Malware Detection under Concept Drift: Science and Engineering

Marcus Botacin[1]

[1]Texas A&M University (TAMU), USA
botacin@tamu.edu
@MarcusBotacin

Whoami

Education

- Assistant Professor @ TAMU (Since 2022)
- CS PhD @ UFPR, Brazil (2021)
- CSE/ECE BSc. + CS MSC @ UNICAMP, Brazil (2015, 2017)

Research

- **Malware** at high-level: ML-based detectors.
- **Malware** at mid-level: Sandboxes and tracers.
- **Malware** at low-level: HW-based detectors.

Current Project

- NSF SaTC: Hardware Performance Counters as the next-gen AVs.

Publication

SPRINGER NATURE Link

Find a journal | Publish with us | Track your research | Search

Home > Detection of Intrusions and Malware, and Vulnerability Assessment > Conference paper

Towards Explainable Drift Detection and Early Retrain in ML-Based Malware Detection Pipelines

Conference paper | First Online: 10 July 2025

pp 3–24 | Cite this conference paper

Figure: **Source:** `https://link.springer.com/chapter/10.1007/978-3-031-97623-0_1`

Agenda

Agenda

1. Concept Drift
   - It is a long-term trend
   - The trend is in the data
2. Understanding Drift
   - Classes are affected differently
   - Our theory of drift events
3. Testing the Hypothesis
   - How to monitor drift events
   - The results do make sense
4. Demonstration
   - Explaining events by examples
   - Generalization
5. Real-World Considerations
   - Performance Drawbacks
   - Labeling Issues
6. Engineering Solutions
7. Conclusions
   - Closing Remarks

It is a long-term trend

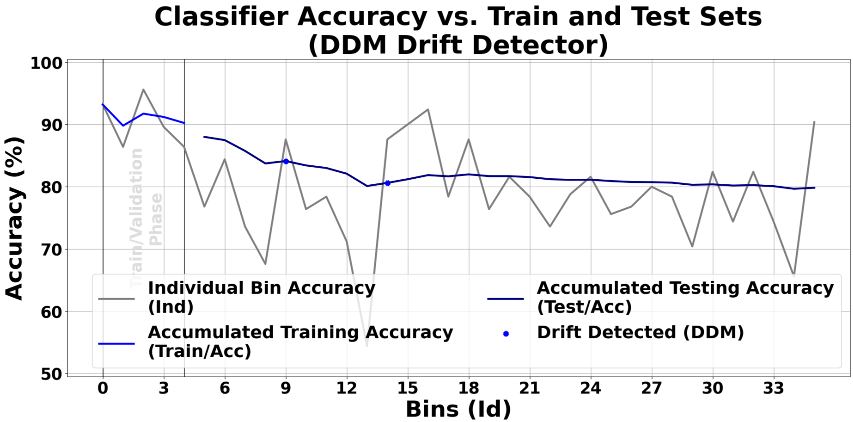


Figure: **Drift tendency vs. instantaneous detection.** Drift points reported by ADWIN.

Agenda

You increase the training size, but the drift is always there!

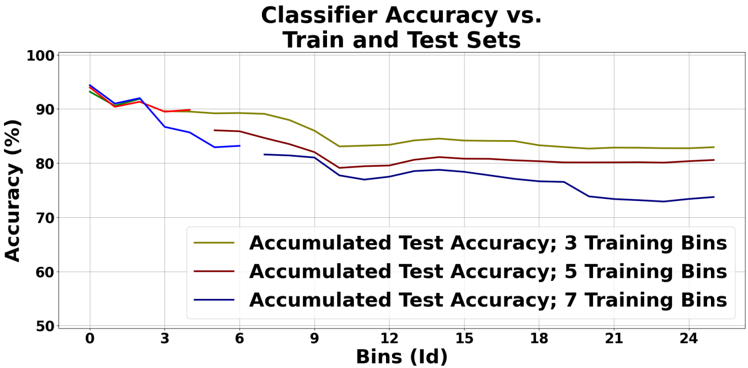


Figure: **Concept drift in practice.** The classification accuracy decreases regardless of the initial training set size/period.

You change the policy, but the drift is always there!

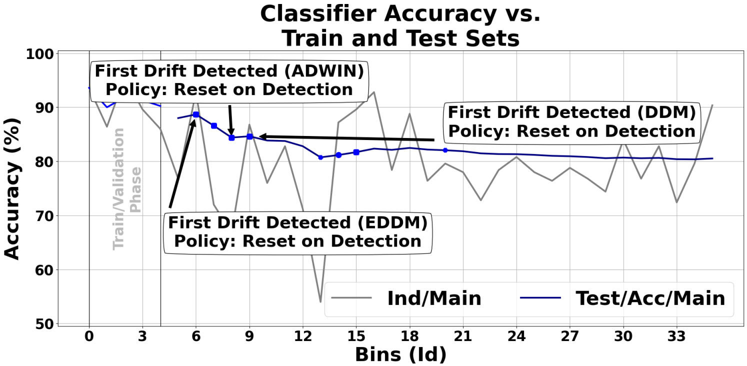


Figure: **Comparing algorithms and policies.** Each one detects a different number of drift points/events and at different times.

Agenda

Classes are affected differently

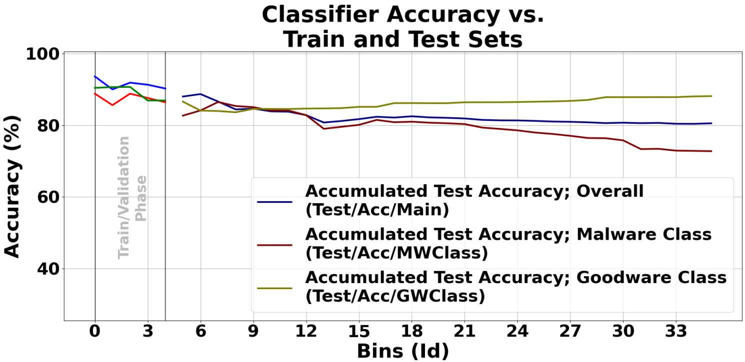


Figure: **Separating detection rates per class** reveals that the drift in the MW class causes the global performance degradation.

Classes drift differently

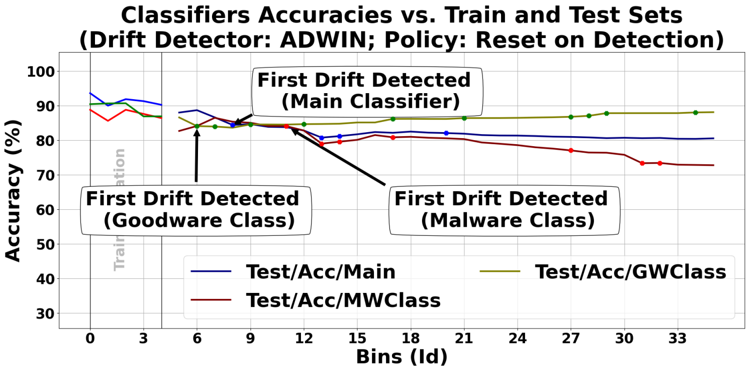


Figure: **Main class vs. sub-classes.** A different number of drift points is identified in each class and at different epochs.

Agenda

Key Hypothesis: Concepts and Frontiers are different things

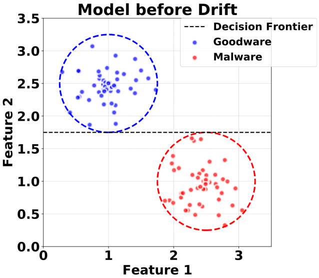


Figure: **Initial Training.**

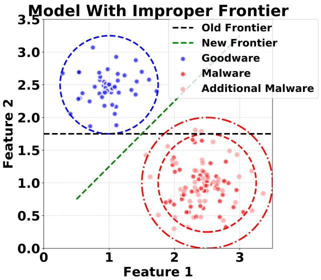


Figure: **Additional Data.**

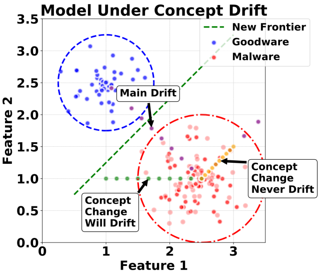


Figure: **Multiple Drifts.**

Proposing a drift taxonomy (1/2)

- **Type 1:** Main Classifier Drift. It detects whether a significant number of samples of any class crossed the detection frontier or not within a sampling window to the point of already harming the final classification result.
- **Type 2:** Sub-Class Drift. It detects whether a significant number of samples of a specific class crossed the detection frontier or not within a sampling window to the point of being noticeable but without guarantees that it affects the final classification result (contingent upon Type 1 detection).
- **Type 3:** Concept Change. It detects if a significant number of samples of a specific class do not match the previous knowledge the classifier had about that class, regardless of the correct class assignment (Type 1 and 2 events).

The concept evolution direction matters

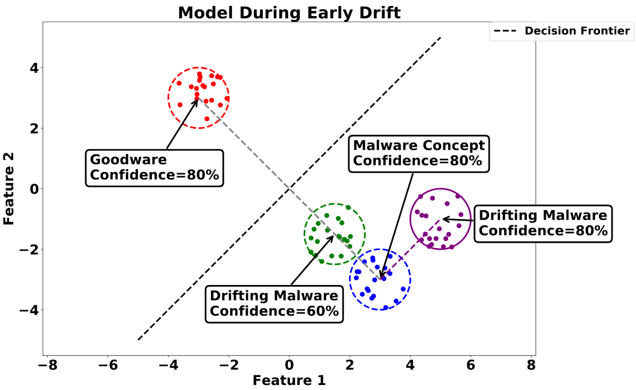


Figure: **Direction-Change Drift Detection.**

Proposing a drift taxonomy (2/2)

- **Type 3:** Concept Change. It detects if a significant number of samples of a specific class do not match the previous knowledge the classifier had about that class, regardless of the correct class assignment (Type 1 and 2 events). The implications of the concept change causing drift or not are contingent on the following cases:
  - **Case A:** Concept change without drift risk. If the concept changes in a direction that does not go toward the decision frontier, it cannot cause drift events.
  - **Case B:** Concept change with imminent drift risk. If the concept changes towards the decision frontier (Type 2), it will eventually cause drift when crossing the frontier (Type 1). This point is a candidate for early retraining.
  - **Case C:** Current Drift due to concept change. If the concept changes towards the frontier (Type 2) and crosses it (Type 1), concept drift is detected late.

The final drift taxonomy

Table: **Explaining Drift Events.** Information types for each combination of triggered detectors. Representing Triggered Detectors ($\checkmark$) and Possible ($\triangle$) and Not-Applicable ($\emptyset$) cases. Omitting Impossible cases.

Main Type			Cases			Conclusion
Type 1	**Type 2**	**Type 3**	**Case A**	**Case B**	**Case C**	
				$\emptyset$		Normal Operation
		$\checkmark$	$\triangle$			Early Concept Change with no impact on frontier
		$\checkmark$		$\triangle$		Early Concept Change with imminent impact on frontier
	$\checkmark$			$\emptyset$		Bad Frontier detected without concept change hold by imbalance in main class
	$\checkmark$	$\checkmark$		$\triangle$		Bad Frontier detected with concept change hold by imbalance in main class
$\checkmark$				$\emptyset$		False Positive Drift Detection
$\checkmark$		$\checkmark$	$\triangle$			False Positive with concept change in non-impactful direction
$\checkmark$	$\checkmark$			$\emptyset$		Bad Frontier detected without concept change, with impact in the main class
$\checkmark$	$\checkmark$	$\checkmark$			$\triangle$	Concept Change with Immediate Impact and Identification

Agenda

1. Concept Drift
   - It is a long-term trend
   - The trend is in the data
2. Understanding Drift
   - Classes are affected differently
   - Our theory of drift events
3. Testing the Hypothesis
   - How to monitor drift events
   - The results do make sense
4. Demonstration
   - Explaining events by examples
   - Generalization
5. Real-World Considerations
   - Performance Drawbacks
   - Labeling Issues
6. Engineering Solutions
7. Conclusions
   - Closing Remarks

Agnostic model monitor with external meta-models

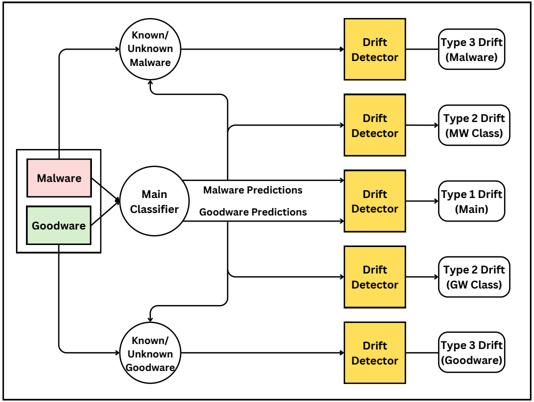


Figure: **Drift-Explainable Architecture.**

Agenda

The concept classes indeed measure different things

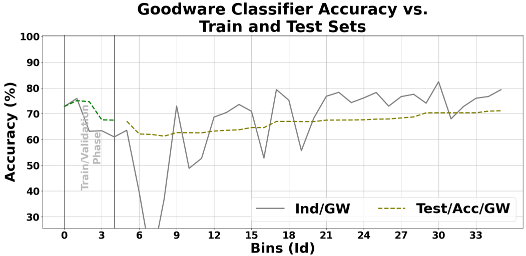


Figure: **GW class self-recognition rate.**

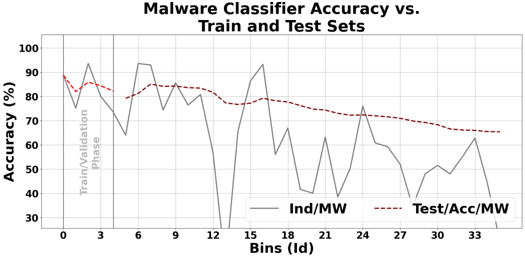


Figure: **MW class self-recognition rate.**

The concept classes indeed drift

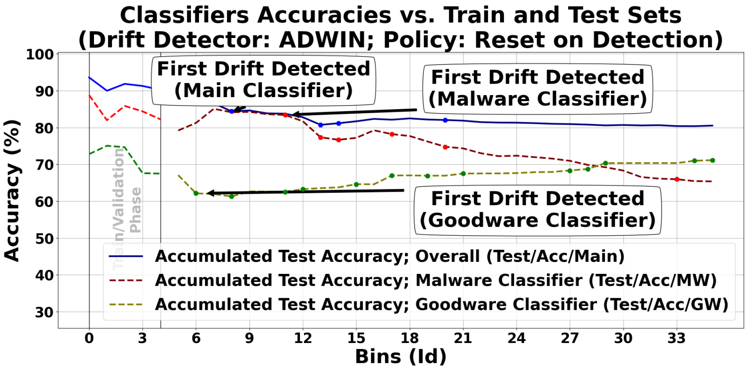


Figure: **Drift in the classes self-recognition rates.** Drifts are represented both for the MW and GW meta-classifiers.

Agenda

Every point can be explained.

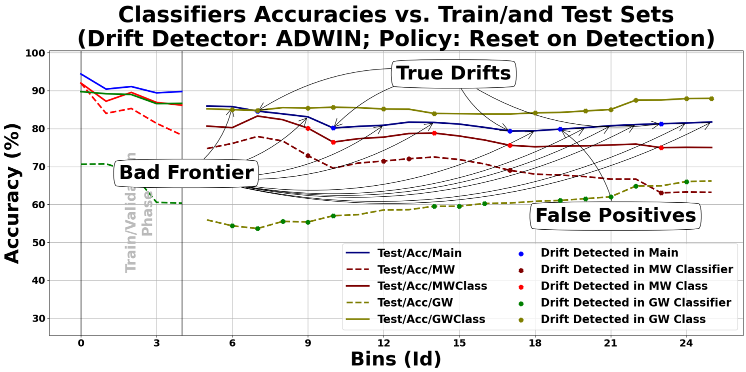


Figure: **Explaining all operational points and all drift occurrences.** Omitting points of normal operation.

The effect of dataset imbalance is also explained.

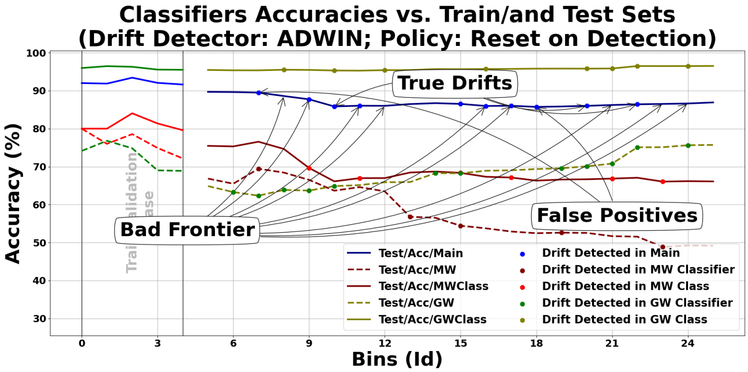


Figure: **Explaining all drift detection points (2:1 balance).** We observe fewer frontier problems and more False Positives.

Calibrating drift detectors is essential

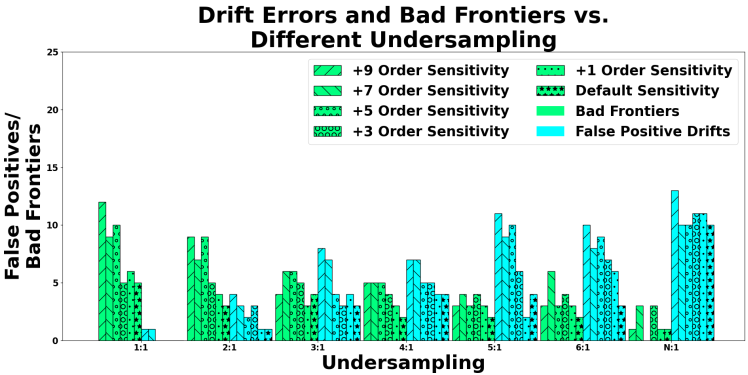


Figure: **Drift Detector Calibration.** False Positives and bad frontiers are explained by the proposed approach.

Explaining concepts lead to better accuracies via early retrain

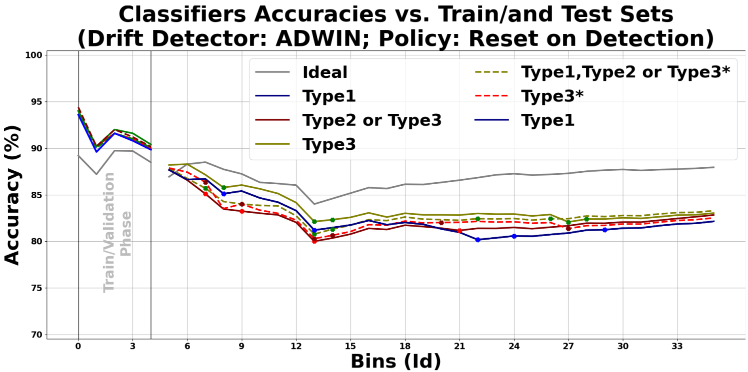


Figure: **Early retraining on concept changes** leads to improved accuracy than retraining only upon main class drift.

Early retrain is more effective

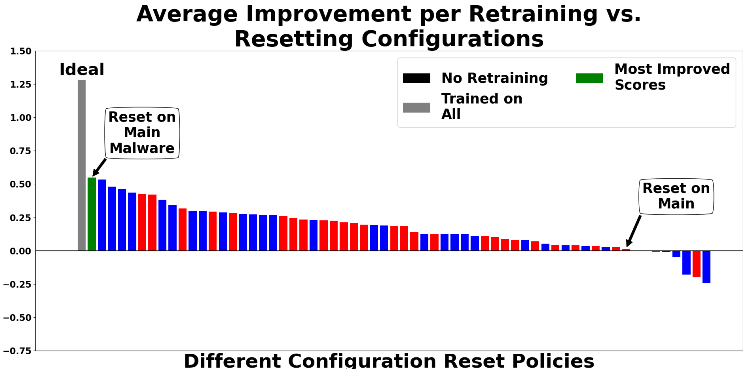


Figure: **Retrain Effectiveness.** The vast majority of the proposed drift detection triggers lead to increased accuracy gains than the original Type 1 drift detector trigger.

Early retrain is more efficient

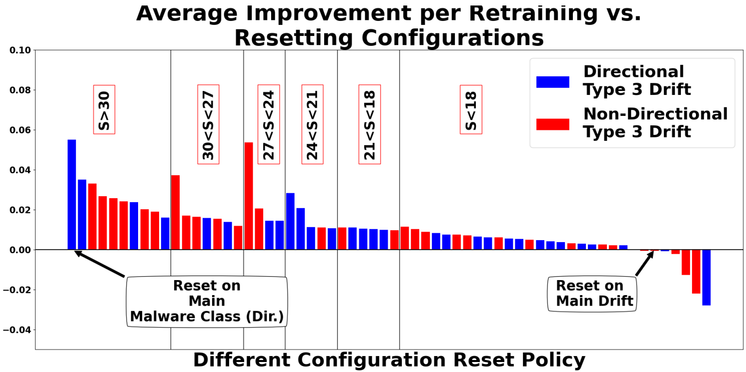


Figure: **Retrain Efficiency.** The best cost-benefit between the amount of retrains and accuracy increase is achieved by identifying concept changes.

Agenda

False Positives always increase with imbalances

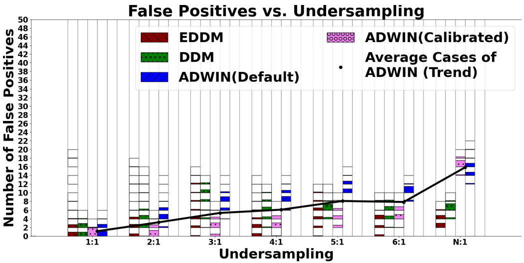


Figure: **DREBIN: FP Results Distribution for different imbalances.** FPs grow with the imbalance for most detectors.

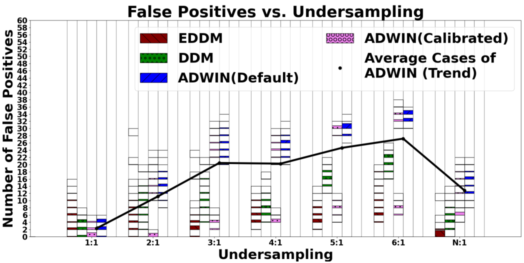


Figure: **ANDROZOO: FP Results Distribution for different imbalances.** FPs grow with the imbalance for most detectors.

Bad Frontiers always decrease with imbalances

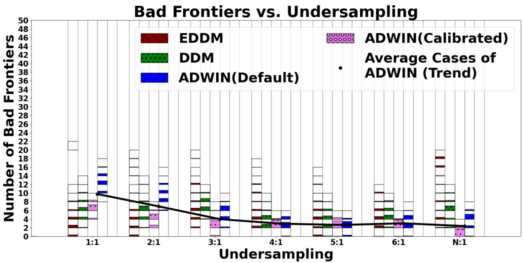


Figure: **DREBIN: Detectors' Results Distribution.** Frontier problems for different undersamplings. The bigger the undersampling, the more frontier problems.

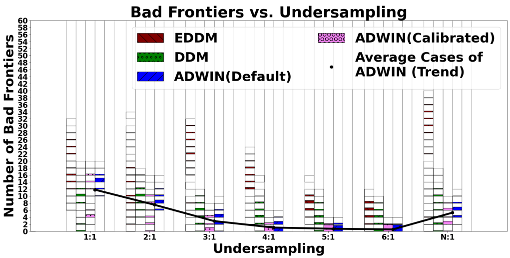


Figure: **ANDROZOO: Detectors' Results Distribution.** Frontier problems for different undersamplings. The bigger the undersampling, the more frontier problems.

Early retrain is always the best option

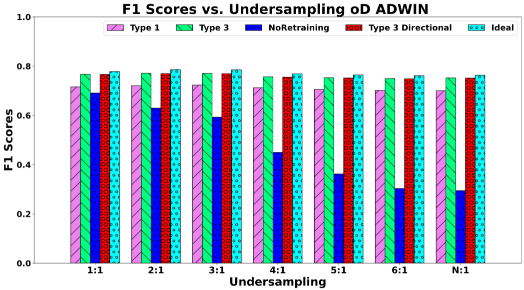


Figure: **DREBIN: Average Retraining Results.** Average F1-score Under Time increase over the baseline when triggering retrains using different policies vs. the multiple dataset imbalances.

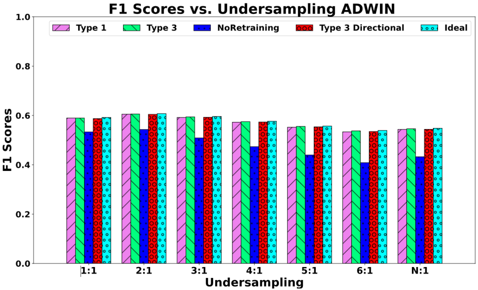


Figure: **ANDROZOO: Average Retraining Results.** The imbalance effect is less pronounced in this dataset, but the Type-3 retraining strategy is still the superior one in all scenarios.

Agenda

Explanation comes at the performance cost

Table: **Runtime Performance Overhead for DREBIN and AndroZoo.** The cost of individual retrains is reduced, but the total execution time cost increases.

		DREBIN		AndroZoo	
Models (#)	**Retrain Policy**	**Total Time / Retrains (#)**	**Cost / Overhead / Normalized**	**Total Time / Retrains (#)**	**Cost / Overhead / Normalized**
1	Type 1	11.65s / 5	2.73s / 0x / 0x	470.4s / 5	94s / 0x / 0x
3	Type 1	53.53s / 5	10.7s / 3.92x / 3.92x	1688.6 / 5	337.7s / 3.6x / 3.6x
3	Type 2	105.77s / 13	8.13s / 7.74x / 2.98x	3563.9s / 15	237.6s / 7.5x / 2.5x
3	Tyoe 3	102.70s / 13	7.90s / 7.52x / 2.89x	4161.3s / 19	219s / 8.84x / 2.32x

Agenda

Limitations & Future Works

- Heterogeneous Architectures.
- Virtual Drifts.
- Intra-Class Drifts.
- **Label Delays.**

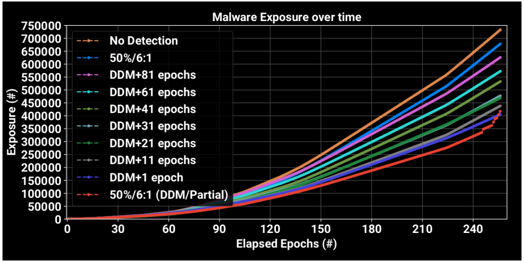
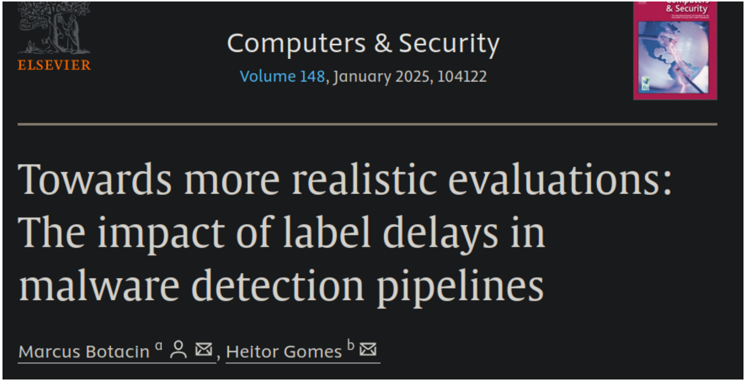


Figure: **Label Delays.** Too much delay nullifies the retraining benefits.

Publications


ELSEVIER

Computers & Security

Volume 148, January 2025, 104122

Towards more realistic evaluations: The impact of label delays in malware detection pipelines

Marcus Botacin [a], Heitor Gomes [b]

Figure: **Source:** https://www.sciencedirect.com/science/article/abs/pii/S0167404824004279

How to Evaluate the Classification Impact?

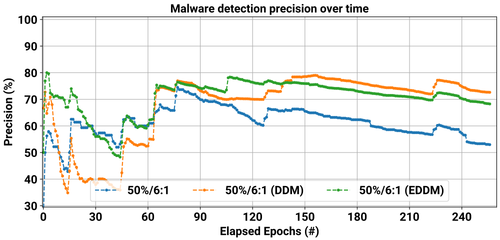


Figure: **Classification Precision.**

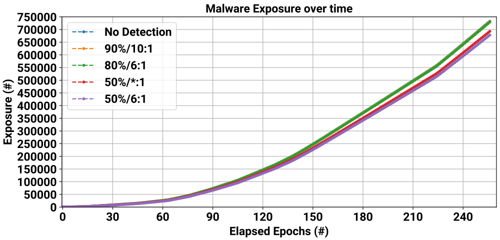


Figure: **Absolute Exposure.**

How to Evaluate the Drift Impact?

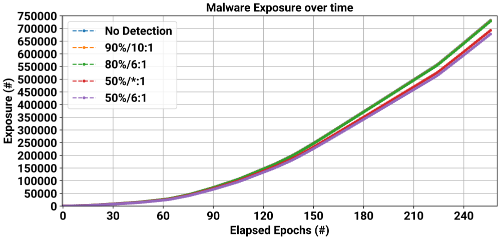


Figure: **Absolute Exposure.**

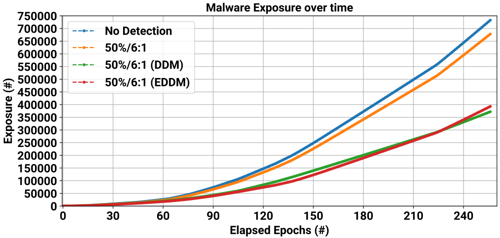


Figure: **Absolute Exposure and Drift.**

What if Labels are not available?

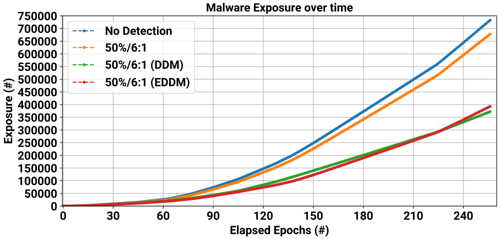


Figure: **Absolute Exposure and Drift.**

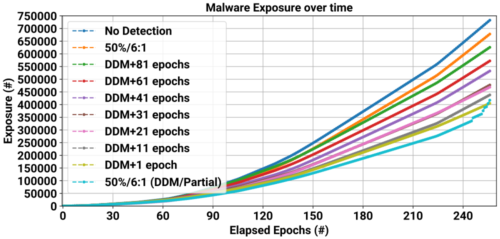


Figure: **Delayed ground-truth labels.**

The Pseudo-Label Delay Mitigation

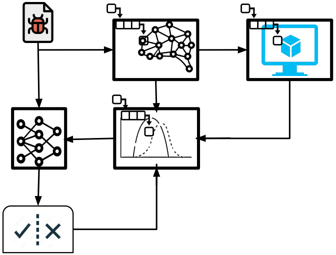


Figure: **Architecture with Pseudo-Labels.**

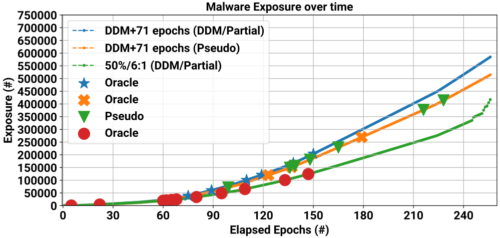


Figure: **New Drift Points.**

Agenda

Recap

Science

- Explaining drift is not the same as explaining the classification.
- Classifier concept and frontier are not the same thing.
- Meta-classifiers can separate concepts and frontiers.
- We can explain every drift event.

Engineering

- Labels are not immediately available.
- Too long delays eliminate the benefits of classifier retrain.
- Pseudo-Labels mitigate label delays.

Thanks!

Questions? Comments?

botacin@tamu.edu
@MarcusBotacin