Novelty Assessment Report

Paper: LAMDA: A Longitudinal Android Malware Benchmark for Concept Drift Analysis

PDF URL: https://openreview.net/pdf?id=1FnCrZtBNQ

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

Machine learning (ML)-based malware detection systems often fail to account for the dynamic nature of real-world training and test data distributions. In practice, these distributions evolve due to frequent changes in the Android ecosystem, adversarial development of new malware families, and the continuous emergence of both benign and malicious applications. Prior studies have shown that such concept drift—distributional shifts in benign and malicious samples, leads to significant degradation in detection performance over time. Despite the practical importance of this issue, existing datasets are often outdated and limited in temporal scope, diversity of malware families, and sample scale, making them insufficient for the systematic evaluation of concept drift in malware detection.

To address this gap, we present LAMDA, the largest and most temporally diverse Android malware benchmark to date, designed specifically for concept drift analysis. LAMDA spans 12 years (2013–2025, excluding 2015), includes over 1 million samples (approximately 37\% labeled as malware), and covers 1,380 malware families and 150,000 singleton samples, reflecting the natural distribution and evolution of real-world Android applications. We empirically demonstrate LAMDA's utility by quantifying the performance degradation of standard ML models over time and analyzing feature stability across years. As the most comprehensive Android malware dataset to date, LAMDA enables in-depth research into temporal drift, generalization, explainability, and evolving detection challenges.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Concept Drift in Android Malware Detection Over Time**

A total of **50 papers** were analyzed and organized into a taxonomy with **26 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Drift Detection and Characterization**
- **Drift Adaptation Strategies**
- **Robust Representation Learning**
- **Feature Engineering and Selection**
- **Specialized Detection Contexts**
- **Malware Evolution and Ecosystem Analysis**
- **Emerging Techniques and Future Directions**

Complete Taxonomy Tree

- Concept Drift in Android Malware Detection Over Time Survey Taxonomy
- Drift Detection and Characterization
    - Drift Detection Mechanisms (3 papers)
    - [9] Transcend: Detecting concept drift in malware classification models (Roberto Jordaney, 2017) View paper
    - [18] Towards Explainable Drift Detection and Early Retrain in ML-Based Malware Detection Pipelines (Jayesh Tripathi, 2025) View paper
    - [21] Active Learning-Based Mobile Malware Detection Utilizing Auto-Labeling and Data Drift Detection (Zhe Deng, 2024) View paper
    - Drift Cause Analysis (4 papers)
    - [8] Data drift in android malware detection (Luca Minnei, 2024) View paper
    - [16] Drift forensics of malware classifiers (T.T. Chow, 2023) View paper
    - [19] Understanding Concept Drift with Deprecated Permissions in Android Malware Detection (Jarrar, 2025) View paper
    - [28] Is it overkill? analyzing feature-space concept drift in malware detectors (Zhi Chen, 2023) View paper
    - Temporal Evaluation and Benchmarking ★ (5 papers)
    - [0] LAMDA: A Longitudinal Android Malware Benchmark for Concept Drift Analysis (Anon et al., 2026) View paper
    - [23] Empirical Evaluation of Concept Drift in ML-Based Android Malware Detection (Jarrar, 2025) View paper
    - [24] Revisiting Temporal Inconsistency and Feature Extraction for Android Malware Detection (Maryam Tanha, 2024) View paper
    - [25] Breaking Out from the TESSERACT: Reassessing ML-based Malware Detection under Spatio-Temporal Drift (Chow, 2025) View paper
    - [27] Aurora: Are Android Malware Classifiers Reliable under Distribution Shift? (Herzog, 2025) View paper
- Drift Adaptation Strategies
    - Active Learning-Based Adaptation (1 papers)
    - [3] Experts still needed: boosting long-term android malware detection with active learning (Alejandro Guerra-Manzanares, 2024) View paper
    - Self-Training and Pseudo-Labeling (2 papers)

  - [15] MORPH: Towards Automated Concept Drift Adaptation for Malware Detection (Alam, 2024) View paper
  - [35] ADAPT: A Pseudo-labeling Approach to Combat Concept Drift in Malware Detection (Alam, 2025) View paper
  - Continuous and Incremental Learning (3 papers)
  - [4] Continuous Learning for Android Malware Detection (Chen Yizheng, 2023) View paper
  - [6] Fast & furious: On the modelling of malware detection as an evolving data stream (F Ceschin, 2023) View paper
  - [7] Temporal-Incremental Learning for Android Malware Detection (Tiezhu Sun, 2024) View paper
  - Retraining and Model Update Strategies (2 papers)
  - [22] Efficient Concept Drift Handling for Batch Android Malware Detection Models (Borja Molina-Coronado, 2023) View paper
  - [43] A pragmatic android malware detection procedure (Paolo Palumbo, 2017) View paper
  - Self-Evolving and Autonomous Systems (1 papers)
  - [1] Droidevolver: Self-evolving android malware detection system (Ke Xu, 2019) View paper
  - Reinforcement Learning-Based Adaptation (1 papers)
  - [45] DRMD: Deep Reinforcement Learning for Malware Detection under Concept Drift (Foley, 2025) View paper
- Robust Representation Learning
  - Invariant Feature Learning (2 papers)
  - [5] Learning Temporal Invariance in Android Malware Detectors (ZHENG Xinran, 2025) View paper
  - [39] Robust Machine Learning for Malware Detection over Time (Angioni, 2022) View paper
  - Domain Adaptation Approaches (1 papers)
  - [14] Domain Adaptation-Based Deep Learning Framework for Android Malware Detection Across Diverse Distributions (Shuguang Xiong, 2024) View paper
  - Semantic and Graph-Based Representations (3 papers)
  - [13] Combating Concept Drift with Explanatory Detection and Adaptation for Android Malware Classification (He Yiling, 2024) View paper
  - [31] Mitigating Distribution Shift in Graph-Based Android Malware Classification via Function Metadata and LLM Embeddings (Tran, 2025) View paper
  - [38] Heterogeneous temporal graph transformer: An intelligent system for evolving android malware detection (Yujie Fan, 2021) View paper
  - Uncertainty-Aware and Confidence-Based Methods (1 papers)
  - [11] Malcertain: Enhancing Deep Neural Network Based Android Malware Detection by Tackling Prediction Uncertainty (Haodong Li, 2024) View paper
- Feature Engineering and Selection
  - Static Feature Analysis (1 papers)
  - [37] Leveraging the first line of defense: a study on the evolution and usage of android security permissions for enhanced android malware detection (Alejandro Guerra-Manzanares, 2023) View paper
  - Dynamic and Hybrid Feature Extraction (2 papers)
  - [29] Advancing Android Malware Detection: A Unified Framework with Dynamic Feature Extraction and Privacy-Preserving Collaboration (Manadadi Sriya Reddy, 2024) View paper
  - [48] Android malware concept drift using system calls: Detection, characterization and challenges (Alejandro Guerra-Manzanares, 2022) View paper
  - Feature Stability and Evolution Analysis (1 papers)
  - [36] Assessing and improving malware detection sustainability through app evolution studies (Haipeng Cai, 2020) View paper
- Specialized Detection Contexts
  - Resource-Constrained and Multilevel Detection (1 papers)
  - [10] Hybrid multilevel detection of mobile devices malware under concept drift (Andrea Augello, 2025) View paper
  - Privacy-Preserving and Federated Learning (2 papers)
  - [41] Enabling Privacy-Preserving Cyber Threat Detection with Federated Learning (Bi Yu, 2024) View paper
  - [47] Comprehensive Android Malware Detection Based on Federated Learning Architecture (Wenbo Fang, 2023) View paper
  - Cross-Device and Cross-Distribution Challenges (2 papers)
  - [33] Corrigendum to Concept drift and cross-device behavior: Challenges and implications for effective android malware detection Computers & Security, Volume 120, 102757 (Alejandro Guerra-Manzanares, 2023) View paper
  - [44] Concept drift and cross-device behavior: Challenges and implications for effective android malware detection (Alejandro Guerra-Manzanares, 2022) View paper
  - Specialized Threat Detection (3 papers)
  - [26] BPFDex: Enabling Robust Android Apps Unpacking via Android Kernel (Mingyang Li, 2025) View paper
  - [49] TSDroid: A novel android malware detection framework based on temporal & spatial metrics in IoMT (Gaofeng Zhang, 2023) View paper
  - [50] Android ransomware detection based on a hybrid evolutionary approach in the context of highly imbalanced data (Iman Almomani, 2021) View paper
- Malware Evolution and Ecosystem Analysis
  - Malware Family Evolution and Characterization (2 papers)
  - [17] Demystifying the evolution of Android malware variants (Lihong Tang, 2023) View paper
  - [34] Addressing malware family concept drift with triplet autoencoder (Numan Halit Guldemir, 2025) View paper
  - Ecosystem-Wide Trend Analysis (2 papers)
  - [30] Dissecting android malware: Characterization and evolution (Yajin Zhou, 2012) View paper
  - [46] Eight years of rider measurement in the android malware ecosystem: evolution and lessons learned (Suarez-Tangil, 2018) View paper
- Emerging Techniques and Future Directions
  - Large Language Model-Based Detection (1 papers)
  - [40] Poster: LLMalware: An LLM-Powered Robust and Efficient Android Malware Detection Framework (Zijing Ma, 2025) View paper
  - Predictive and Proactive Approaches (1 papers)
  - [20] Counteracting Concept Drift by Learning with Future Malware Predictions (Bosansky, 2024) View paper
  - Cross-Domain Applications and Methodological Transfers (2 papers)
  - [2] Cluster analysis and concept drift detection in malware: A. Mishra, M. Stamp (A Mishra, 2025) View paper

- [12] Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic â¦ (S Malik, 2025) View paper
- Novel Detection Paradigms (2 papers)
- [32] Novel nature-inspired optimization approach-based SVM for identifying the android malicious data (Bhawani Sankar Panigrahi, 2024) View paper
- [42] Android malware detection using time-aware machine learning approach (Anas Alsobeh, 2024) View paper

Narrative

Core task: concept drift in Android malware detection over time. The field addresses how malware evolves continuously, causing trained detectors to degrade as new attack patterns emerge and old features become obsolete. The taxonomy organizes research into seven main branches. Drift Detection and Characterization focuses on identifying when and how distribution shifts occur, often through temporal evaluation frameworks and empirical benchmarking studies. Drift Adaptation Strategies encompasses incremental learning, active learning, and domain adaptation methods that update models without full retraining. Robust Representation Learning seeks feature encodings that remain stable across time, while Feature Engineering and Selection refines input signals to minimize sensitivity to evolving malware tactics. Specialized Detection Contexts examines drift in particular settings such as privacy-preserving or federated environments. Malware Evolution and Ecosystem Analysis investigates the underlying causes of drift by studying how malware families, permissions, and app ecosystems change. Emerging Techniques and Future Directions explores novel paradigms including large language models and graph-based embeddings that may offer new resilience against temporal shifts.

Several active lines of work reveal key trade-offs between detection accuracy, adaptation speed, and computational overhead. Continuous learning approaches like Continuous Learning Android[4] and Temporal Incremental Learning[7] enable models to absorb new samples incrementally, yet they must balance plasticity against catastrophic forgetting. In contrast, works emphasizing temporal invariance such as Temporal Invariance Android[5] and TESSERACT[25] aim to learn representations that generalize across time windows, reducing the need for frequent updates but potentially sacrificing responsiveness to abrupt shifts. LAMDA[0] sits within the Temporal Evaluation and Benchmarking cluster, alongside Empirical Drift Evaluation[23] and Temporal Inconsistency Revisited[24], providing rigorous experimental protocols to measure drift effects. Compared to Aurora[27], which also emphasizes systematic temporal assessment, LAMDA[0] offers a complementary perspective on how to structure longitudinal experiments and interpret performance degradation patterns, helping researchers understand whether observed drift stems from feature obsolescence, label noise, or adversarial evolution.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Empirical Evaluation of Concept Drift in ML-Based Android Malware Detection

Authors: Jarrar, Radi, Ahmed Sabbah, Radi Jarrar, Mohaisen, et al. (8 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Despite outstanding results, machine learning-based Android malware detection models struggle with concept drift, where rapidly evolving malware characteristics degrade model effectiveness. This study examines the impact of concept drift on Android malware detection, evaluating two datasets and nine machine learning and deep learning algorithms, as well as Large Language Models (LLMs). Various feature types--static, dynamic, hybrid, semantic, and image-based--were considered. The results showed ...

Relationship Analysis

Both papers belong to the Temporal Evaluation and Benchmarking category, focusing on empirical evaluation of model degradation under concept drift in Android malware detection. They overlap in their core objective of quantifying performance decline over time using temporal splits and multiple ML models. However, the original paper (LAMDA) primarily contributes a large-scale longitudinal benchmark dataset spanning 12 years with 1M+ samples and detailed drift characterization, while the candidate paper focuses on a comparative empirical study across multiple datasets, feature types (static, dynamic, hybrid, semantic, image-based), and detection methods including LLMs, without introducing a new benchmark.

2. Revisiting Temporal Inconsistency and Feature Extraction for Android Malware Detection

Authors: Maryam Tanha, Arunab Singh, Gavin Knoke | **Year/Venue**: 2024 | **URL**: View paper

Abstract

The growing popularity and ease of access have turned Android applications into prime targets for malicious attackers. Within the security research community, machine learning has become an essential instrument for conducting Android malware detection and analysis. However, there are potential threats to validity of existing studies, mainly resulting from their used datasets. One of the primary issues is temporal inconsistency (also called temporal bias) that is caused by incorrect time splits o...

Relationship Analysis

Both papers belong to the Temporal Evaluation and Benchmarking category, focusing on datasets and empirical studies for evaluating Android malware detection under temporal constraints. They overlap in addressing temporal inconsistency issues and creating temporally-organized datasets from AndroZoo for drift analysis. However, the original paper (LAMDA) presents a large-scale longitudinal benchmark spanning 12 years with over 1 million samples for comprehensive concept drift analysis, while the candidate paper focuses specifically on addressing temporal bias through precise release time indicators (Google Play Store upload year) and proposes a filtering methodology to improve dataset temporal consistency.

3. Breaking Out from the TESSERACT: Reassessing ML-based Malware Detection under Spatio-Temporal Drift

Authors: Chow, Theo, Linhardt, Lorenz, Arp, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Several recent works focused on the best practices for applying machine learning to cybersecurity. In the context of malware, TESSERACT highlighted the impact of concept drift on detection performance and suggested temporal and spatial constraints to be enforced to ensure realistic time-aware evaluations, which have been adopted by the community. In this paper, we demonstrate striking discrepancies in the performance of learning-based malware detection across the same time frame when evaluated o...

Relationship Analysis

Both papers belong to the Temporal Evaluation and Benchmarking category, focusing on datasets and empirical studies for evaluating Android malware detection under temporal drift. They overlap in addressing concept drift through longitudinal dataset construction, temporal evaluation protocols, and performance degradation analysis over time. However, LAMDA introduces a new 12-year benchmark (2013-2025) with over 1 million samples emphasizing dataset scale and feature stability analysis, while the candidate paper critically examines existing evaluation methodologies (TESSERACT constraints) and identifies five novel spatio-temporal bias factors (timestamp types, temporal luck, app markets, VT thresholds, dataset size) that affect realistic evaluations across existing datasets like APIGraph and Transcendent.

4. Aurora: Are Android Malware Classifiers Reliable under Distribution Shift?

Authors: Herzog, Alexander, Alexander Herzog, Cavallaro, Lorenzo, et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The performance figures of modern drift-adaptive malware classifiers appear promising, but does this translate to genuine operational reliability? The standard evaluation paradigm primarily focuses on baseline performance metrics, neglecting confidence-error alignment and operational stability. While TESSERACT established the importance of temporal evaluation, we take a complementary direction by investigating whether malware classifiers maintain reliable and stable confidence estimates under di...

Relationship Analysis

Both papers belong to the Temporal Evaluation and Benchmarking category, focusing on evaluating Android malware detection systems under realistic temporal constraints and concept drift. They overlap in their emphasis on longitudinal evaluation, temporal dataset construction, and measuring model degradation over time. However, LAMDA focuses on creating a large-scale benchmark dataset spanning 12 years (2013-2025) with over 1 million samples to enable drift analysis, while Aurora focuses on evaluating the reliability and stability of existing classifiers' confidence estimates under distribution shift using metrics like AURC, rather than dataset construction.

Contributions Analysis

Overall novelty summary. The paper introduces LAMDA, a large-scale longitudinal Android malware benchmark spanning 12 years and over 1 million samples, designed to facilitate systematic concept drift analysis. Within the taxonomy, it resides in the 'Temporal Evaluation and Benchmarking' leaf under 'Drift Detection and Characterization'. This leaf contains five papers total, indicating a moderately populated research direction. The sibling papers—Empirical Drift Evaluation, Temporal Inconsistency Revisited, Aurora, and one other—similarly focus on measuring model degradation over time, suggesting that temporal benchmarking is an established but not overcrowded subfield within concept drift research.

The taxonomy reveals that LAMDA's leaf sits within a broader branch dedicated to drift detection and characterization, which also includes 'Drift Detection Mechanisms' and 'Drift Cause Analysis'. Neighboring branches address adaptation strategies (active learning, incremental learning, retraining) and robust representation learning (invariant features, domain adaptation). The scope note for LAMDA's leaf explicitly excludes adaptation methods, clarifying that its contribution lies in providing evaluation infrastructure rather than proposing new model update techniques. This positioning suggests the work complements rather than competes with adaptation-focused research, offering a shared resource for testing drift mitigation approaches.

Among the three contributions analyzed, the dataset itself (Contribution A) examined 10 candidates with zero refutable prior work, suggesting strong novelty in scale and temporal scope. However, the empirical demonstration of concept drift (Contribution B) examined 10 candidates and found 6 refutable matches, indicating that performance degradation under temporal shifts is well-documented in prior studies. The multi-faceted analysis framework (Contribution C) examined only 1 candidate with no refutations, though the limited search scope makes it difficult to assess novelty conclusively. Overall, the dataset contribution appears more distinctive than the empirical findings, which align with established observations in the field.

Based on the limited search of 21 candidates, LAMDA's primary novelty lies in its dataset scale and temporal coverage rather than in demonstrating drift effects, which prior work has extensively characterized. The analysis does not cover exhaustive literature beyond top-K semantic matches, so additional related benchmarks or longitudinal studies may exist outside this scope. The contribution's value likely centers on enabling more rigorous comparative evaluations rather than introducing fundamentally new insights about concept drift mechanisms.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: LAMDA: A large-scale longitudinal Android malware benchmark dataset

Description: The authors introduce LAMDA, a comprehensive Android malware dataset spanning 12 years (2013–2025, excluding 2015) with over 1 million samples covering 1,380 malware families and 150,000 singleton samples. The dataset is specifically structured to enable systematic evaluation of concept drift in malware detection systems.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. On the relativity of time: Implications and challenges of data drift on long-term effective android malware detection

URL: View paper

Brief Assessment

The candidate paper (Relativity of Time[53]) focuses on data drift implications in Android malware detection but does not present a competing longitudinal benchmark dataset. Without access to the candidate's full text, no evidence of prior work that would refute LAMDA's novelty can be established.

2. Assessing and improving malware detection sustainability through app evolution studies

URL: View paper

Brief Assessment

App Evolution Studies[36] focuses on sustainability of malware detectors over time using 8-year longitudinal data (13,627 benign, 12,755 malware), but does not present a comprehensive benchmark dataset. LAMDA's scale (1M+ samples, 1,380 families, 12 years) and explicit design for concept drift analysis represent a distinct contribution.

3. LongCGDroid: Android malware detection through longitudinal study for machine learning and deep learning

URL: View paper

Brief Assessment

LongCGDroid[54] uses a 200k sample dataset spanning 2013-2022, which is smaller in scale and temporal scope than LAMDA's 1 million samples across 2013-2025. The candidate focuses on image-based call graph representations rather than static feature-based concept drift analysis, representing a different methodological approach to longitudinal malware detection.

4. FL-MalDrift: a federated learning framework for malware detection under local concept drift

URL: View paper

Brief Assessment

FL-MalDrift[52] focuses on federated learning for malware detection under concept drift, not on creating longitudinal benchmark datasets. The candidate uses existing datasets (Drebin, CICMalDroid 2020) for evaluation rather than introducing new benchmark data.

5. Learning Temporal Invariance in Android Malware Detectors

URL: View paper

Brief Assessment

Temporal Invariance Android[5] does not present a benchmark dataset. It proposes TIF, a training framework for temporal invariance, and evaluates on 'a decade-long dataset' without claiming to introduce that dataset as a contribution.

6. Temporal-Incremental Learning for Android Malware Detection

URL: View paper

Brief Assessment

Temporal Incremental Learning[7] uses the MalNet dataset organized chronologically for temporal-incremental learning, not for systematic concept drift analysis. The focus is on class-incremental learning with temporal shifts rather than comprehensive drift benchmarking.

7. One step forward, two steps back: ML-based malware detection under concept drift

URL: View paper

Brief Assessment

One Step Forward[51] focuses on evaluating retraining strategies for concept drift in malware detection, not on creating a large-scale longitudinal benchmark dataset. The candidate's scope is methodological evaluation rather than dataset construction.

8. Continuous Learning for Android Malware Detection

URL: View paper

Brief Assessment

Continuous Learning Android[4] focuses on active learning methods for continuous malware detection rather than dataset construction. While it uses existing datasets (apigraph, androzoo), it does not claim to introduce a large-scale longitudinal benchmark dataset for concept drift analysis.

9. Experts still needed: boosting long-term android malware detection with active learning

URL: View paper

Brief Assessment

Active Learning Android[3] focuses on active learning strategies for maintaining non-stationary malware detection models over a 7-year period, not on creating a large-scale longitudinal benchmark dataset with comprehensive family labels and structured temporal splits for concept drift analysis.

10. DRMD: Deep Reinforcement Learning for Malware Detection under Concept Drift

URL: View paper

Brief Assessment

DRMD[45] focuses on deep reinforcement learning methods for malware detection under concept drift, not on creating longitudinal benchmark datasets. The candidate uses existing datasets (Transcendent and Hypercube) for evaluation rather than introducing new benchmark datasets for concept drift analysis.

Contribution 2: Empirical demonstration of concept drift and performance degradation

Description: The authors conduct comprehensive empirical evaluations showing how machine learning-based malware detectors degrade over time due to concept drift. They analyze performance degradation patterns, feature stability, and temporal shifts using multiple evaluation methodologies including supervised learning experiments and distributional analysis.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Hybrid multilevel detection of mobile devices malware under concept drift

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

2. Transcending transcend: Revisiting malware classification in the presence of concept drift

URL: View paper

Prior Art Analysis

Transcending Transcend[56] demonstrates empirical evaluation of concept drift and performance degradation in malware detection systems prior to the ORIGINAL paper. The candidate paper presents comprehensive experiments showing how machine learning-based malware classifiers degrade over time due to concept drift, including performance degradation patterns, feature stability analysis, and temporal shifts. The candidate evaluates multiple classifiers on a dataset spanning 5 years (2014-2019), showing rapid performance decay where classifiers become worse than random in under one year, and demonstrates that concept drift causes the incoming test distribution to diverge from the original training distribution.

Evidence

Evidence 1 - **Rationale**: Both papers describe the same fundamental problem: ML-based malware detection systems degrade over time due to concept drift caused by evolving malware and changing distributions. The candidate paper establishes this phenomenon prior to the original work. - **Original**: machine learning (ml)-based malware detection systems often fail to account for the dynamic nature of real-world training and test data distributions. in practice, these distributions evolve due to frequent changes in the android ecosystem, adversarial development of new malware families, and the co... - **Candidate**: machine learning (ml) algorithms have displayed superhuman performance across a wide range of classification tasks such as computer vision [23] and natural language processing [17]. however, a great deal of this success is conditional on one central assumption: that the training and test data are dr...

Evidence 2 - **Rationale**: Both papers empirically demonstrate performance degradation of ML models over time. The candidate paper shows classifiers becoming worse than random in under one year, demonstrating the same type of temporal performance analysis claimed as novel in the original paper. - **Original**: we empirically demonstrate lamda's utility by quantifying performance degradation of standard ml models over time and analyzing feature stability across years. - **Candidate**: figure 7 shows the the f1, precision, and recall (rows 1-3) for each of the novel evaluators (columns). the middle dashed line shows the baseline performance when no rejection is enforced. this is the performance decay caused by concept drift present in the dataset that results from an evolving mali...

Evidence 3 - **Rationale**: Both papers analyze feature distribution shifts over time. The candidate paper uses KL divergence to quantify distributional shifts and demonstrates temporal changes in feature distributions, which is part of the comprehensive drift analysis claimed in the original paper. - **Original**: we conduct comprehensive drift analysis, including per-feature distribution shifts, feature stability analysis across malware families (zhang et al., 2020), temporal label flipping analysis, and shap-based explanation (lundberg & lee, 2017) drift that reveal temporal changes in feature importance. - **Candidate**: figure 6 shows a visibly significant covariate shift in the distribution of features for training and test malware examples from jordaney et al. [20], with a kullback-leibler (kl) divergence [24]-an unbounded measure of distribution difference-of 696.66. the covariate shift in our dataset is much mo...

3. Breaking Out from the TESSERACT: Reassessing ML-based Malware Detection under Spatio-Temporal Drift

URL: View paper

Prior Art Analysis

TESSERACT[25] demonstrates that machine learning-based malware detectors experience significant performance degradation over time due to concept drift, predating the original paper's claims. The candidate paper presents comprehensive empirical evaluations showing temporal performance decay, distributional shifts, and feature instability across multiple years of Android malware data. Both papers analyze how ML models degrade when tested on temporally distant data, use similar evaluation methodologies (temporal train/test splits), and examine feature-level changes over time. TESSERACT[25] explicitly states it 'demonstrated the impact of concept drift on detection performance' and provides evidence of performance degradation patterns that align with the original paper's contribution.

Evidence

Evidence 1 - **Rationale**: TESSERACT[25] explicitly demonstrated the impact of concept drift on malware classifiers before the original paper, directly challenging the novelty of empirically demonstrating concept drift and performance degradation in malware detection. - **Original**: we empirically demonstrate lamda's utility by quantifying performance degradation of standard ml models over time and analyzing feature stability across years - **Candidate**: t esseract [29], inspired by other works [6], [27], demonstrated the impact of concept drift on malware classifiers, and provided constraints and suggestions to eliminate experimental bias and have more realistic evaluations

Evidence 2 - **Rationale**: Both papers empirically quantify performance degradation of ML malware detectors over time using F1-score metrics and temporal evaluation splits, showing that TESSERACT[25] already demonstrated this phenomenon. - **Original**: all detectors perform strongly under iid conditions, but their effectiveness declines sharply as the temporal gap from training increases. for instance, lightgbm's f1-score on lamda drops from 97.49% (iid) to 59.48% (near) and 47.24% (far) - **Candidate**: evaluation across five state-of-the-art malware detectors reveals substantial discrepancy in f1-score on the two datasets (see figure 1), with detection performance on apigraph being consistently higher than on transcendent across the same time frame

Evidence 3 - **Rationale**: The original paper acknowledges that prior studies (including TESSERACT[25]) have already shown concept drift leads to performance degradation, indicating this is not a novel finding. - **Original**: prior studies have shown that such concept drift-distributional shifts in benign and malicious samples, leads to significant degradation in detection performance over time - **Candidate**: we postulate that the spatio-temporal constraints in prior work [10], [18], [29], [32] do not entirely eliminate experimental bias, and that there are other nuanced factors to consider

Evidence 4 - **Rationale**: TESSERACT[25] conducted thorough evaluations of temporal and spatial factors affecting malware detection performance, including analysis across multiple datasets and classifiers, demonstrating prior comprehensive drift analysis work. - **Original**: we conduct comprehensive drift analysis, including per-feature distribution shifts, feature stability analysis across malware families (zhang et al., 2020), temporal label flipping analysis - **Candidate**: we identify five novel temporal and spatial bias factors that affect realistic evaluations. we thoroughly evaluate the impact of these factors in the android malware domain on two representative datasets and five android malware classifiers

4. On the limitations of continual learning for malware classification

URL: View paper

Prior Art Analysis

Continual Learning Limitations[59] demonstrates that machine learning-based malware detectors degrade over time due to concept drift, providing comprehensive empirical evaluations of performance degradation patterns. The paper explicitly states that malware classification models 'suffer from catastrophic forgetting' and shows how 'the adversarial nature of malware and continual evolution of benign software (goodware) makes for an inherently non-stationary problem.' The candidate paper evaluates performance degradation across temporal splits, showing that models experience significant accuracy drops over time, with detailed analysis of how 'the distribution shift must be captured while avoiding catastrophic forgetting.' This work predates the original paper and provides similar empirical demonstrations of concept drift effects on malware detection performance.

Evidence

Evidence 1 - **Rationale**: Both papers identify concept drift as causing performance degradation in malware detectors over time, with the candidate explicitly discussing distribution shifts requiring model retraining. - **Original**: prior studies have shown that such concept drift-distributional shifts in benign and malicious samples, leads to significant degradation in detection performance over time - **Candidate**: the adversarial nature of malware and continual evolution of benign software (goodware) makes for an inherently non-stationary problem. to accommodate shifts in the data distribution over time, the model needs to be retrained regularly to maintain its effectiveness

Evidence 2 - **Rationale**: Both papers empirically investigate how malware classification models degrade over time, with the candidate specifically examining catastrophic forgetting and performance degradation. - **Original**: we empirically demonstrate lamda's utility by quantifying performance degradation of standard ml models over time and analyzing feature stability across years - **Candidate**: in this work, we investigate the extent to which malware classification models suffer from catastrophic forgetting (mccloskey & cohen, 1989; ratcliff, 1990; french, 1999), and whether we can address this using approaches from continual learning research

Evidence 3 - **Rationale**: Both papers provide quantitative measurements of performance degradation over temporal splits, demonstrating empirically how model accuracy declines with concept drift. - **Original**: lightgbm's f1-score on lamda drops from 97.49% (iid) to 59.48% (near) and 47.24% (far), alongside a significant rise in the false negative rate, from 1.47% to 50.51% and 64.10%, respectively,-demonstrating increased difficulty - **Candidate**: the mean accuracies of none and joint baselines over the 12 tasks are 93.1% and 95.9%, respectively. we can see that joint performance trends upward with each incremental task. this may be expected, despite changes in the data distribution, as there is more training data in each additional task

5. Learning Temporal Invariance in Android Malware Detectors

URL: View paper

Prior Art Analysis

Temporal Invariance Android[5] explicitly addresses the same phenomenon: learning-based Android malware detectors degrading over time due to distribution drift. The paper systematically investigates how classifiers trained with empirical risk minimization face challenges against distribution shifts, attributing their shortcomings to inability to learn stable discriminative features. This demonstrates prior recognition and analysis of the performance degradation problem in Android malware detection under temporal drift.

Evidence

Evidence 1 - **Rationale**: Both papers identify and analyze the same core problem: Android malware detectors experiencing performance degradation over time due to distribution drift. Temporal Invariance Android[5] systematically investigates this challenge, demonstrating prior empirical work on concept drift in Android malware detection. - **Original**: machine learning (ml)-based malware detection systems often fail to account for the dynamic nature of real-world training and test data distributions. in practice, these distributions evolve due to frequent changes in the android ecosystem, adversarial development of new malware families, and the co... - **Candidate**: learning-based android malware detectors degrade over time due to natural distribution drift caused by malware variants and new families. this paper systematically investigates the challenges classifiers trained with empirical risk minimization (erm) face against such distribution shifts and attribu...

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

URL: View paper

Brief Assessment

Explainable Drift Detection[18] focuses on explaining drift detection mechanisms in malware pipelines, not on empirically demonstrating concept drift patterns. The candidate addresses drift explanation and early retraining strategies rather than comprehensive empirical evaluation of performance degradation over time.

7. Transcend: Detecting concept drift in malware classification models

URL: View paper

Prior Art Analysis

Transcend[9] demonstrates empirical evaluation of concept drift and performance degradation in malware classification models prior to the ORIGINAL paper. The candidate paper presents comprehensive experiments showing how machine learning-based malware detectors degrade over time, including performance degradation patterns across temporal splits, distributional analysis, and feature stability assessments. Transcend[9] specifically evaluates concept drift using both binary classification (Android malware) and multi-class classification (Windows malware) scenarios, demonstrating performance decay when models are trained on older data and tested on newer samples. The paper provides detailed confusion matrices, precision/recall metrics, and statistical assessments showing model degradation over time, which directly overlaps with the ORIGINAL paper's claimed contribution of empirically demonstrating concept drift and performance degradation.

Evidence

Evidence 1 - **Rationale**: Both papers claim to empirically demonstrate concept drift and performance degradation in malware detection models. Transcend[9] explicitly states it identifies concept drift through case studies, which predates the ORIGINAL paper's similar claim. - **Original**: we empirically demonstrate lamda's utility by quantifying performance degradation of standard ml models over time and analyzing feature stability across years - **Candidate**: we show how transcend can be used to identify concept drift based on two separate case studies on android and windows malware, raising a red flag before the model starts making consistently poor decisions due to out-of-date training

Evidence 2 - **Rationale**: Both papers conduct experiments to detect and demonstrate concept drift using temporal evaluation methodologies. Transcend[9] shows performance decay through empirical experiments, which is the same type of contribution claimed by the ORIGINAL paper. - **Original**: we perform a detail concept drift detection using structured temporal splits (dragoi et al., 2022) to show that lamda exhibits pronounced distributional shift than prior benchmark - **Candidate**: this section presents a number of experiments to show how transcend identifies concept drift and correctly marks as untrustworthy the decisions the ncm-based classifier predicts erroneously. we first show how the performance of the learning model introduced in [2] decays in the presence of concept drif...

Evidence 3 - **Rationale**: Both papers present empirical results showing performance degradation through confusion matrices and metrics when models are tested on temporally distant data. Transcend[9] demonstrates this with specific precision/recall values showing degradation, similar to the ORIGINAL paper's approach. - **Original**: table 2 summarizes the performance of malware detectors on both lamda and apigraph under the iid, near, and far evaluation splits. all detectors perform strongly under iid conditions, but their effectiveness declines sharply as the temporal gap from training increases - **Candidate**: table 3a: confusion matrix when the model is trained on drebin and tested on marvin. the confusion matrix in table 3a clearly shows how the model is affected by concept drift as it reports low precision and recall for the positive class representing malicious objects

Evidence 4 - **Rationale**: Both papers conduct comprehensive drift analysis including feature-level assessments. Transcend[9] performs alpha assessment to evaluate algorithm quality and data distribution, which overlaps with the ORIGINAL paper's feature stability and distributional analysis. - **Original**: we conduct comprehensive drift analysis, including per-feature distribution shifts, feature stability analysis across malware families (zhang et al., 2020), temporal label flipping analysis - **Candidate**: the alpha assessment analysis takes into account how appropriate is the similarity-based algorithm when applied to a dataset. it can detect if the final algorithm results still suffer from over-fitting issues despite the efforts of minimizing it using common and well known techniques

8. Fesad: Ransomware detection with machine learning using adaption to concept drift

URL: View paper

Brief Assessment

Fesad[55] focuses on ransomware detection with concept drift adaptation methods. The provided context contains only the title page and metadata, lacking technical content about empirical evaluations, performance degradation patterns, or distributional analysis that would be needed to assess novelty claims.

9. FeSA: Feature selection architecture for ransomware detection under concept drift

URL: View paper

Brief Assessment

FeSA[57] focuses on feature selection for ransomware detection under concept drift, not on comprehensive empirical evaluation of malware detector degradation patterns across multiple methodologies and temporal analysis as in the original paper.

10. Adapting to concept drift in malware detection

URL: View paper

Prior Art Analysis

Adapting Concept Drift[58] demonstrates that machine learning-based malware classifiers experience performance degradation over time due to concept drift, predating the ORIGINAL paper's claims. The candidate shows that when using temporally ordered files, classifier accuracy drops from 98.7% (cross-validation) to 96.7%, explicitly demonstrating concept drift in malware detection. This empirical evaluation of temporal performance degradation and the presence of concept drift was conducted prior to the ORIGINAL paper's work.

Evidence

Evidence 1 - **Rationale**: Both papers identify concept drift as a fundamental challenge in malware detection, with the candidate explicitly stating that classifiers cannot learn future malware patterns due to evolution. - **Original**: prior studies have shown that such concept drift-distributional shifts in benign and malicious samples, leads to significant degradation in detection performance over time. - **Candidate**: malware and other software are constantly evolving, which makes classifiers unable to learn what malware in the future will be like. therefore, methods are required for adapting to this concept drift in malware detection.

Evidence 2 - **Rationale**: The candidate provides concrete empirical evidence of performance degradation (98.7% to 96.7%) when temporal ordering is considered, demonstrating concept drift through quantitative evaluation. - **Original**: we empirically demonstrate lamda's utility by quantifying performance degradation of standard ml models over time and analyzing feature stability across years. - **Candidate**: this classifier achieves an accuracy of 98.7% when using crossvalidation to evaluate the model. in contrast, when using temporally ordered files, the classifier has an accuracy of 96.7% showing the presence of concept drift in malware detection.

Evidence 3 - **Rationale**: Both papers acknowledge that malware evolution drives concept drift, with the candidate explicitly stating that constant evolution prevents classifiers from learning future patterns. - **Original**: prior studies have shown that malware families (i.e., clusters of samples exhibiting similar behavioral traits) play a central role in driving such drifts - **Candidate**: machine learning models have been shown to be effective in classifying files as either malicious or benign. however, malware and other software are constantly evolving, which makes classifiers unable to learn what malware in the future will be like.

Contribution 3: Multi-faceted concept drift analysis framework

Description: The authors develop and apply a comprehensive analytical framework for studying concept drift that includes multiple complementary techniques: per-feature distribution analysis, family-wise feature stability assessment, temporal label drift tracking, and SHAP-based explanation drift analysis to reveal how feature importance changes over time.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Open Challenges of Malware Detection under Concept Drift

URL: View paper

Brief Assessment

Open Challenges Drift[60] discusses concept drift detection and explanation methods at a high level but does not present a comprehensive analytical framework combining per-feature distribution analysis, family-wise stability assessment, temporal label drift tracking, and SHAP-based explanation drift analysis as implemented in the original paper.

Appendix: Text Similarity Detection

Textual similarity detection checked 23 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Continuous Learning for Android Malware Detection

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] LAMDA: A Longitudinal Android Malware Benchmark for Concept Drift Analysis View paper
- [1] Droidevolver: Self-evolving android malware detection system View paper
- [2] Cluster analysis and concept drift detection in malware: A. Mishra, M. Stamp View paper
- [3] Experts still needed: boosting long-term android malware detection with active learning View paper
- [4] Continuous Learning for Android Malware Detection View paper
- [5] Learning Temporal Invariance in Android Malware Detectors View paper
- [6] Fast & furious: On the modelling of malware detection as an evolving data stream View paper
- [7] Temporal-Incremental Learning for Android Malware Detection View paper
- [8] Data drift in android malware detection View paper
- [9] Transcend: Detecting concept drift in malware classification models View paper
- [10] Hybrid multilevel detection of mobile devices malware under concept drift View paper
- [11] Malcertain: Enhancing Deep Neural Network Based Android Malware Detection by Tackling Prediction Uncertainty View paper
- [12] Advancing educational data mining for enhanced student performance prediction: a fusion of feature selection algorithms and classification techniques with dynamic â¦ View paper
- [13] Combating Concept Drift with Explanatory Detection and Adaptation for Android Malware Classification View paper
- [14] Domain Adaptation-Based Deep Learning Framework for Android Malware Detection Across Diverse Distributions View paper
- [15] MORPH: Towards Automated Concept Drift Adaptation for Malware Detection View paper
- [16] Drift forensics of malware classifiers View paper
- [17] Demystifying the evolution of Android malware variants View paper
- [18] Towards Explainable Drift Detection and Early Retrain in ML-Based Malware Detection Pipelines View paper
- [19] Understanding Concept Drift with Deprecated Permissions in Android Malware Detection View paper
- [20] Counteracting Concept Drift by Learning with Future Malware Predictions View paper
- [21] Active Learning-Based Mobile Malware Detection Utilizing Auto-Labeling and Data Drift Detection View paper
- [22] Efficient Concept Drift Handling for Batch Android Malware Detection Models View paper
- [23] Empirical Evaluation of Concept Drift in ML-Based Android Malware Detection View paper
- [24] Revisiting Temporal Inconsistency and Feature Extraction for Android Malware Detection View paper
- [25] Breaking Out from the TESSERACT: Reassessing ML-based Malware Detection under Spatio-Temporal Drift View paper
- [26] BPFDex: Enabling Robust Android Apps Unpacking via Android Kernel View paper
- [27] Aurora: Are Android Malware Classifiers Reliable under Distribution Shift? View paper
- [28] Is it overkill? analyzing feature-space concept drift in malware detectors View paper
- [29] Advancing Android Malware Detection: A Unified Framework with Dynamic Feature Extraction and Privacy-Preserving Collaboration View paper
- [30] Dissecting android malware: Characterization and evolution View paper
- [31] Mitigating Distribution Shift in Graph-Based Android Malware Classification via Function Metadata and LLM Embeddings View paper

- [32] Novel nature-inspired optimization approach-based SVM for identifying the android malicious data View paper
- [33] Corrigendum to Concept drift and cross-device behavior: Challenges and implications for effective android malware detection Computers & Security, Volume 120, 102757 View paper
- [34] Addressing malware family concept drift with triplet autoencoder View paper
- [35] ADAPT: A Pseudo-labeling Approach to Combat Concept Drift in Malware Detection View paper
- [36] Assessing and improving malware detection sustainability through app evolution studies View paper
- [37] Leveraging the first line of defense: a study on the evolution and usage of android security permissions for enhanced android malware detection View paper
- [38] Heterogeneous temporal graph transformer: An intelligent system for evolving android malware detection View paper
- [39] Robust Machine Learning for Malware Detection over Time View paper
- [40] Poster: LLMalware: An LLM-Powered Robust and Efficient Android Malware Detection Framework View paper
- [41] Enabling Privacy-Preserving Cyber Threat Detection with Federated Learning View paper
- [42] Android malware detection using time-aware machine learning approach View paper
- [43] A pragmatic android malware detection procedure View paper
- [44] Concept drift and cross-device behavior: Challenges and implications for effective android malware detection View paper
- [45] DRMD: Deep Reinforcement Learning for Malware Detection under Concept Drift View paper
- [46] Eight years of rider measurement in the android malware ecosystem: evolution and lessons learned View paper
- [47] Comprehensive Android Malware Detection Based on Federated Learning Architecture View paper
- [48] Android malware concept drift using system calls: Detection, characterization and challenges View paper
- [49] TSDroid: A novel android malware detection framework based on temporal & spatial metrics in IoMT View paper
- [50] Android ransomware detection based on a hybrid evolutionary approach in the context of highly imbalanced data View paper
- [51] One step forward, two steps back: ML-based malware detection under concept drift View paper
- [52] FL-MalDrift: a federated learning framework for malware detection under local concept drift View paper
- [53] On the relativity of time: Implications and challenges of data drift on long-term effective android malware detection View paper
- [54] LongCGDroid: Android malware detection through longitudinal study for machine learning and deep learning View paper
- [55] Fesad: Ransomware detection with machine learning using adaption to concept drift View paper
- [56] Transcending transcend: Revisiting malware classification in the presence of concept drift View paper
- [57] FeSA: Feature selection architecture for ransomware detection under concept drift View paper
- [58] Adapting to concept drift in malware detection View paper
- [59] On the limitations of continual learning for malware classification View paper
- [60] Open Challenges of Malware Detection under Concept Drift View paper