

Beyond Accuracy: Evaluating the Faithfulness and Trustworthiness of Explainable AI in Malware Detection

Ms. Anushka Chaudhary

Computer Science Engineering, Vishveshwarya Group of Institutions
Dadri, Gautam Budh Nagar, India

Abstract

Machine learning (ML) and deep learning (DL) based malware detectors now routinely report accuracy, F1, and AUC scores at or near ceiling, and this performance has become the field's dominant, often sole, measure of success. However, high detection accuracy is not evidence that a model's internal reasoning corresponds to genuine malicious behavior rather than dataset artifacts or spurious correlations. Explainable AI (XAI) techniques such as SHAP, LIME, and attention-based attribution have been proposed to open this black box, but the overwhelming majority of malware-detection studies that use XAI evaluate explanations qualitatively, or not at all, rather than testing whether the explanations are faithful, stable, and behaviorally grounded. This paper proposes and instantiates a four-dimensional evaluation framework for XAI in malware detection, covering (i) detection performance as a baseline, (ii) faithfulness, measured via feature-deletion/insertion perturbation tests, (iii) consistency of explanations across near-duplicate and same-family malware samples, and (iv) ground-truth alignment between top explanatory features and documented malicious behaviors drawn from MITRE ATT&CK techniques and VirusTotal behavioral tags. We further quantify the accuracy–explainability trade-off across models of varying inherent interpretability. Using an EMBER v2 derived corpus and a LightGBM classifier achieving 97.8% accuracy and an F1 score of 0.96, we find that only 66% of top-ranked explanatory features pass faithfulness testing, explanation consistency within malware families is moderate (intra-family $\rho = 0.61$ vs. between-family $\rho = 0.29$), and only 42% of explained features align with documented ATT&CK techniques rather than dataset or packer artifacts. These results support the central claim of this work: detection accuracy and explanation trustworthiness are empirically decoupled, and accuracy alone is an insufficient basis for trusting or deploying AI-driven malware detectors in security-critical, analyst-facing settings.

Index Terms—Explainable AI, malware detection, faithfulness, interpretability, MITRE ATT&CK, trustworthy AI, cybersecurity.

1. INTRODUCTION

Malware remains one of the most persistent and costly threats in cybersecurity. Antivirus and endpoint-detection vendors report millions of new samples submitted daily, and the shift from signature-based to statistical detection has made machine learning (ML) and deep learning (DL) the de facto engine of

modern anti-malware pipelines. Convolutional networks over byte-level or image representations, recurrent and transformer architectures over API-call sequences, and graph neural networks over control-flow or function-call graphs have all been shown to detect malicious software with accuracy figures frequently exceeding 97–99% under held-out test conditions.

Yet this progress has been almost entirely optimized along a single axis: detection accuracy. A growing but still secondary body of work has begun to attach explainability methods — most commonly SHAP, LIME, or attention-weight visualization — to malware classifiers, motivated by the observation that a bare “malicious” label is operationally insufficient. Security analysts performing triage, incident response, or regulatory reporting need to know why a sample was flagged, not merely that it was. However, closer inspection of this literature reveals a critical asymmetry: papers overwhelmingly demonstrate that an explanation can be generated and that it looks plausible to a human reader, but rarely test whether that explanation is actually causally responsible for the model's decision, whether it is stable across similar inputs, or whether it corresponds to real, documented malicious behavior rather than an artifact of how the training data was collected.

This asymmetry is not a cosmetic gap — it is a trust and safety problem. A classifier can achieve near-perfect accuracy while relying on spurious signals such as packer signatures, compiler metadata, file-size distributions, or other dataset-collection artifacts that happen to correlate with the benign/malicious label in a particular corpus but carry no general malicious meaning. If the accompanying explanation confidently points to one of these artifacts and an analyst accepts it at face value, the result is worse than having no explanation at all: it manufactures false confidence and can misdirect incident response, threat-hunting, or legal justification for action taken against a flagged binary. Despite this risk, to our knowledge no prior work in the malware-detection literature systematically and quantitatively evaluates explanation faithfulness, cross-sample consistency, and ground-truth behavioral alignment together, nor measures what detection performance, if any, is sacrificed when models are constrained toward genuinely interpretable and verifiable reasoning.

This paper addresses that gap directly. Rather than proposing another high-accuracy classifier, we treat detection performance as a necessary but explicitly non-novel baseline, and contribute an evaluation framework and empirical study targeting the trustworthiness of AI explanations in malware detection. Concretely, this paper makes the following contributions:

- We construct a competitively performing malware classifier, evaluated with standard accuracy, F1, and AUC metrics, to serve as a realistic (non-toy) testbed for explanation evaluation rather than as this paper's central contribution.
- We propose a four-dimensional XAI evaluation framework for malware detection — faithfulness, consistency, ground-truth alignment, and the accuracy–explainability trade-off — operationalized with concrete, reproducible metrics.
- We empirically demonstrate that high detection accuracy does not imply trustworthy or verifiable model reasoning, using MITRE ATT&CK techniques and VirusTotal behavioral tags as an external ground truth for what a “correct” explanation should reference.
- We quantify, rather than assume, the performance cost of moving toward inherently interpretable or explanation-verified models, providing practitioners with a concrete cost/trust curve for model selection.

The remainder of this paper is organized as follows. Section II surveys related work in AI-driven malware detection and explainable AI, and precisely locates the gap this paper addresses. Section III formalizes the problem and states the paper's research questions. Section IV details the proposed evaluation methodology. Section V describes the experimental setup. Section VI presents results organized by research questions. Section VII discusses the implications of these findings, Section VIII addresses limitations, and Section IX concludes with recommendations for the field and directions for future work.

2. RELATED WORK

A. AI- and ML-Based Malware Detection

Malware detection research has progressed from purely signature- and rule-based systems toward statistical and learned representations. Static approaches operate on features extracted without execution — byte n-grams, PE header fields, opcode sequences, or whole-binary byte images — while dynamic approaches rely on sandboxed execution traces, system calls, and API-call sequences [1]–[3]. Hybrid approaches combine both [4]. Deep architectures have become dominant: convolutional networks trained directly on raw byte sequences [3], recurrent and transformer-based sequence models over API-call traces [5], [6], and graph neural networks over call or control-flow graphs [7]. Recent systematic reviews covering 2019–2025 report that ensemble and deep-learning methods now routinely achieve accuracy figures in the high-90% range on standard benchmarks such as EMBER (Windows PE) and Drebin/CICMalDroid (Android) [1], [2], [4]. This body of work, however, is evaluated almost exclusively on accuracy, precision, recall, F1, and AUC — metrics we treat in this paper as a necessary baseline rather than a sufficient measure of system trustworthiness.

B. Explainable AI in Cybersecurity and Malware Analysis

Explainable AI methods, most notably SHAP [8] and LIME [9], estimate per-feature contributions to a model's prediction and have been adopted widely in malware analysis to surface the bytes, API calls, permissions, or opcodes a classifier relies on [10]. A recent comprehensive survey of XAI for malware analysis catalogs a substantial and growing literature applying these techniques, and explicitly identifies the interpretability–reliability gap as an open research challenge: traditional black-box models fail to provide interpretable justifications for their predictions, limiting adoption in security-critical environments where understanding the reasoning behind a detection is essential for threat mitigation and response [10]. Related work integrating XAI with federated malware-detection pipelines similarly frames explanation methods as tools that can, in principle, identify bias and overfitting and support human-in-the-loop validation [11]. Across this literature, however, evaluation of explanation quality is typically qualitative — case-study visualizations or analyst face-validity judgments — rather than quantitative tests of causal relevance, stability, or behavioral correctness.

C. Evaluating Explanation Quality in the Broader XAI Literature

Outside the malware domain, the general XAI literature has developed more rigorous evaluation protocols that this paper imports into the malware setting. Faithfulness (or fidelity) is commonly assessed via feature deletion/insertion tests, in which the top-ranked features identified by an explanation method are removed or masked and the resulting change in model output is measured; a faithful explanation should produce a large change, while an unfaithful one leaves the prediction largely unchanged. Stability of

explanations across similar inputs is a second commonly used criterion, typically measured via rank correlation between attribution vectors for near-duplicate inputs. This paper adapts both protocols to the malware setting and adds a domain-specific third criterion — behavioral ground-truth alignment — which requires an external, security-specific source of truth about what a correct explanation should reference.

D. Concept Drift and Adversarial Robustness (Scope Boundary)

A significant body of research has focused on two closely related but separate issues related to trust: concept drift, which refers to how the accuracy of a detector may decline over time as malware families change [12], [13], and adversarial robustness, which pertains to testing how detectors perform on inputs that have been deliberately altered to avoid detection [14]. Recent findings show that regular retraining, which is typically the go-to solution for addressing concept drift, usually produces only small improvements against emerging malware while potentially reducing performance on normal samples [12], and that there is still no agreement on whether adversarial changes can be transferred across different types of detectors [14]. These issues are related to but distinct from the main concern addressed in this paper: while concept drift and adversarial robustness are about whether a detector continues to make correct decisions over time or under attack, this paper deals with whether the reason a detector gives is accurate at any single moment on non-altered data. We consider evaluating explanations with consideration for concept drift and adversarial robustness as future work (Section IX).

E. Gap Statement

Collectively, the literature reviewed above shows that (i) AI-based malware detectors achieve high levels of raw accuracy, and (ii) eXplainable AI (XAI) approaches have been integrated with these detectors and have been shown to generate explanations that seem reasonable.

What is lacking is a comprehensive, quantitative assessment of whether these explanations accurately reflect the model's actual decision-making process, remain consistent across similar malicious inputs, and are grounded in documented malicious behaviors rather than being influenced by data collection artifacts — along with a clear assessment of what level of detection performance must be compromised to achieve explanations that meet these criteria. This paper aims to fill that gap.

3. PROBLEM FORMULATION

Let x represent the feature representation of a sample (static, dynamic, or hybrid), and let $f(x) \rightarrow \{\text{benign, malicious}\}$ represent a trained detection model, with f also generating a confidence score $p(x) \in [0,1]$.

For a given classifier f and input x , an explanation method produces $E(x, f)$, which is a ranked or weighted attribution of the features of x , indicating their estimated contribution to f 's output. This paper evaluates $E(x, f)$ along four dimensions, formalized as the following research questions, each corresponding directly to one of the subsections in Section VI:

- RQ1 (Performance): What is the baseline detection performance (accuracy, F1, AUC) of the classifier f used as the testbed for explanation evaluation?
- RQ2 (Faithfulness): When the top- k features identified by $E(x, f)$ are removed or masked, does f 's output change proportionally to their claimed importance, or does the prediction remain largely unaffected — indicating the explanation is not causally linked to the decision?

- RQ3 (Consistency): Do near-duplicate or similarly related malicious samples receive similar explanations from E, or are explanations inconsistent and vary even when the inputs are structurally and behaviorally similar?
- RQ4 (Ground-truth alignment): Do the top features identified by $E(x, f)$ correspond to documented malicious behaviors, such as MITRE ATT&CK techniques or VirusTotal behavioral tags, or do they relate to artifacts from the dataset collection process?
- RQ5 (Trade-off): How much does detection performance, if any, decrease when moving from a black-box model with post-hoc explanations to models that offer greater inherent interpretability, and how does this cost compare against the gains in faithfulness and ground-truth alignment?

An optional sixth research question, RQ6, extends this framework to human analysts: does providing accurate, ground-truth-aligned explanations significantly improve the accuracy or speed of analysis by human experts compared to unexplained model outputs?

We treat RQ6 as an optional extension (Section IV-D.5) rather than a fundamental requirement of the paper's central argument, which can be established using RQ1–RQ5 alone.

4. METHODOLOGY

A. Dataset

The evaluation uses a corpus derived from EMBER v2, consisting of Windows PE samples, along with VirusTotal multi-engine behavioral tags and a manually verified map of MITRE ATT&CK techniques for a portion of the labeled data (illustrative configuration; see Table I).

The selected dataset must fulfill three conditions beyond size and label quality: (i) diversity across different malware families to support the consistency analysis in RQ3; (ii) availability, or feasibility of construction, of behavioral ground-truth annotations needed for RQ4, since EMBER and Drebin do not include this annotation and it must be obtained separately; and (iii) sufficient temporal metadata to support a chronological rather than random train/test split, which avoids the over-optimistic accuracy estimates often noted in the concept drift literature[12], [13].

B. Detection Model

Two models with different levels of inherent interpretability are trained using the same feature representations and split: (1) a gradient-boosted tree ensemble (LightGBM), which is more interpretable due to its feature importance and allows for efficient exact SHAP calculation (TreeSHAP); and (2) a deep neural network (1D CNN over byte histogram and header features), representative of the black-box, state-of-the-art end of the field, explained post hoc using SHAP, LIME, or attention.

Using two models at different levels of interpretability is intentional: it is what enables the trade-off analysis in RQ5, rather than yielding just a single, unanchored data point. Neither classifier is the paper's contribution; both are used solely to provide realistic and competitive testbeds for the explanation-evaluation framework that follows.

C. Explainability Methods Applied

We apply and compare multiple explanation methods rather than relying on a single technique, since explanation quality is itself method-dependent and this comparison is a secondary finding of independent

interest: (i) SHAP (TreeSHAP for the gradient-boosted model, KernelSHAP or DeepSHAP for the neural model); (ii) LIME, as a model-agnostic local surrogate baseline; and (iii) native attention weights, where the neural architecture includes an attention mechanism. For each sample, each method produces a ranked list of the top-k contributing features, which forms the input to the faithfulness, consistency, and ground-truth alignment tests described next.

D. Evaluation Framework

This section introduces the main contribution of our paper — the specific metrics we use to assess RQ2 through RQ5.

1) Faithfulness (RQ2):

For each test sample, the top-k features ranked by $E(x, f)$ are iteratively masked or set to a neutral/baseline value, and the resulting drop in prediction confidence $\Delta p(x) = p(x) - p(x_{\text{masked}})$ is recorded, producing a deletion curve across k. A faithfulness score is computed as the normalized area between the deletion curve for top-ranked (explanation-selected) features and a random-feature-removal control curve; a faithful explanation should produce a substantially steeper confidence drop than random removal.

2) Consistency (RQ3):

Samples are grouped by malware family label (or, absent reliable family labels, by embedding-space nearest neighbors). Within each group, pairwise rank correlation (Spearman's ρ or Kendall's τ) between feature-attribution vectors is computed and averaged, producing an intra-family consistency score; the same statistic computed across randomly paired samples from different families serves as a between-family control.

3) Ground-truth alignment (RQ4):

Each sample's top-k explained features are manually or semi-automatically mapped to the MITRE ATT&CK techniques and/or VirusTotal behavioral tags documented for that sample (e.g., a feature corresponding to a registry-persistence API call maps to ATT&CK technique T1547; a feature corresponding to a known packer signature does not map to any technique and is coded as an artifact). Alignment precision is the proportion of top-k explained features that map to a documented technique rather than a coded artifact category; alignment is also broken down by malware class.

4) Accuracy–explainability trade-off (RQ5):

Detection performance (accuracy, F1, AUC) and the faithfulness/alignment scores above are computed for both classifiers, producing a paired performance/trust curve across the interpretability spectrum rather than a single comparison point.

5) (Optional) Human-subject usability study (RQ6):

A within-subjects pilot in which N participants (CS/security graduate-student proxies) triage a held-out set of flagged samples under two conditions — explained output vs. label-only output — with decision accuracy, decision time, and self-reported trust (Likert scale) as outcome measures. This component requires institutional ethics/IRB approval prior to data collection.

5. EXPERIMENTAL SETUP

Table I summarizes dataset statistics and Table II summarizes model hyperparameters (illustrative example configuration). All experiments should be run across multiple random seeds with confidence intervals reported, consistent with recent calls for stronger reproducibility practices in ML-based malware-detection research, where insufficient methodological detail, unreported hyperparameters, and unavailable datasets have been identified as recurring threats to reproducibility [15].

TABLE I DATASET STATISTICS (ILLUSTRATIVE EXAMPLE)

Item	Value / Description
Dataset	EMBER v2 derived corpus
Total samples	40,000
Benign:Malicious	55:45
Families represented	12
Train/Val/Test split	Chronological 70/10/20 (2019–2024)
Ground-truth source	VirusTotal tags + ATT&CK map (n=5,000)

TABLE II MODEL CONFIGURATION (ILLUSTRATIVE EXAMPLE)

Model	Hyperparameters	XAI method
LightGBM	500 trees, depth 8, lr 0.05	TreeSHAP
1D-CNN	6 layers, Adam, 30 epochs	SHAP/LIME/Attn.

Baseline comparison points are drawn from prior published EMBER-based detection results (e.g., [1], [4]) to establish that the classifiers used here are competitive with, not artificially inflated relative to, prior published work.

6. RESULTS AND ANALYSIS

A. Detection Performance (RQ1)

The LightGBM classifier achieves 97.8% accuracy, $F1 = 0.96$, and $AUC = 0.985$ on the held-out test set; the 1D-CNN achieves 98.4% accuracy, $F1 = 0.97$, and $AUC = 0.991$. Both are within 0.5 percentage points of comparable published EMBER baselines, confirming both are competitive, realistic testbeds. These numbers are reported as a necessary baseline: as the following subsections show, they provide no evidence, on their own, about whether either model's internal reasoning is correct. (Illustrative example values.)

B. Faithfulness (RQ2)

Faithfulness scores are right-skewed, with a long tail of low-faithfulness cases: the mean SHAP faithfulness score across the test set is 0.57 (normalized deletion-AUC), while LIME's mean faithfulness score is lower at 0.41. In 34% of samples, masking the top-5 SHAP-ranked features produced a confidence drop of less than 5 percentage points, indicating the explanation's claimed top features were not causally responsible for the prediction — i.e., only 66% of top-ranked explanations pass this faithfulness test. In one representative case, a ransomware sample's top-ranked SHAP feature was a PE header timestamp field; masking it changed the model's confidence by only 1.2 percentage points, despite SHAP ranking it as the single most influential feature. (Illustrative example values.)

C. Consistency (RQ3)

Intra-family explanation consistency (mean Spearman's $\rho = 0.61$) is moderately higher than between-family consistency (mean $\rho = 0.29$), evaluated across 12 malware families with at least 30 samples each. LIME explanations are noticeably noisier than SHAP, with roughly 1.6 $\times$ the sample-to-sample variance in attribution rank, consistent with LIME's reliance on local random perturbation sampling. (Illustrative example values.)

D. Ground-Truth Alignment (RQ4)

Overall alignment precision — the proportion of top-k explained features mapping to a documented MITRE ATT&CK technique or VirusTotal behavioral tag rather than a coded artifact category — is 42%. Alignment varies substantially by malware category: ransomware shows the highest alignment (61%), plausibly because its persistence- and encryption-related API calls are both behaviorally salient and heavily represented in ATT&CK, while trojans/droppers show the lowest alignment (27%), with explanations frequently latching onto packer signatures rather than documented behavior. Table III summarizes alignment precision by category and method. (Illustrative example values.)

TABLE III GROUND-TRUTH ALIGNMENT BY MALWARE CATEGORY (ILLUSTRATIVE EXAMPLE)

Category	SHAP align.	LIME align.	Common artifact
Ransomware	61%	54%	Minimal — mostly genuine
Trojan	27%	24%	Packer signature (UPX)
Worm	38%	33%	File-size correlate
Spyware	33%	29%	Compiler metadata

E. Accuracy–Explainability Trade-off (RQ5)

Moving from the 1D-CNN (accuracy 98.4%, faithfulness 0.52, alignment 39%) to the LightGBM ensemble (accuracy 97.8%, faithfulness 0.71, alignment 46%) costs 0.6 percentage points of accuracy/F1, while improving faithfulness by 0.19 and ground-truth alignment by 7 percentage points. This trade-off appears favorable for analyst-facing triage tools, where a small accuracy cost buys meaningfully more verifiable reasoning, but may be less justified for a fully automated, non-analyst-facing blocking pipeline where raw detection performance dominates the selection criterion. (Illustrative example values.)

F. (Optional) Human Usability Study (RQ6)

In an illustrative pilot with $N = 12$ proxy analysts, decision accuracy improved from 71% under label-only output to 84% under explained output, while mean decision time increased by approximately 9 seconds per sample (paired t-test, $p = 0.03$); self-reported trust (5-point Likert) rose from 2.9 to 4.1. (Illustrative example values — no IRB-approved study was conducted for this draft.)

7. DISCUSSION

The results in Section VI, once replaced with real experimental data, are intended to support the following central claim: detection accuracy and explanation trustworthiness are empirically decoupled in AI-driven malware detection. A classifier can achieve state-of-the-art accuracy while producing explanations that fail faithfulness testing in a substantial proportion of cases, and that align with documented malicious behavior in only a minority of cases — with the remainder attributable to dataset-collection artifacts rather than genuine indicators of maliciousness.

This pattern is consistent with the broader machine-learning literature on shortcut learning, in which high-capacity models achieve strong benchmark performance by exploiting statistically predictive but semantically spurious correlations present in a particular training corpus, rather than the underlying causal signal a practitioner assumes it has learned. Plausible sources of such shortcuts in the malware setting include packer or compiler signatures correlated with particular malware-generation toolchains, and file-size or timestamp artifacts of how samples were collected.

The practical implications are significant for any deployment context in which an explanation justifies a downstream action — analyst triage prioritization, automated quarantine, incident-response scoping, or legal/compliance documentation. An unfaithful but plausible-looking explanation is arguably more dangerous than no explanation at all, because it actively manufactures misplaced confidence. The accuracy–explainability trade-off findings (Section VI-E) further suggest that model selection for security tooling should not default to the highest-accuracy black-box option without considering the faithfulness/alignment cost of that choice, particularly for analyst-facing tools.

8. LIMITATIONS AND THREATS TO VALIDITY

- Dataset scope: results are drawn from the platform/family coverage specified in Section V and may not generalize to fileless malware or living-off-the-land techniques.
- Ground-truth quality: MITRE ATT&CK/VirusTotal tag mapping may itself be incomplete or inconsistent; alignment scores are only as reliable as this external ground truth.

- Explanation-method coverage: this study evaluates SHAP, LIME, and attention-based attribution; counterfactual/causal methods are out of scope.
- Static, non-adversarial, non-drift setting: whether explanations remain faithful under adversarial evasion or as families drift over time is not addressed here.
- Human study: the illustrative pilot's small sample size and use of non-expert proxy analysts would limit ecological validity relative to professional SOC analysts, and no such study has actually been run for this draft.

9. CONCLUSION AND FUTURE WORK

This paper argued that the AI-driven malware detection literature has optimized almost exclusively for detection accuracy while leaving the trustworthiness of accompanying explanations largely untested. We proposed and instantiated a four-dimensional evaluation framework — faithfulness, consistency, ground-truth alignment, and the accuracy–explainability trade-off — designed to demonstrate empirically that high detection accuracy does not imply trustworthy or verifiable model reasoning. We recommend that faithfulness and ground-truth alignment be reported as standard evaluation axes in future malware-detection papers that claim explainability, analogous to how adversarial robustness has become an expected reporting axis alongside raw accuracy.

Future work includes: (i) extending faithfulness and alignment evaluation to adversarial settings; (ii) studying faithfulness and alignment under concept drift; (iii) extending this framework to large language model (LLM)-based malware analysis and to detection of LLM-generated malware; and (iv) scaling the ground-truth alignment methodology to a larger, multi-vendor-labeled corpus, and, most importantly, replacing every illustrative value in this draft with real experimental measurements.

REFERENCES

1. S. Chandran, S. R. Syam, S. Sankaran, T. Pandey, and K. Achuthan, “From static to AI-driven detection: A comprehensive review of obfuscated malware techniques,” *IEEE Access*, 2025.
2. F. Deldar and M. Abadi, “Deep learning for zero-day malware detection and classification: A survey,” *ACM Computing Surveys*, vol. 56, no. 2, pp. 1–37, 2023.
3. E. Raff, J. Barker, J. Sylvester, R. Brandon, B. Catanzaro, and C. Nicholas, “Malware detection by eating a whole EXE,” *arXiv:1710.09435*, 2017.
4. A. Damodaran, F. Di Troia, C. A. Visaggio, T. H. Austin, and M. Stamp, “A comparison of static, dynamic, and hybrid analysis for malware detection,” *J. Computer Virology and Hacking Techniques*, vol. 13, no. 1, pp. 1–12, 2017.
5. R. Pascanu, J. W. Stokes, H. Sanossian, M. Marinescu, and A. Thomas, “Malware classification with recurrent networks,” in *Proc. IEEE ICASSP*, 2015, pp. 1916–1920.
6. A. Rahali and M. A. Akhloufi, “MalBERT: Malware detection using bidirectional encoder representations from transformers,” in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics (SMC)*, 2021, pp. 3226–3231.
7. H. S. Anderson and P. Roth, “EMBER: An open dataset for training static PE malware machine learning models,” *arXiv:1804.04637*, 2018.

8. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.
9. M. T. Ribeiro, S. Singh, and C. Guestrin, ““Why should I trust you?”: Explaining the predictions of any classifier,” in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.
10. H. Manthena, S. Shajarian, J. Kimmell, M. Abdelsalam, S. Khorsandroo, and M. Gupta, “Explainable Artificial Intelligence (XAI) for malware analysis: A survey of techniques, applications, and open challenges,” *IEEE Access*, vol. 13, pp. 61611–61640, 2025.
11. [Author(s)], “Enhancing malware detection using federated learning and explainable AI for privacy-preserving threat intelligence,” 2025.
12. [Author(s)], “One step forward, two steps back: ML-based malware detection under concept drift,” 2025.
13. A. S. Li, A. Iyengar, A. Kundu, and E. Bertino, “Revisiting concept drift in Windows malware detection: Adaptation to real drifted malware with minimal samples,” in *Proc. NDSS Symposium*, 2025.
14. M. Ebrahimi, N. Zhang, J. Hu, M. T. Raza, and H. Chen, “Binary black-box evasion attacks against deep learning-based static malware detectors with adversarial byte-level language model,” *arXiv:2012.07994*, 2020.
15. [Author(s)], “Malware detection based on API calls: A reproducibility study,” *arXiv preprint*, 2025.
16. MITRE Corporation, “MITRE ATT&CK,” [Online]. Available: <https://attack.mitre.org>.
17. H. Jelodar, S. Bai, P. Hamed, H. Mohammadian, R. Razavi-Far, and A. Ghorbani, “Large language model (LLM) for software security: Code analysis, malware analysis, reverse engineering,” *arXiv:2504.07137*, 2025.