Co-funded by the European Union

Investing in your future

IEEE CSR Cyber Security and Resilience

TAL TECH

Active learning-based mobile malware detection utilizing auto-labeling and data drift detection

Zhe Deng, Arthur Hubert, Sadok Ben Yahia, Hayretdin Bahsi

zhe.deng@taltech.ee
arthur.hubert@uni.lu
sadok.ben@taltech.ee
hayretdin.bahsi@taltech.ee

2024/09/02

Motivation

- The ubiquity of Mobile Devices and Concerns about safety and privacy
- Challenges in Malware Detection
  - Mobile malware's dynamic and static characteristics are constantly changing, making it challenging to maintain accurate detection models over time
- Non-stationary model
  - Active learning can help in retraining models periodically to adapt to these changes and manage data drift
- Cost of Data Labeling
  - The scarcity of labeled cybersecurity data due to privacy concerns and high labeling costs makes auto-labeling crucial for efficient and cost-effective mobile malware detection.

TAL TECH

Overview

- Introduction
- Background
- Methods
- Results
- Conclusion

Introduction

This study introduces a novel pool-based active learning method combined with auto-labeling to adapt to evolving malware threats, achieving high detection accuracy with minimal labeled data.

Active learning is a semi-supervised strategy that is particularly useful when collecting data is easy but labeling data is expensive.

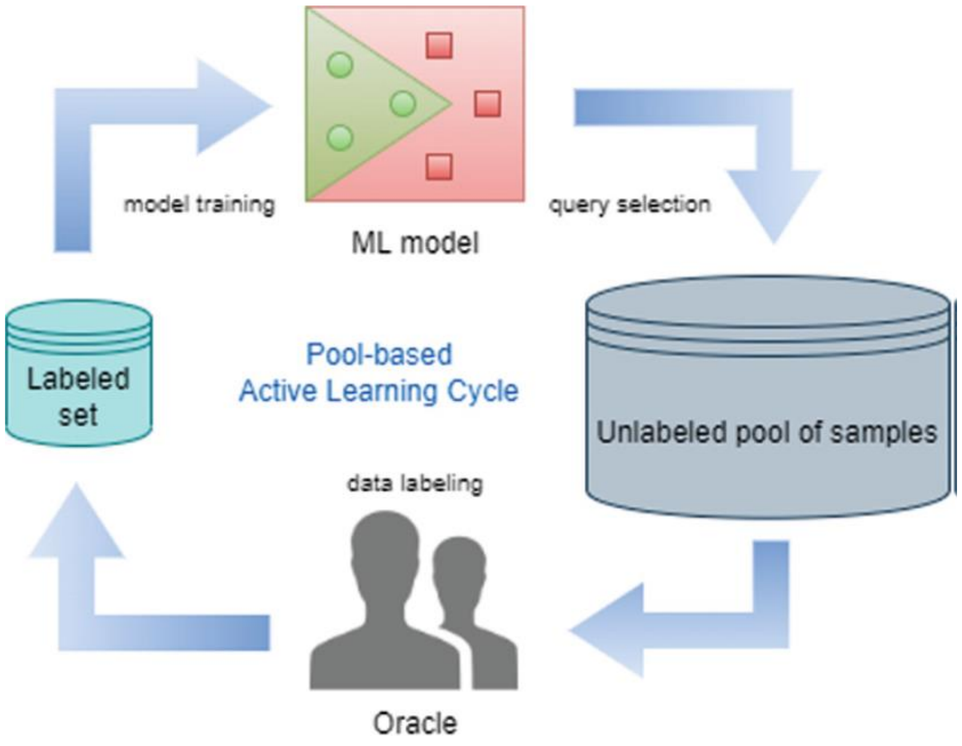
This data is used to train our model by selecting informative samples from a large, unlabeled pool.

TAL TECH

Background

Pool-based active learning:

- 1. Initialization & Model Training
- 2. Query Strategy & Data Selection
- 3. Labeling (Annotation)
- 4. Model Update & Evaluation
- Iterations



Guerra-Manzanares, A., Bahsi, H. (2023). On the Application of Active Learning to Handle Data Evolution in Android Malware Detection.

Methods

Dataset:

- *KronoDroid*

Two basic balancing techniques:

- Oversampling
- Undersampling

Divide the dataset into 44 smaller sub-datasets following the timeline of the samples

Data	Size	Description
Benign samples	36,755	Time frame: 2008-2020
Malware samples	41,382	Time frame: 2008-2020
Permissions	166	Categorical (binary) features
System calls	288	Numeric features
Hybrid (perms+ syscalls)	454	Binary and numeric features
Timestamps	-	*First Seen* and *Last Modification*

Guerra-Manzanares, A. & Bahsi, H. & Nõmm, S. (2021). KronoDroid: Time-based Hybrid-featured Dataset for Effective Android Malware Detection and Characterization. Computers & Security.

TAL TECH

Methods

Active learning-based mobile malware detection utilizing auto-labeling and data drift detection

Three types of training:

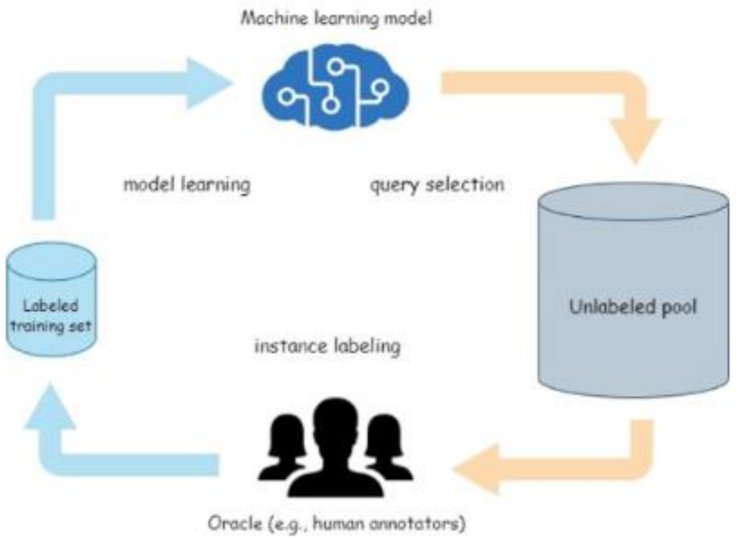
- Active learning with uncertainty sampling
  - Our main strategy

$$U(x) = 1 - P(y * |x)$$

- Batch retraining
  - As an upper limit we aim to reach
- Active learning with random sampling
  - A lower limit

Auto-labelling

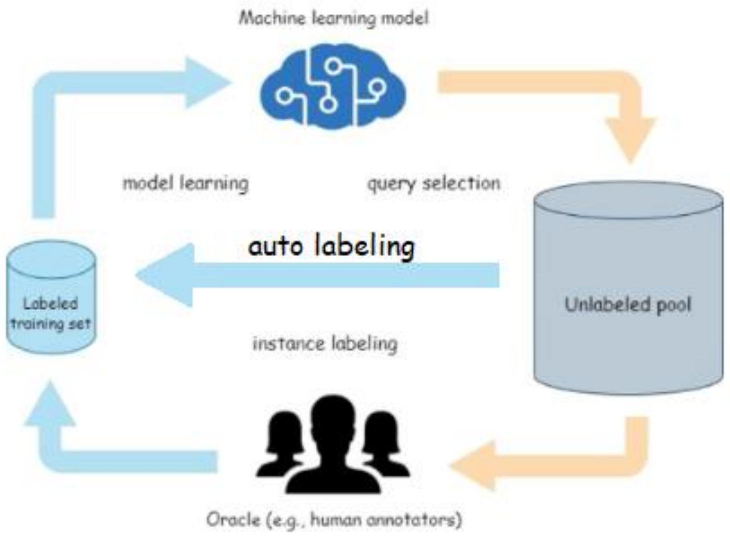
Data drift

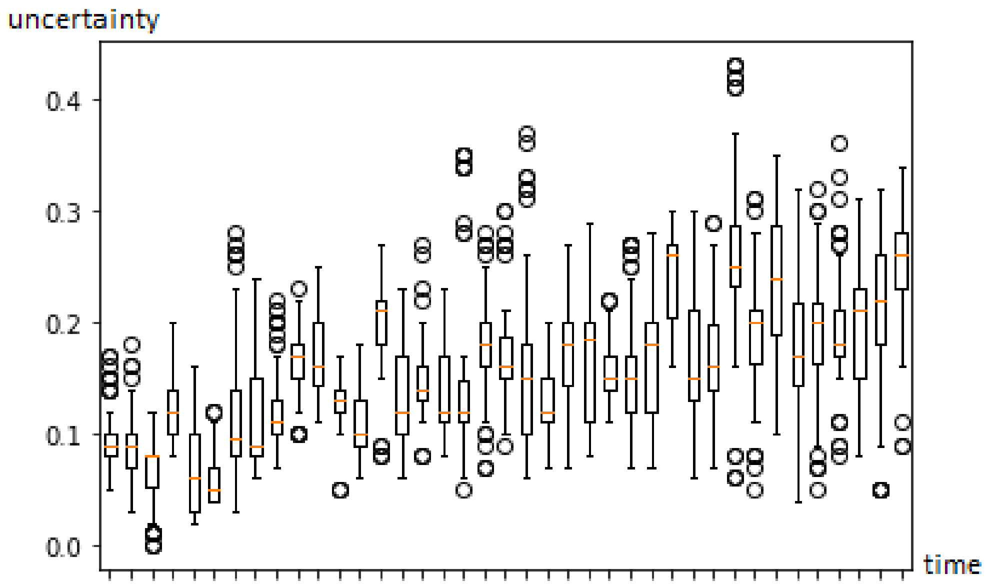


Methods

Active learning-based mobile malware detection utilizing auto-labeling and data drift detection

Auto-labelling





Data drift

$$P_t(x) \neq P_{t+1}(x) \rightarrow P_t(y|x) \neq P_{t+1}(y|x)$$

Results

Baseline:

TABLE I: Baseline: benchmark training results

Feature	Balancing Method	Training Strategy	Label Numbers	Label Proportion(%)	F1(%)	Accuracy(%)
Permission	Oversampling	Batch	31723	100.0	99.0	98.1
		Random	3020	9.5	93.9	95.2
		Uncertainty	1501	4.7	95.0	96.0
	Undersampling	Batch	31723	100.0	99.0	98.1
		Random	3292	10.4	93.6	95.6
		Uncertainty	1983	6.2	94.3	95.8
System call	Oversampling	Batch	31723	100.0	98.1	96.5
		Random	9841	31.0	89.6	90.5
		Uncertainty	6694	21.1	91.2	92.7
	Undersampling	Batch	31723	100.0	98.1	96.5
		Random	10108	31.8	89.1	89.6
		Uncertainty	7470	23.5	90.9	92.2
Hybrid	Oversampling	Batch	31723	100.0	99.3	98.6
		Random	5073	16.0	95.1	95.9
		Uncertainty	2264	7.1	96.9	96.8
	Undersampling	Batch	31723	100.0	99.3	98.6
		Random	4496	14.2	96.2	96.4
		Uncertainty	2182	**6.9**	**97.4**	97.2

TAL TECH

Results

Auto-labelling with threshold

- **Static threshold**
- Time dynamic threshold
- Iteration dynamic threshold

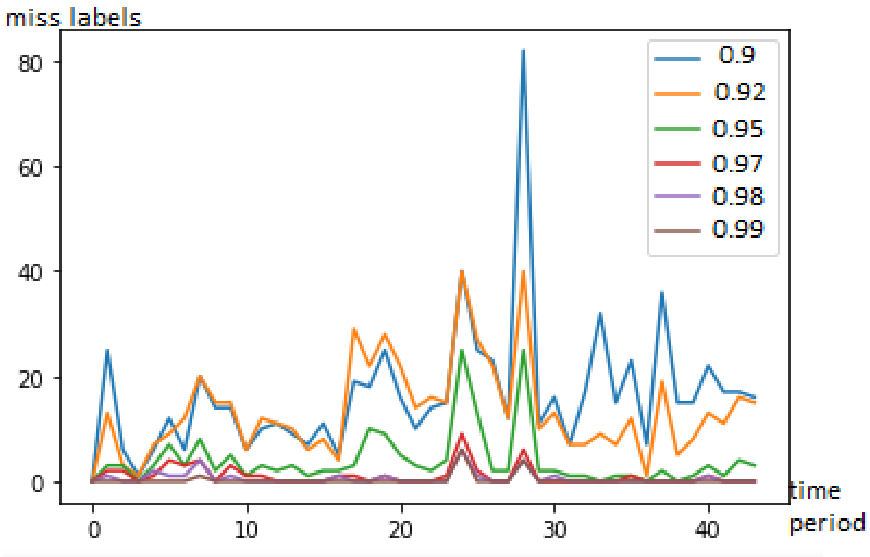


TABLE II: Training results for different static thresholds values (Hybrid, Undersampling)

Static Threshold	Label Numbers	Label Proportion(%)	F1(%)	Accuracy(%)	Auto-label Numbers	Miss-label Numbers
0.90	946	2.98	89.9	91.2	27550	972
0.92	1079	3.40	91.0	92.5	26392	698
0.95	1673	5.27	92.4	94.1	21825	419
0.97	1655	5.21	94.3	95.7	13790	55

TAL TECH

Results

Auto-labelling with threshold

- Static threshold
- **Time dynamic threshold**
- Iteration dynamic threshold

TABLE III: Training results for thresholds increasing or decreasing through time (Hybrid)

Balancing Method	Dynamic Threshold	Label Numbers	Label Proportion(%)	F1(%)	Accuracy(%)	Auto-label Numbers	Miss-label Numbers
Oversampling	Ascending	1439	4.53	92.0	93.3	27466	592
	Descending	1145	3.60	**94.3**	**95.1**	10306	**181**
Undersampling	Ascending	1405	4.60	92.6	94.0	24792	387
	Descending	927	3.00	90.9	94.0	18751	323

TAL TECH

Results

Auto-labelling with threshold

- Static threshold
- Time dynamic threshold
- **Iteration dynamic threshold**

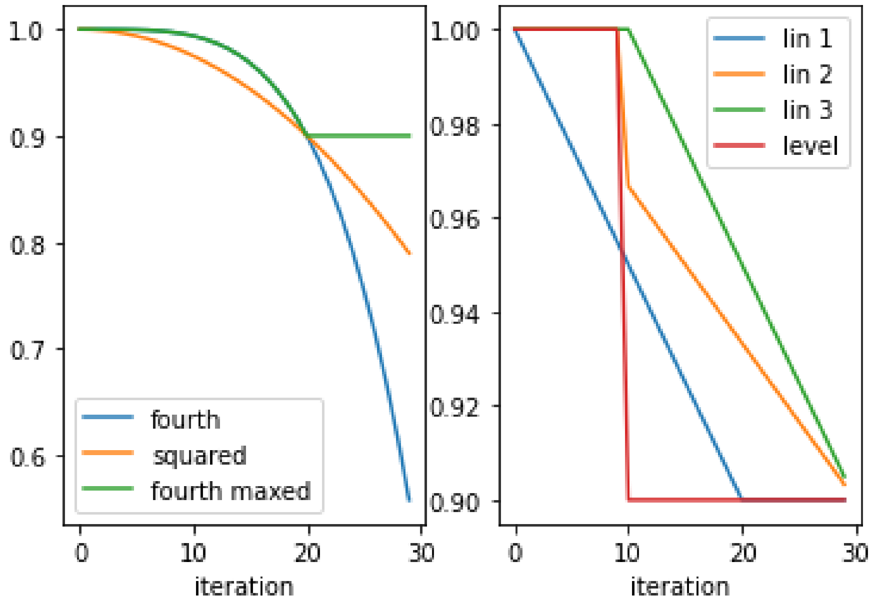


TABLE IV: Comparison of optimized shape results

Shapes	F1 (%)	Accuracy (%)	Label Numbers	Label Proportion(%)
no auto-labeling	98.30	97.40	1334	4.27
fourth	97.20	96.00	502	1.61
fourth maxed	97.70	96.70	718	2.30
squared	97.30	96.10	531	1.70
lin 1	97.50	96.30	711	2.28
lin 2	97.70	96.60	716	2.29
lin 3	97.70	96.60	576	1.85
level	**97.97**	96.90	662	2.12
T_desc	97.40	96.30	897	2.87

Results

Auto-labelling with threshold

- Static threshold
- Time dynamic threshold
- Iteration dynamic threshold

Auto-labelling driven by data drift detection

It achieves a 97.9% F1 score using only 2.37% of the labels

TABLE V: Auto-labeling driven by drift detection (*level*)

Drift Threshold	F1 (%)	Accuracy (%)	Label	
			Numbers	Proportion (%)
0.20	**97.8**	96.7	699	**2.24**
0.25	**97.9**	**96.9**	741	**2.37**
0.30	**97.7**	96.6	734	**2.35**
0.50	**97.8**	96.7	691	**2.21**

TAL TECH

Conclusion

- Active learning in mobile malware detection reduces labeling costs while improving model performance.
- The approach prioritizes acquiring highly informative data at minimal cost.
- Auto-labeling high-confidence data points expands the training set and adapts to non-stationary environments, maintaining performance over time.
- Careful management of thresholds is essential to prevent mislabeling and optimize model cost-effectiveness.
- Future improvements include enhancing data drift prediction through continuously monitoring the statistical properties and incorporating adaptive algorithms for timely model updates.

THANK YOU!

zhe.deng@taltech.ee
arthur.hubert@uni.lu
sadok.ben@taltech.ee
hayretdin.bahsi@taltech.ee

Co-funded by
the European Union

Investing
in your future

TALLINN UNIVERSITY
OF TECHNOLOGY