Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39343

# *Robustness and Probability Calibration of Multi-class Memory Malware Classification under Distribution Shift and Class Imbalance: A Cross-Dataset Empirical Study*

**Fangke Fu**

*Department of Computer Science, Beijing University of Civil Engineering and Architecture, Beijing, China*
*202308010107@stu.bucea.edu.cn*

***Abstract.*** Memory malware classification faces distribution shift and class imbalance in real-world deployment, yet most studies assume identical training and test distributions and overlook probabilistic reliability. This study builds an empirical evaluation framework based on the BCCC-MalNetMemLog-2026 dataset (five time points) and the cross-domain BCCC-mal-netmem-2025 dataset. Four imbalance strategies, two classifiers, three calibration methods, and a conformal rejection mechanism are compared. Results show that class weighting is more stable than SMOTE under shift; calibration improves the Brier score and ECE but reduces Macro-F1 in cross-domain four-class classification; conformal rejection based on maximum softmax probability fails cross-domain, with accepted-set Macro-F1 decreasing from 0.205 to 0.157 as the significance level rises from 0.01 to 0.5; and labeled target-domain samples are the most critical factor, with Macro-F1 rising from 0.191 to 0.358 after random 30% sampling, outperforming active sampling.

***Keywords:*** memory malware classification, distribution shift, class imbalance, probability calibration, selective prediction

## 1. Introduction

Cevallos-Salas [1] performed multiclass classification on CIC-MalMem-2022 using SMOTE but validated only under identical distributions; TESSERACT [2] and Transcendent [3] proposed fairness-oriented evaluation and conformal prediction rejection but targeted static samples, leaving the coupled shift-and-imbalance setting for memory dumps unexamined. Existing studies focus on Macro-F1 and ROC-AUC while neglecting probabilistic reliability (Brier score, ECE) and minority-class performance (Worst Recall), and cross-dataset evaluation remains limited. This study contributes: (1) an evaluation framework for coupled shift and imbalance; (2) systematic comparison of four imbalance strategies, two classifiers, three calibration methods, and a conformal rejection mechanism; (3) evaluation of CORAL, active sampling, and adaptive rejection, including failure phenomena in cross-domain four-class classification.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39343

## 2. Related work

Cevallos-Salas [1] demonstrated memory-dump features but ignored shift, calibration, and cross-domain generalization. SMOTE, class weighting, and importance weighting are common yet most assume identical distributions. TESSERACT [2] emphasized fairness under distribution change but used static Android samples; this study covers covariate, class-prior, and temporal shift. Transcendent [3] used conformal prediction but focused on static samples without analyzing rejection-rate trade-offs. Platt [4] and Zadrozny and Elkan [5] mitigate overconfidence, but coupled memory malware studies remain scarce. CORAL [6] performs unsupervised domain adaptation via covariance alignment, yet joint evaluation in cross-dataset memory malware is lacking; this study evaluates SourceOnly through +CORAL+Active30+Calib+Rej configurations.

## 3. Method

### 3.1. Data, drift, and imbalance strategies

Datasets were cleaned and label-unified into four classes (Backdoor, Trojan, Rootkit, Hoax), with 96-dimensional features standardized. Three drift types were constructed: covariate shift (Gaussian noise, sigma=0.3.0), class-prior shift (imbalance ratios 1, 3, 8, 15), and temporal shift (train/val/test = time points 1/2/5). Drift was quantified by KS statistic, Wasserstein distance, and MMD. Four imbalance strategies were compared: none, SMOTE, class weighting, and importance weighting. RandomForest and XGBoost were used as classifiers, with calibration methods: none, Platt scaling, and isotonic regression.

### 3.2. Selective prediction and cross-domain adaptation

A conformal rejection mechanism implements selective prediction with the uncertainty score

$$S_i = 1 - \underset{c}{max} P_{ic} \tag{1}$$

where $S_i$ refers to the uncertainty score of sample $i$ and $P_{ic}$ refers to the predicted probability that sample $i$ belongs to class $c$. The threshold is fitted on the validation set, and the significance level takes values of 0.01, 0.05, 0.10, 0.20, 0.30, and 0.50. For cross-domain adaptation... Alignment, fine-tuning, calibration, and threshold estimation were performed only on the adaptation pool, with the test set excluded from all fitting to prevent data leakage.

## 4. Experiments

BCCC-MalNetMemLog-2026 contains five time points, 9,404 samples, and 96-dimensional features; time points 1/2/5 were used for training, validation/calibration, and testing. BCCC-malnetmem-2025 (1,934 samples) was used only for cross-dataset testing. Five experiment groups were run: no-drift baseline, covariate shift, class-prior shift, temporal shift, and cross-dataset testing. Classification was evaluated by Macro-F1, Macro-Recall, Balanced Accuracy, MCC, Macro-PR-AUC, and Worst Recall; probabilistic reliability by Brier score, ECE, and LogLoss; selective prediction by rejection rate and accepted-set metrics.

DOI: 10.54254/2755-2721/2026.39343

## 5. Results and analysis

### 5.1. Imbalance strategies under covariate and class-prior shifts

Figure 1 shows that as covariate shift increases, Macro-F1 and Balanced Accuracy degrade for all strategies, but probabilistic reliability behaves differently. SMOTE and class weighting are robust: at zero shift SMOTE reaches Macro-F1 ~0.512 and class weighting ~0.510, remaining around 0.510 and 0.514 at maximum shift, whereas none and importance weighting fall to ~0.442 and ~0.436. Counterintuitively, Brier score and ECE decrease for SMOTE (0.247→0.208, 0.344→0.286) and class weighting (0.237→0.203, 0.324→0.278) as shift increases, but increase for importance weighting. Class weighting maintains the most stable rejection rate (0.23-0.26), and accepted-set Macro-F1 stays around 0.51-0.53 for SMOTE and class weighting but falls from ~0.48 to ~0.40 for others.

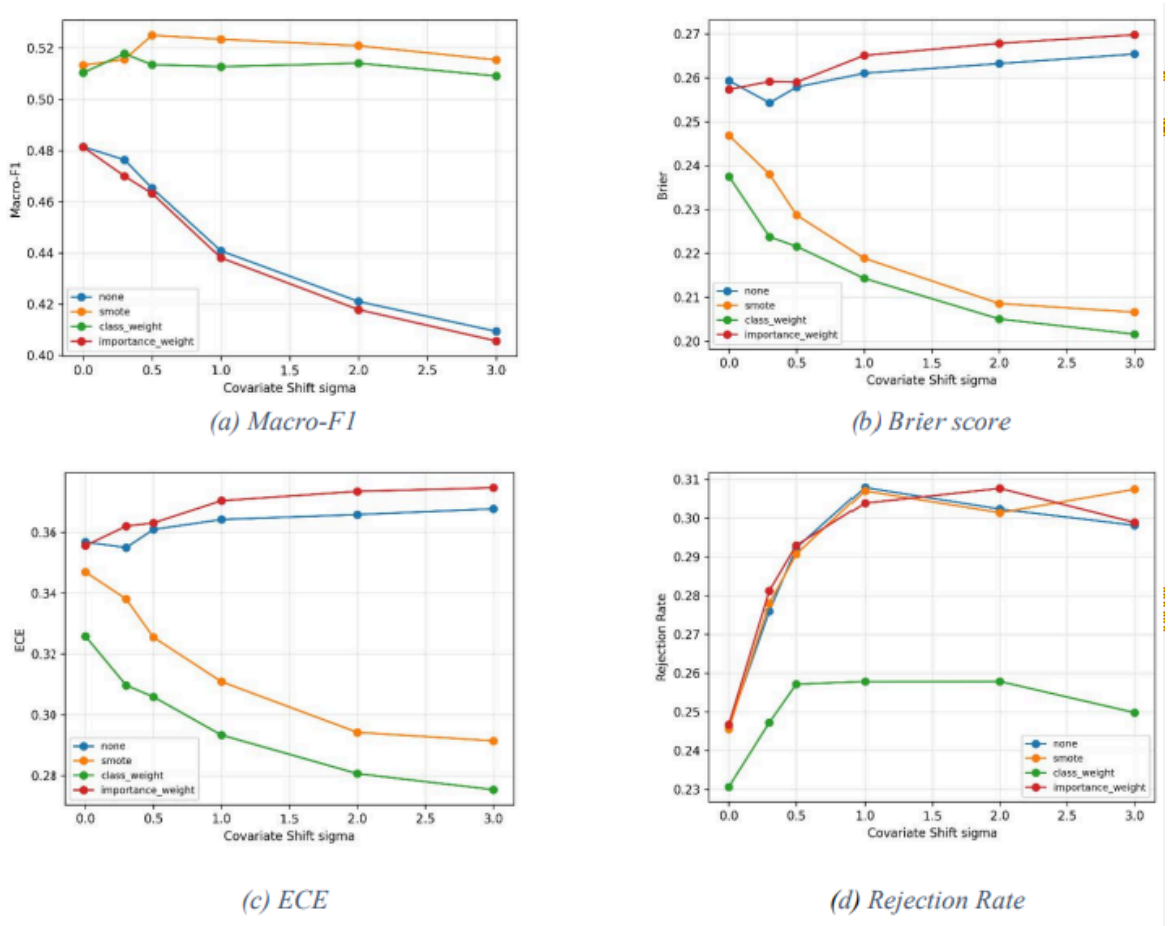



Figure 1. Effect of covariate shift intensity on imbalance strategies

### 5.2. Cross-domain ablation: domain adaptation, sampling, and calibration

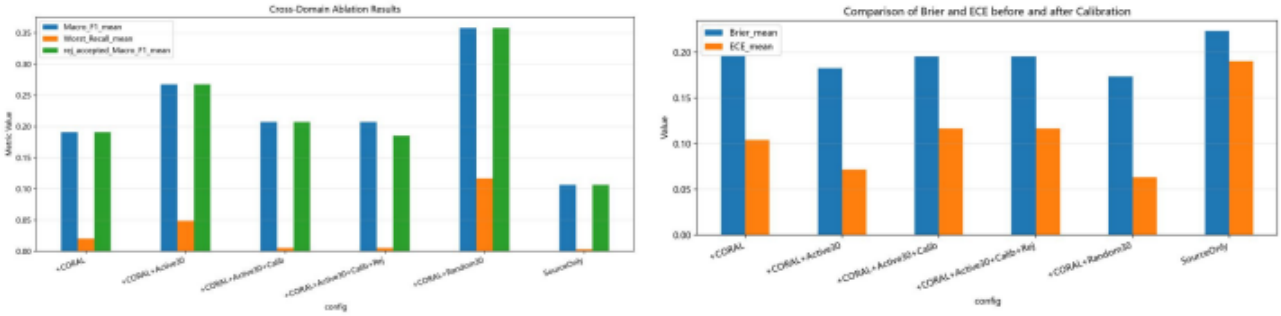
Table 1 and Figure 2 report the cross-domain ablation, averaged over three random seeds.

DOI: 10.54254/2755-2721/2026.39343

Table 1. Cross-domain ablation results (average over three seeds)

| Configuration | Macro-F1 | Worst Recall | Brier | ECE | Acc.-Set F1 |
|---|---|---|---|---|---|
| SourceOnly | 0.106 | 0.003 | 0.224 | 0.190 | 0.106 |
| +CORAL | 0.191 | 0.019 | 0.196 | 0.104 | 0.191 |
| +CORAL+Random30 | 0.358 | 0.117 | 0.174 | 0.063 | 0.358 |
| +CORAL+Active30 | 0.268 | 0.049 | 0.183 | 0.072 | 0.268 |
| +CORAL+Active30+Calib | 0.207 | 0.005 | 0.196 | 0.117 | 0.207 |
| +CORAL+Active30+Calib+Rej | 0.207 | 0.005 | 0.196 | 0.117 | 0.185 |

CORAL provides substantial improvement: Macro-F1 rises from 0.106 to 0.191, Worst Recall from 0.003 to 0.019, Brier from 0.224 to 0.196, ECE from 0.190 to 0.104. Labeled target-domain samples are most critical: random 30% sampling jumps Macro-F1 to 0.358 (~3.4× SourceOnly), Worst Recall to 0.117, and Brier/ECE to 0.174/0.063 — the only configuration bringing minority-class Worst Recall above 0.1. Active sampling underperforms random sampling (0.268 vs 0.358). Calibration and rejection have negative effects: calibration drops Macro-F1 from 0.268 to 0.207 and Worst Recall from 0.049 to 0.005; rejection further lowers accepted-set Macro-F1 to 0.185.




(a) Cross-domain ablation results

(b) Brier score and ECE before and after calibration under cross-domain configurations

Figure 2. Cross-domain ablation results and calibration effects

## 5.3. Failure analysis of calibration and rejection

Figure 2(b) compares Brier score and ECE before/after calibration. SourceOnly: Brier 0.224, ECE 0.190; +CORAL: 0.196/0.104; +CORAL+Random30: 0.174/0.063. However, adding calibration to +CORAL+Active30 increases Brier from 0.183 to 0.192, ECE from 0.072 to 0.112, and drops Macro-F1 from 0.268 to 0.207. This calibration paradox occurs because isotonic calibration alters the argmax decision boundary, turning correct predictions into errors; in cross-domain four-class settings with substantial per-class overlap, calibration amplifies minority-class probabilities, introducing false positives. Calibration should be used only for downstream decisions, not to replace the classifier.

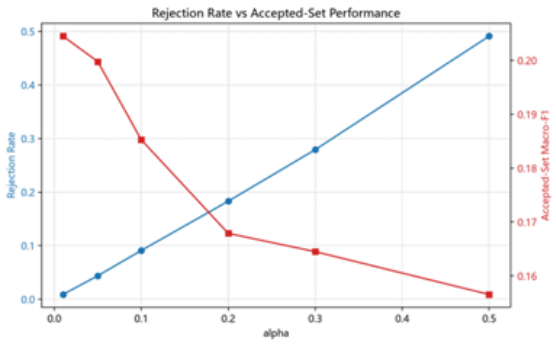
Figure 3(a) shows that as significance level increases from 0.01 to 0.50, rejection rate rises from 0.009 to 0.491, but accepted-set Macro-F1 decreases monotonically from 0.205 to 0.157. Conformal rejection based on maximum softmax probability thus fails in cross-domain classification: the rejected set contains many originally correct samples while difficult minority samples are not

DOI: 10.54254/2755-2721/2026.39343

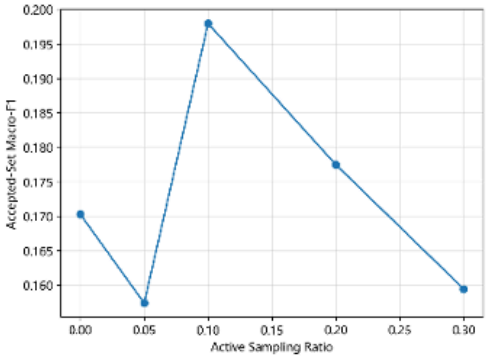
preferentially rejected. The model cannot distinguish shift-induced low confidence from truly difficult samples, so increasing rejection only reduces accepted-set performance. Class-conditional conformal prediction or cost-sensitive rejection is needed.

## 5.4. Active-sampling-ratio ablation

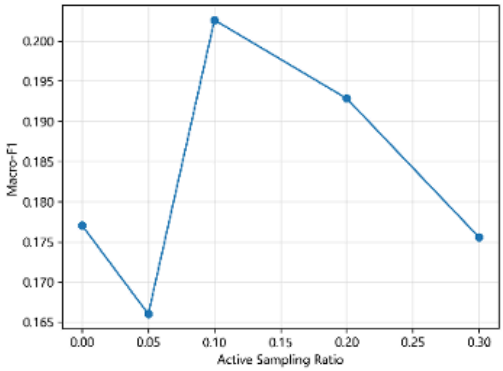
Figure 3(b)-(d) report active-sampling-ratio ablation. As ratio increases to 0.30, Macro-F1 peaks at 0.207 but fluctuates between 0.168-0.207; Worst Recall remains near zero (max 0.005); ECE reaches minimum 0.102 at 0.20 but rises to 0.30. Compared with +CORAL+Random30 (Macro-F1 0.358, Worst Recall 0.117), uncertainty-based active sampling is clearly inferior to random sampling.



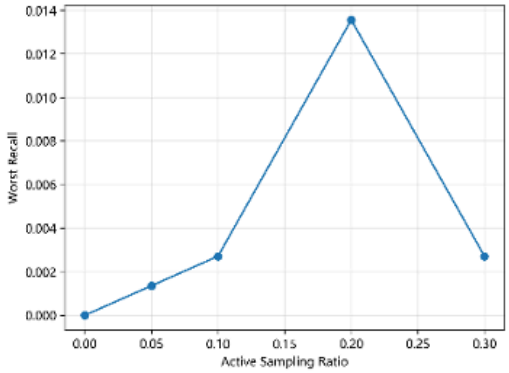

(a) Effect of alpha on rejection rate and accepted-set Macro-F1




(b) Effect of active sampling ratio on accepted-set Macro-F1




(c) Effect of active sampling ratio on Macro-F1




(d) Effect of active sampling ratio on Worst Recall

Figure 3. Effects of alpha and active sampling ratio on model performance

## 5.5. Summary

Class weighting is most stable under covariate shift, followed by SMOTE; SMOTE and class weighting improve Macro-F1 and Worst Recall under class-prior shift. Cross-domain gains come mainly from labeled target-domain samples (random 30% sampling reaches Macro-F1 0.358, better than active sampling), while calibration improves Brier/ECE but can harm Macro-F1 and conformal rejection based on maximum softmax probability fails cross-domain.

DOI: 10.54254/2755-2721/2026.39343

## 6. Discussion and limitations

Isotonic calibration alters the argmax boundary and amplifies minority-class probabilities, so calibration should be used only for downstream decisions, not to replace the classifier. Conformal rejection fails because maximum softmax probability cannot separate shift-induced low confidence from genuinely difficult samples; class-conditional or cost-sensitive rejection is needed. CORAL helps but is limited; labeled target-domain samples remain most critical. Limitations: simulated drift; adaptation-set-tuned thresholds; weak active-learning comparison; limited target-domain samples; and no comparison with stronger domain-adversarial methods. Worst Recall remains extremely low—minority malware families are nearly undetectable cross-domain, the main bottleneck.

## 7. Conclusions and future work

This study built an empirical framework for coupled distribution shift and class imbalance in memory malware on BCCC-MalNetMemLog-2026 and BCCC-mal-netmem-2025. Contributions: (1) an empirical framework for coupled shift and imbalance; (2) systematic comparison of method effectiveness and failure boundaries; (3) correction of data leakage and cross-dataset validation; (4) identification of failure phenomena in active sampling, calibration, and conformal rejection. Future work will collect real temporal malware data, study online drift detection and unknown-family identification, and explore class-conditional conformal prediction, cost-sensitive rejection, and stronger domain adaptation.

## References

[1] Cevallos-Salas, D., Grijalva, F., Estrada-Jimenez, J., Benitez, D., & Andrade, R. (2024). Obfuscated privacy malware classifiers based on memory dumping analysis. *IEEE Access, 12*, 17481–17498.

[2] Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., & Cavallaro, L. (2019). TESSERACT: Eliminating experimental bias in malware classification across space and time. In *Proceedings of the 28th USENIX Security Symposium (USENIX Security 19)*, Santa Clara, CA, USA, pp. 729–746.

[3] Barbero, F., Pendlebury, F., Pierazzi, F., & Cavallaro, L. (2022). Transcending TRANSCEND: Revisiting malware classification in the presence of concept drift. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, pp. 805–822.

[4] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, MIT Press, Cambridge, MA, USA. pp. 61–74.

[5] Zadrozny, B., & Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, Edmonton, AB, Canada, pp. 694–699.

[6] Sun, B., Feng, J., & Saenko, K. (2017). CORAL: Correlation alignment for unsupervised domain adaptation. In *Domain Adaptation in Computer Vision Applications*, Cham, Switzerland: Springer, pp. 153–171.