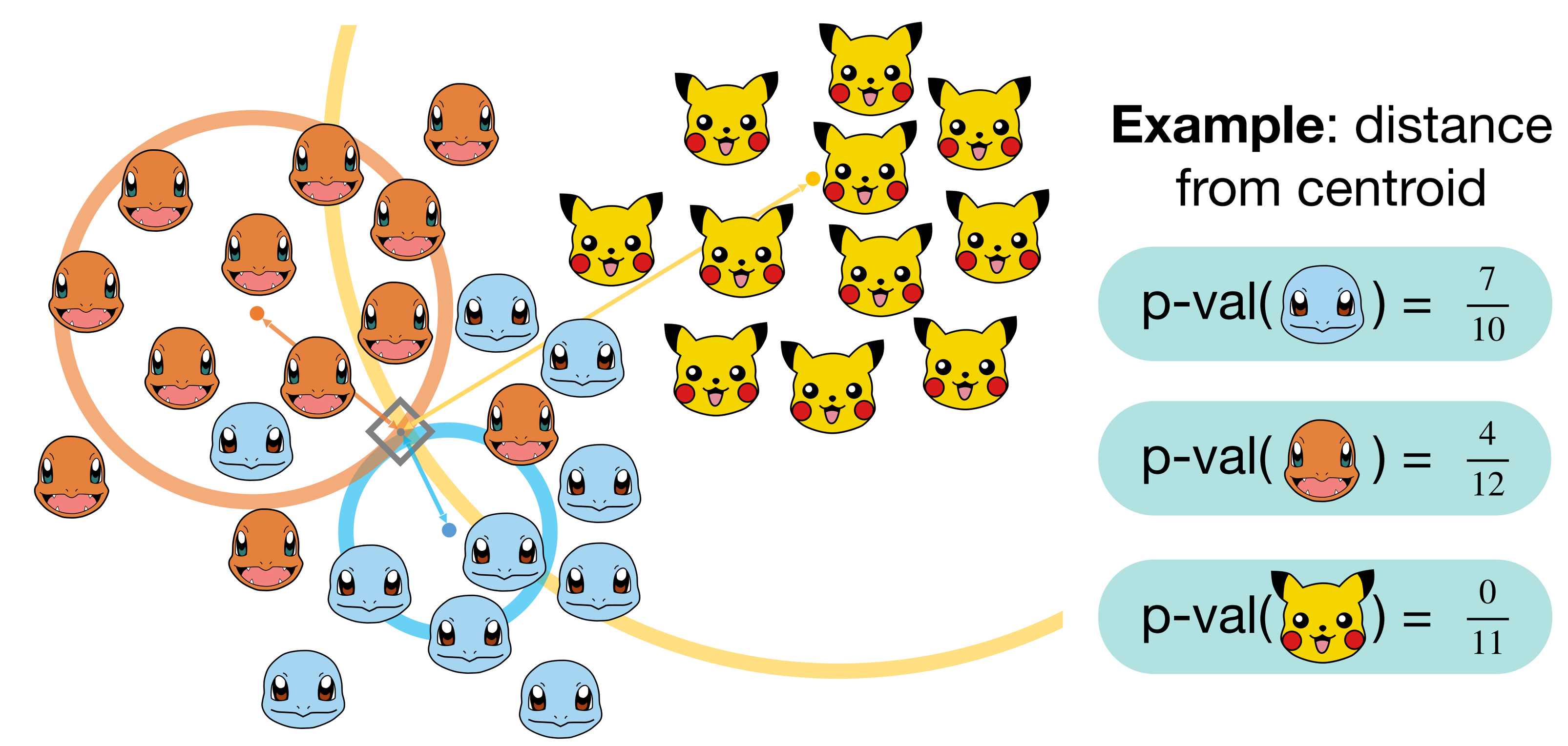
Identify “**aging**” in ML models for malware classification – *which predictions can be trusted?*

Concept Drift: After a model is trained, malware may evolve, and new malware families may arise.



Intuition: Do objects “fit” well into the predicted class(es)?

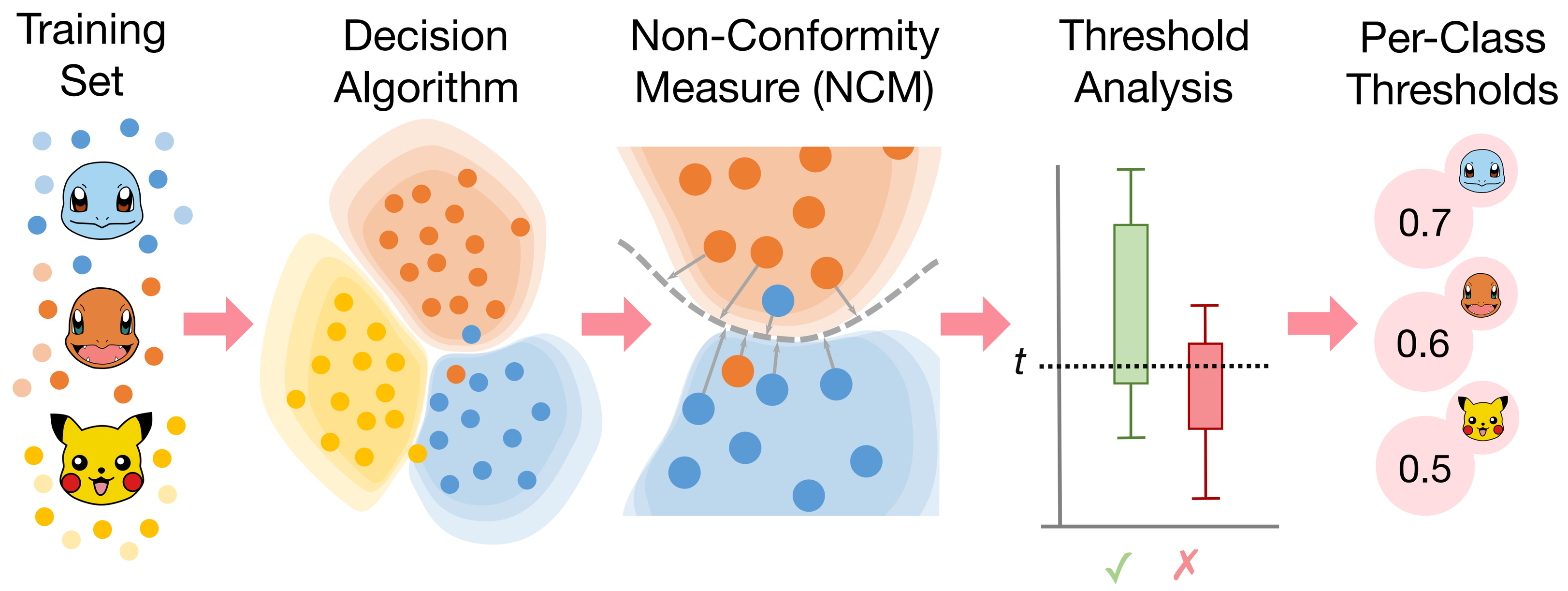
P-value: The ratio of the number of training elements less similar than the element under test to total class members.



Solution

Algorithm-agnostic, uses a score from the ML classifier
Conformal Evaluator (CE) statistically evaluates classifiers
Per-class thresholds identify unreliable predictions

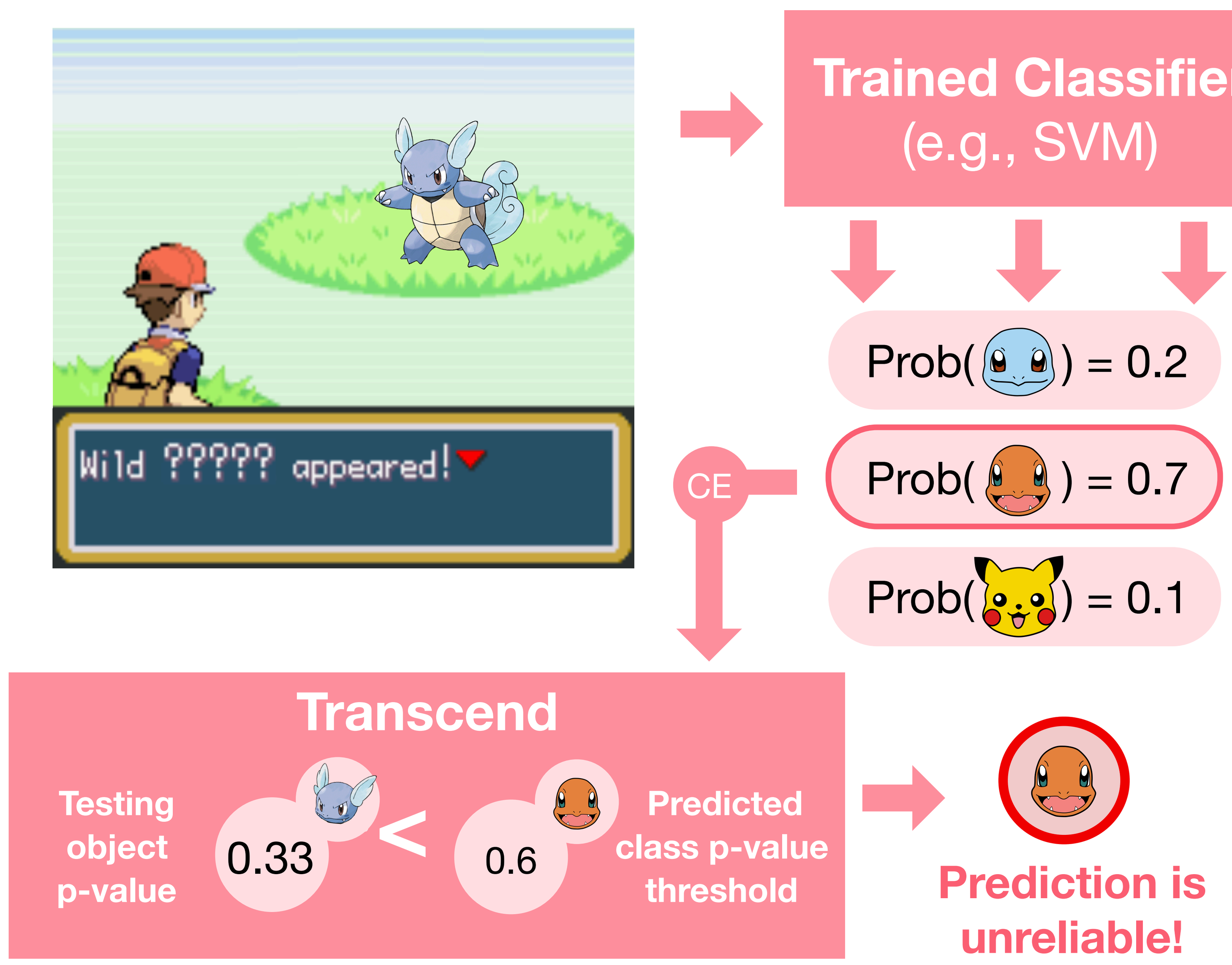
Training and Validation Phase



Transcend derives per-class thresholds by solving an optimization problem to achieve a trade-off between the **performance** and the **number of rejected elements**.

Testing Phase

What happens when a new object arrives?



R. Jordaney, K. Sharad, S. K. Dash, Z. Wang, D. Papini, I. Nouretdinov, L. Cavallaro
Transcend: Detecting Concept Drift in Malware Classification Models.
Proceedings of USENIX Security, 2017 · https://s2lab.kcl.ac.uk/projects/ce/

Experiments on Malware

Binary Classification

Drebin Dataset		Marvin Dataset	
Type	Objects	Type	Objects
Benign	123,456	Benign	9,592
Malware	5,560	Malware	9,179

	TPR reliable predictions		TPR unreliable predictions		FPR reliable predictions		FPR unreliable predictions	
	p-value	prob.	p-value	prob.	p-value	prob.	p-value	prob.
1st quartile	0.9045	0.6654	0.0000	0.3176	0.0007	0.0	0.0000	0.0013
Median	0.8737	0.8061	0.3080	0.3300	0.0000	0.0	0.0008	0.0008
Mean	0.8737	0.4352	0.3080	0.3433	0.0000	0.0	0.0008	0.0018
3rd quartile	0.8723	0.6327	0.3411	0.3548	0.0000	0.0	0.0005	0.0005

Without Transcend

Sample	Benign (Predicted label)	Malicious (Predicted label)	Rec.
Benign	4,498	2	1
Malicious	2,890	1,610	0.36
Prec.	0.61	1	

With Transcend

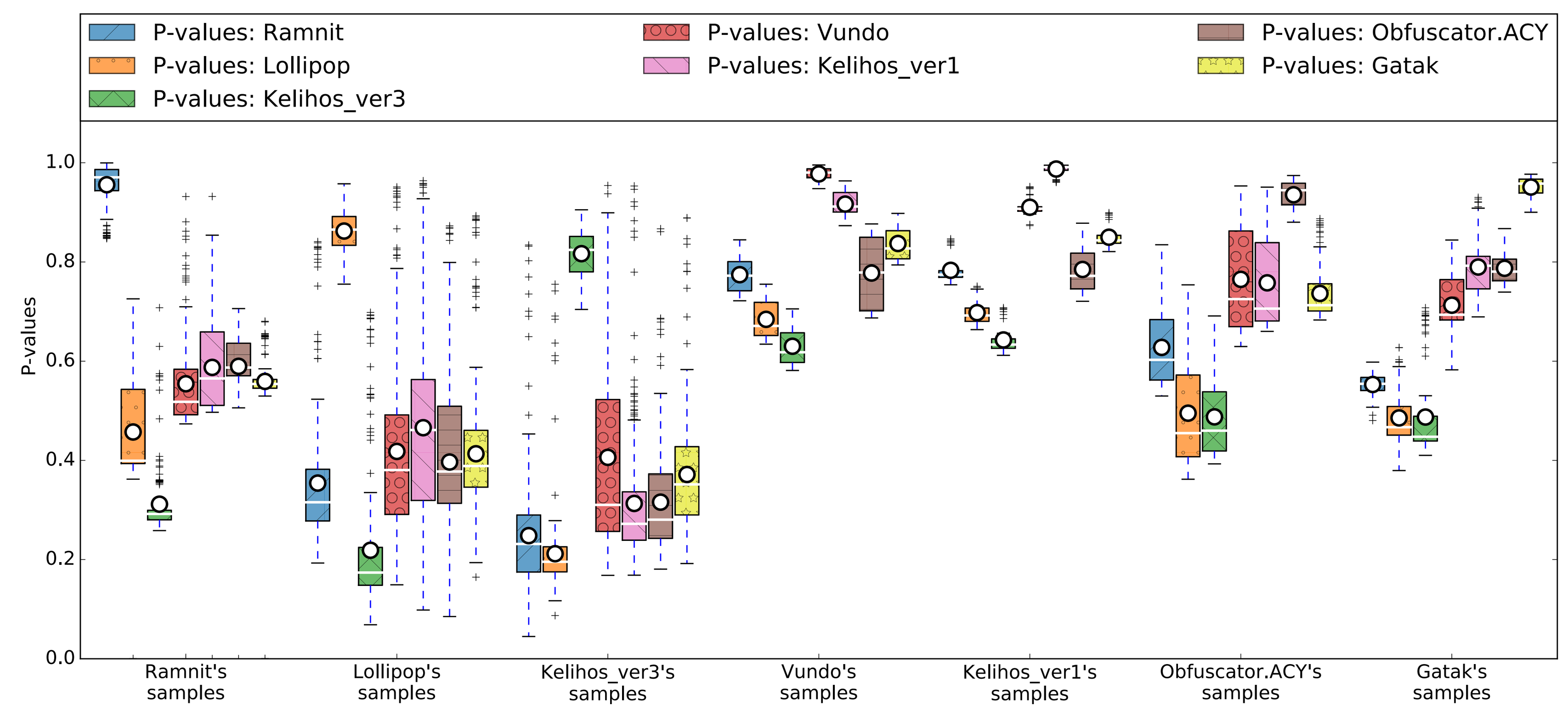
Sample	Benign (Predicted label)	Malicious (Predicted label)	Rec.
Benign	4,257	2	1
Malicious	504	1,610	0.76
Prec.	0.89	1	

Sample	Benign (Predicted label)	Malicious (Predicted label)	Rec.
Benign	4,413	87	0.98
Malicious	225	4,245	0.94
Prec.	0.96	0.98	

Multi-class Classification

Microsoft Malware Classification Challenge Dataset

Malware	Objects	Malware	Objects
Ramnit	1,541	Obfuscator.ACY	1,228
Lollipop	2,478	Gatak	1,013
Kelihos`ver3	2,942	Kelihos`ver1	398
Vundo	475	Tracur	751



CE’s p-values help to reveal the quality of predictions made by the decision algorithm.

This research has been partially supported by UK EPSRC grants EP/K033344/1, EP/L022710/1, and EP/K006266/1, as well as the NVIDIA Corporation.

NVIDIA · EPSRC Pioneering research and skills · SYSTEMS SECURITY RESEARCH LAB