Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 18, No. 2, 2026

The Measurement Framework for Underestimated Failure Modes of Past-only Query Policies in Open-World Malware Family Classification

Jiali Che, Haoqing Ren and Tingting Sun

Sias University, Xinzheng, Zhengzhou, Henan, 451100, China

Abstract: Laboratories that evaluate open-world malware family classifiers typically ship a query or abstention policy chosen on a historical proxy window. The work argues that this past-only selection step constitutes an underestimated failure mode: the proxy-optimal policy can be unsafe on the next unseen-family window even when the candidate grid is fixed. This problem differs from training a better classifier, per-sample rejection, and learning a detection-and-labeling policy with future adaptation. It presents a rigorous measurement framework: a frozen past-only protocol, a ledger of negative selectors, split safety and efficacy estimands, and one registered instance named MPSR-RSG that never reads current future labels. On development windows covering BODMAS PE and EMBER2024 with 18 comparisons, the instance exhibits zero failures. Across 12 baseline-failure windows, its hierarchical-bootstrap 95% CI for failure-rate reduction is [0.25, 1.00].For the preregistered held-out efficacy population consisting of Avast-CTU CAPEv2 and MOTIF with 9 baseline-failure windows, the corresponding CI is [0.22, 1.00].BODMAS Phase 0a cutoffs are retained within the safety monitor rather than the confirmatory efficacy population. This arrangement follows considerations of source role, chain integrity, and estimand identifiability, rather than an attempt to improve the headline interval by excluding these data. The core claim supported by experiments is that past-only selection requires high-priority evaluation under these constraints, and the work does not prove the existence of a universal selector.

Keywords: Malware Family Classification; Open-world Evaluation; Selective Classification; Temporal Evaluation.

1. Introduction

Malware family classification supports incident response and sample triage. Unlike image categories, families change packers, payloads, and naming conventions, so a model that looks strong on a historical validation window can fail when previously unseen families arrive [1]. The literature has responded with better temporal splits, per-sample rejection, adaptive detection policies, and audits of confidence under drift. A second failure mode is still easy to miss: laboratories do not ship a single frozen classifier, they ship a policy-which uncertainty score to trust, whether to query an oracle, and how to abstain-chosen on a proxy window. If that proxy-optimal policy is unsafe on the next unseen-family window, the evaluation protocol itself amplified risk. This phenomenon is termed past-only selection mismatch.

This paper does not claim a new state-of-the-art family classifier, nor the identification of the uniquely correct selector. The research investigates whether past-only selection is measurable as a first-class problem and whether a disciplined protocol can separate "do no harm on already-safe windows" from "reduce failures where the proxy already failed," without reading the current future. The work in a budgeted open-set setting: each cutoff has a labeled past, a proxy evaluation window, and a future window whose unseen-family labels are forbidden to the selector. A candidate is safe if both overall and unseen-family accuracy meet a floor of 0.9; otherwise its utility is treated as zero. The baseline is single-window select.

Always switching to a multi-window robust candidate harms sources that were already safe. Switching only when a past oracle would have failed misses cases where the proxy is already failing. MPSR-RSG is one registered instance: it keeps the proxy by default and intervenes only when past evidence shows a minimax selection-regret candidate is no worse than the proxy and the proxy's past safety pattern is jointly safe with material past regret, jointly unsafe, or unsafe on the most recent past future. It is evidence under constraints, not a proof of a general solution.

2. Contributions

1. A problem statement: past-only selection of query policies is distinct from classifier training, per-sample rejection, online adaptation, and confidence auditing.

2. A split protocol: safety non-inferiority on already-safe windows, efficacy only on baseline-failure windows, hierarchical bootstrap over source → domain → cutoff, and a frozen ledger of negative selectors.

3. One past-only instance (MPSR-RSG) that introduces no new free hyperparameters, inheriting the accuracy floor of 0.9 and adopting two past windows. This method is evaluated on frozen development windows including BODMAS and EMBER2024, alongside a preregistered held-out efficacy population consisting of Avast-CTU and MOTIF.

4. An explicit, three-part rationale for keeping prospective BODMAS 2020-03/04/05 cutoffs as a safety monitor rather than a third confirmatory efficacy source.

This work does not treat EMBER file types as independent sources, does not reopen a burned Win32 confirmation, and does not search hyperparameters on confirmation labels.

3. Related Work

3.1. Temporal Evaluation of Malware Classifiers

TESSERACT formalizes spatial and temporal bias in security classifiers[2].Transcend monitors drift via conformal evaluation[3].Subsequent datasets organize PE malware in

time[4,5,6].MOTIF supplies expert family labels for a long-tail PE corpus[7].Avast-CTU CAPEv2 releases sandbox reports for behavioral research[8].Label noise can change conclusions[9].These works motivate strict time splits; they do not specify how to choose a query policy from proxy windows.

3.2. Rejection, Adaptation, and Confidence: TRANSCENDENT, DRMD, Aurora

TRANSCENDENT revisits Transcend as a practical rejection framework: conformal evaluators score each test sample and quarantine examples likely to be misclassified under drift[10].The decision unit is the sample; the training set and scoring model are given.DRMD instead trains a deep RL agent that jointly classifies, rejects, and requests labels under a one-step MDP, with time-aware AUT gains on Android detection[11].Adaptation and labeling are part of the learned policy; the agent is not a past-only selector over a frozen candidate grid.Aurora asks whether selective methods remain reliable: whether confidence ranks errors (AURC) and whether rejection quotas stay stable under shift[12].The decision unit is again a model's confidence profile, not which of several query policies to deploy from past windows.

Table 1. Related systems versus past-only policy selection

Work	Decision	Uses current future / new labels?
TRANSCENDENT	Reject or accept a sample via conformal p-values	Scores the test point; does not pick a policy from past windows
DRMD	Classify, reject, or query under a learned MDP	Policy is trained for ongoing adaptation and AL
Aurora	Audit confidence–error ranking and rejection stability	Evaluation of a given scorer, not a past-only switch
This paper	Choose among frozen query policies	Past proxy + two past futures only

Table 1 states the measured research gap. TRANSCENDENT can reject a sample that a deployed policy would have classified; it does not choose among entropy/KNN/SAGE query rules using only past futures. DRMD can spend a labeling budget to track drift; the evaluated selector is forbidden from reading the current future or retraining. Aurora can show that a confidence function is miscalibrated; it does not output a past-only switch between already-trained candidates. The three lines of work are complementary. Treating them as substitutes would hide the selection mismatch.

3.3. Open-world Recognition and Selective Classification

Open-set recognition and unknown-family detection are active [13,14,15]. Selective classification studies when to abstain [16]. TRANSCENDENT instantiates rejection for malware drift; Aurora questions whether the resulting confidence remains trustworthy. The objective of this work is not a new open-set score, but which already-trained candidate formed by combining uncertainty metrics and query rules to deploy when the future family mix changes.

3.4. Model Selection under Distribution Shift

Domain generalization literature shows that validation risk can be a poor selector and that selection protocols are first-class [17,18]. Minimax-regret thinking prefers equalizing regret across environments rather than raw risk. This idea is instantiated at the policy layer: candidates are query policies, environments are past futures, and the default action remains the current proxy choice.

3.5. Negative Selection Rules

Always switching to a multi-window robust candidate harms sources whose proxy was already safe. Switching only after a past oracle fails to transfer, or only after the most recent proxy fails, repeats those two error types. MPSR-RSG is a separately registered instance, not a threshold tweak of a burned negative.

4. Setting

4.1. Windows, Candidates, and Safety

For a source × domain time chain, each cutoff t yields a proxy matrix and a future matrix over the same candidate grid.Candidates combine a tree classifier which adopts LightGBM and CatBoost where available, an uncertainty score measured by entropy, maximum softmax probability or KNN, and a query rule including direct, cluster medoid and SAGE-style open-set query, under a small label budget that accounts for 1% of assigned mass in the reported matrices.Let $U(c, w)$ be the safe utility of candidate c on window w: if overall accuracy and unseen-family accuracy are both at least 0.9, utility is correct labels per queried label; otherwise utility is 0 and c is unsafe. Windows with no unseen assigned samples are excluded, not counted as successes.

Single-window selection c_0 = select (proxy) maximizes proxy safe utility with the frozen tie-break of the protocol. Selection regret on a window is

$$R_w(\mathrm{c}) = U(\mathrm{c}^*_w, w) - U(\mathrm{c}, w),\ \mathrm{c}^*_w = \arg\max_{\mathrm{c}'} U(\mathrm{c}', w) \quad (1)$$

4.2. Past-only Constraint

The selector may read the current proxy and the futures of the two immediately preceding cutoffs on the same chain.It must not read the current future, confirmation labels, or any hyperparameter grid.File types of EMBER2024 (.NET, Win64) count as one source.

4.3. Development, Confirmation, and Prospective BODMAS Cutoffs

Development uses BODMAS windows_pe (frozen chain including the 2019-11 and 2020-01 cutoffs) and EMBER2024 dotnet/win64.Later BODMAS months 2020-03/04/05 are not inserted into that chain: doing so would rewrite the past futures of already frozen development comparisons and would un-freeze the development claim. Those three cutoffs can appear only as prospective windows after the frozen chain.

Confirmation efficacy is preregistered on Avast-CTU windows_pe_cape and MOTIF windows_pe_long_tail—sources that were not used to invent the rule.BODMAS is already a development source. Counting it again as a third confirmatory efficacy source would not be an independent

replication. The prospective windows are therefore scored as safety monitors on the development source: introduced failures must remain zero, and unrepaired baseline failures are reported in full.

A third reason is identifiability, not a post-hoc dislike of a wide interval.Efficacy asks whether a past-only rule can repair a baseline future failure.Older-only past unsafety, which manifests as being unsafe on $past_1$ and safe on $past_2$, is the same past observable as windows where switching has introduced a future failure under weaker rules, for instance MOTIF year_2020.The registered invariant is therefore: do not intervene on older-only.On such windows, repair is not identified from past information.Mixing unidentified comparisons into a source-equal hierarchical efficacy bootstrap answers a different question, namely a mean that pools zeros that no past-only rule is allowed to fix.Those windows are retained in the safety estimand, and raw outcomes are provided to enable recomputation of a pooled three-source statistic when needed."The pooled CI would include zero" is not adopted as the exclusion criterion, as this practice would be outcome-dependent.The exclusion is role, chain integrity, and identifiability, all stated in the method registration before the confirmatory run.

5. Method: MPSR-RSG

Let $past_1$ be the older past future and $past_2$ the more recent one. Write S for candidates safe on both past futures. Let c_{rob} be the minimax-regret member of S (ties broken as in the frozen worst-window robust selector of the protocol).Let $\delta = 1 - 0.9 = 0.1$ be the inherited materiality threshold. Algorithmically:

1. c_0 =select (current proxy).
2. If $S = \phi$, output c_0.
3. If $\max_w R_w(c_{rob}) > \max_w R_w(c_0)$ $\max_w$, output c_0 (regret gate).
4. If c_0 is safe on both past futures: if any $R_w(c_0) < \delta$, output c_0 (materiality; do not leave a nearly oracle-safe proxy); else output c_{rob}.
5. Else if c_0 is unsafe on both past futures or unsafe on $past_2$, output c_{rob}.
6. Else (c_0 unsafe only on $past_1$), output c_0 (older-only hold).

Unlike a consensus gate that required c_{rob} to equal that worst-window robust selector, MPSR-RSG does not. Consensus blocked switches when the proxy was already recently unsafe solely because two jointly safe candidates disagreed on a secondary tie-break—a development-visible mechanism, not a confirmation-tuned threshold.

6. Evaluation Protocol

Two estimands are distinguished in this work.

Safety non-inferiority. Population: comparisons with baseline future failure false. Estimand: count of introduced failures (rule unsafe, baseline safe). Pass: zero on development and on all confirmation windows, including BODMAS monitors.

Efficacy superiority. Population: comparisons with baseline future failure true. Estimand: mean failure reduction (1 if the rule is safe and the baseline is not, 0 if both fail).Pass: at least two independent sources contribute a baseline-failure window, and the hierarchical bootstrap 95% CI lower bound of the mean is strictly positive.20,000 replicates are used for resampling sources, then domains, then cutoffs, with seed 20260904.Conditional mean regret reduction on the same population is secondary.

Table 2. Development split

Estimand	Mean	95% CI
Introduced failure rate (already-safe windows)	0.00	[0.00, 0.00]
Failure-rate reduction (baseline-failure windows)	0.65	[0.25, 1.00]
Conditional mean regret reduction	0.69	[0.30, 1.00]

Eligibility: development has 12 baseline-failure windows on BODMAS and EMBER2024.The confirmatory efficacy population has 9 baseline-failure windows on Avast-CTU and MOTIF. Prospective BODMAS 2020-03/04/05 windows are eligible for the safety estimand and ineligible for confirmatory efficacy, as argued in Section 3.3.

7. Results

7.1. Development

Eighteen two-window comparisons are usable after excluding zero-unseen windows.Six are already safe under the proxy; MPSR-RSG introduces zero failures there.Of 12 baseline-failure windows, 6 are repaired.Table 2 reports hierarchical bootstrap summaries.These numbers describe this registered instance under the frozen chain.

Typical holds are materiality, older-only hold, empty jointly-safe set, and the regret gate.Switches occur when the proxy is jointly past-safe with material regret, or when it is unsafe on both past futures or on the recent past future.

7.2. Confirmation

The full confirmation set has 25 comparisons (Avast-CTU, MOTIF, and three BODMAS prospective monitors). Introduced failures: 0 on every source, including the monitors.

The preregistered confirmatory efficacy population is Avast-CTU and MOTIF, which include 22 comparisons, 9 baseline-failure windows, among which 5 are repaired.See Table 3 for detailed quantitative results.Figure 1 shows every confirmation window.There are no introduced failures.FIX rows cluster on Avast 2016-09/11, Avast 2017-05, and MOTIF 2017-H1/2019-H1; STILL rows are baseline failures the rule did not repair.MOTIF 2020 remains SAFE under an older-only hold the past signature that earlier was safe turned into an introduced failure.

Table 3. Confirmatory efficacy population: Avast-CTU CAPEv2 and MOTIF

Estimand	Mean	95% CI
Introduced failure rate (already-safe windows in this population)	0.00	[0.00,0.00]
Failure-rate reduction	0.57	[0.22,1.00]
Conditional mean regret reduction	0.52	[0.20,0.83]
Repaired / baseline-failure windows	5/9	-

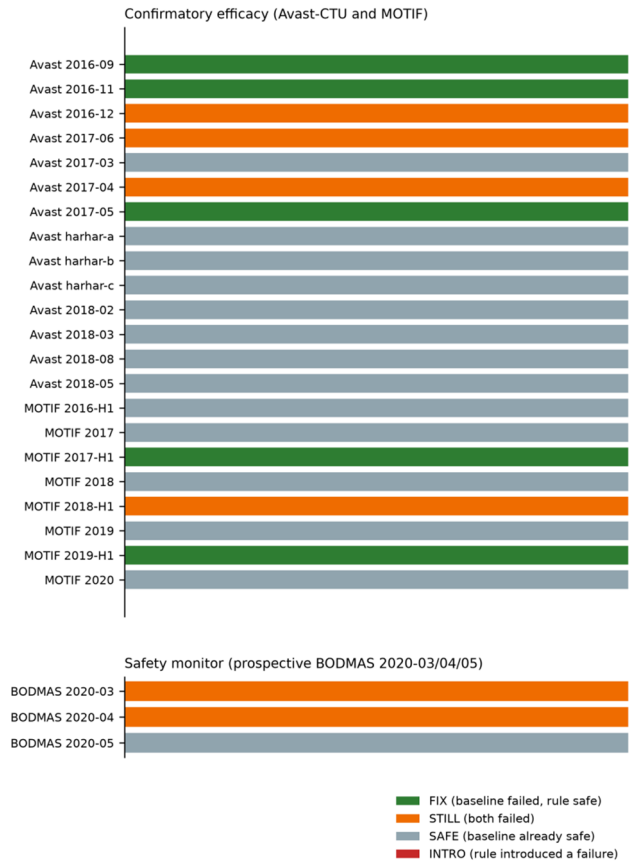


Fig 1. Confirmation windows under the registered past-only instance.

BODMAS prospective monitors (3 comparisons in Fig. 1, lower panel): introduced failures 0; two baseline failures remain unrepaired. Both have an older-only past signature, so the registered rule holds the proxy. They are not omitted from the paper; they are omitted from the confirmatory efficacy bootstrap. Pooling all 11 confirmation baseline-failure windows yields a wider interval whose lower bound can include zero; that pooled figure mixes identified and unidentified repairs and is not the work confirmatory estimand.

8. Discussion

The intended contribution is a problem and a protocol, not a proof that MPSR-RSG is the right selector. Past-only selection mismatch is readily concealed by the statement that the best validation candidate is selected. TRANSCENDENT, DRMD, and Aurora address neighboring operational questions: sample rejection, adaptive detect-and-query, confidence reliability, and should be cited as such, not as having already solved policy choice. Under a frozen grid, it shows that always-robust and failure-triggered selectors are harmful or incomplete, that a split safety/efficacy measurement is feasible, and that one registered past-only instance can pass that measurement on two held-out sources without introducing failures. That is disciplined evidence. It is not a license to claim a general, cross-product selector.

The prospective BODMAS cutoffs represent a potential risk of selective result optimization. The work gives three defenses that do not depend on the numerical width of the confirmatory CI: (i) BODMAS is a development source, so confirmatory efficacy is defined on other sources;(ii) those months cannot enter the frozen development chain;(iii) older-only failures are unidentified for a past-only repair estimand and remain in the safety monitor (Fig. 1). All monitor outcomes are reported.

Limitations remain. Avast-CTU family cardinality is small. MOTIF is long-tail and sparse. Candidate grids are not identical across every legacy matrix. The dual regime was invented from a development ledger of prior negatives; confirmation labels were not used to grid-search thresholds, but the form of the rule is adaptive to that ledger. Win32 confirmation remains burned. Evaluations are not performed for Android, PDF samples, or live vendor telemetry.

9. Conclusion

Open-world family evaluation routinely selects a query policy on a proxy window. That selection step can fail on the next unseen-family window, and it is not the same problem as training a stronger classifier, rejecting a sample (TRANSCENDENT), learning an adaptive detect-and-label policy (DRMD), or auditing confidence (Aurora). A past-only protocol, a negative ledger, and one registered instance with split safety and efficacy evidence across development windows, Avast-CTU and MOTIF are provided. Prospective BODMAS 2020-03/04/05 cutoffs are a safety monitor with a stated identifiability argument, not a silenced third source. The claim to take away is that past-only selection deserves first-class measurement, not that the problem has been solved.

References

[1] Arp, D., Quiring, E., Pendlebury, F., et al. (2022). Dos and don'ts of machine learning in computer security. In Proceedings of the 31st USENIX Security Symposium (pp. 3971–3988). USENIX Association.

[2] Pendlebury, F., Pierazzi, F., Jordaney, R., et al. (2019). TESSERACT: Eliminating experimental bias in malware classification across space and time. In Proceedings of the 28th USENIX Security Symposium (pp. 729–746). USENIX Association.

[3] Jordaney, R., Sharad, K., Dash, S. K., et al. (2017). Transcend: Detecting concept drift in malware classification models. In Proceedings of the 26th USENIX Security Symposium (pp. 625–642). USENIX Association.

[4] Yang, L., Ciptadi, A., Laziuk, I., et al. (2021). BODMAS: An open dataset for learning based temporal analysis of PE malware. In 2021 IEEE Symposium on Security and Privacy Workshops (SPW) (pp. 78–84). IEEE.

[5] Joyce, R. J., Miller, G., Roth, P., et al. (2025). EMBER2024: A benchmark dataset for holistic evaluation of malware classifiers. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 5516–5526). ACM.

[6] Zola, F., Bruse, J. L., Galar, M., et al. (2023). Temporal analysis of distribution shifts in malware classification for digital forensics. In 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) (pp. 439–450). IEEE.

[7] Joyce, R. J., Amlani, D., Nicholas, C., et al. (2023). MOTIF: A malware reference dataset with ground truth family labels. Computers & Security, 124, 102921. https://doi.org/10.1016/j.cose.2022.102921.

[8] Bošanský, B., Kouba, D., Manhal, O., et al. (2022). Avast-CTU public CAPE dataset. arXiv:2209.03188.

[9] Wu, X., Guo, W., Yan, J., et al. (2023). From grim reality to practical solution: Malware classification in real-world noise.

In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 2602–2619). IEEE.

[10] Barbero, F., Pendlebury, F., Pierazzi, F., et al. (2022). Transcending TRANSCEND: Revisiting malware classification in the presence of concept drift. In 2022 IEEE Symposium on Security and Privacy (SP) (pp. 805–823). IEEE.

[11] McFadden, S., Foley, M., D'Onghia, M., et al. (2026). DRMD: Deep reinforcement learning for malware detection under concept drift. In Proceedings of the AAAI Conference on Artificial Intelligence, 40(2), 854–862.

[12] Herzog, A., Eusebi, A., Cavallaro, L. (2025). On the reliability and stability of selective methods in malware classification tasks. arXiv:2505.22843.

[13] Vaze, S., Han, K., Vedaldi, A., et al. (2022). Open-set recognition: A good closed-set classifier is all you need. In Proceedings of the International Conference on Learning Representations (ICLR).

[14] Lu, T., & Wang, J. (2024). DOMR: Toward deep open-world malware recognition. IEEE Transactions on Information Forensics and Security, 19, 1455–1468. https://doi.org/10.1109/TIFS.2024.3357040.

[15] Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the International Conference on Learning Representations (ICLR).

[16] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In Advances in Neural Information Processing Systems 30 (NeurIPS 2017) (pp. 1–12). Curran Associates.

[17] Gulrajani, I., & Lopez-Paz, D. (2021). In search of lost domain generalization. In Proceedings of the International Conference on Learning Representations (ICLR).

[18] Ye, H., Xie, C., Cai, T., et al. (2021). Towards a theoretical framework of out-of-distribution generalization. In Advances in Neural Information Processing Systems 34 (NeurIPS 2021) (pp. 24608–24621). Curran Associates.