ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Multimodal Malware Detection under Obfuscated and Adversarial Attack Conditions Using ResNet-50 and MLP Encoders with Contrastive Learning and Adaptive Cross-Modal Fusion

Peter Sackitey | Computer Science Department, Kwame Nkrumah University of Science and Technology, Ghana | ORCID: 0009-0007-2281-2498

Collinson Colin Mawunyo Agbesi | Computer Science Department, Koforidua Technical University, Ghana | ORCID: 0009-0009-3486-4740

Timothy Graham | Computer Science Department, Kwame Nkrumah University of Science and Technology, Ghana | ORCID: 0009-0009-6018-9128

Abstract

Traditional Android malware detection methods such as basic heuristic and signature-based methods as well as those trained on unimodal data samples are exposed to challenges from contemporary malware that utilizes advanced evasion techniques like adversarial and obfuscation perturbed variants to avoid detection. This work presents a multimodal malware detection method that uses image-based and tabular representations of Android APKs. A ResNet-50 network is used to extract visual embeddings from byteplot images that capture structural patterns resilient to obfuscated variants, while tabular data was analysed by a masked autoencoder pre-trained MLP to model behavioural correlations. An adaptive fusion technique dynamically weights the contributions of each modality per sample while Cross-modal contrastive learning

Received: 23.09.2025

Accepted: 09.03.2026

Published: 19.05.2026

Cite this article as:
P. Sackitey, C.C.M. Agbesi, T. Graham, "Multimodal malware detection under obfuscated and adversarial attack conditions using ResNet-50 and MLP encoders with contrastive learning and adaptive cross-modal fusion," ACIG, vol. 5, no. 2, 2026, doi: 10.60097/ACIG/219075.

Corresponding author:
Peter Sackitey, Computer Science Department, Kwame Nkrumah University of Science and Technology, Ghana; E-mail: psackitey5@st.knust.edu.gh
0009-0007-2281-2498

aligns these embeddings in a shared latent space. The fused representation is fed into a supervised classifier to predict benign and malware categories. Adversarial, obfuscated, and clean scenarios confirmed outstanding results on tests samples from the multimodal CIC-MalDroid2020 dataset.

Evaluation metrics include accuracy, recall, precision, F1-score, and macro-averaged receiver operating characteristic–area under the curve (ROC-AUC). Statistical significance is assessed using McNemar's test. The results show that multimodal integration and adaptive fusion greatly improve detection generalization against evolving malware threats.

Keywords

multimodal malware detection, ResNet-50, MLP, adaptive fusion, contrastive learning

1. Introduction

Malware detection has made a significant progress recently with multimodal approaches as one of the promising emerging paradigms. These approaches fuse tabular, image-based, and sequential features to optimise detection accuracy and generalization. Image-based techniques which convert executable files into greyscale images or other visual representations, capture structural patterns without the need for reverse engineering. Methods such as grayscale or byteplot images have demonstrated high accuracy in classifying both adversarial, obfuscated and normal malware by leveraging intensity patterns and visual textures that are inherently less sensitive to conventional obfuscation and adversarial techniques [1, 2]. Tabular representations, including metadata, system call sequences, and bytecode semantics add to visual features by providing rich static and dynamic behavioural information where these features are often modelled using graph-based models or deep learning architectures to capture semantic relationships and dependencies [3, 4]. Integrating diverse modalities allows malware detectors to exploit complementary feature spaces to achieve a level of insight and resilience that unimodal systems cannot provide.

Notwithstanding, the increasing complexity of adversarial and obfuscated malware exposes the traditional detection systems to severe challenges as malware authors employ techniques such as packing, encryption, irrelevant code insertion, and code metamorphism to disguise malicious intent and evade feature-based or signature-based

detectors [2, 5] and adversarial attacks craft subtle perturbations in malware samples to mislead machine learning models without altering the malware's functionality [6, 7]. These strategies expose the vulnerabilities of unimodal detection systems, which often fail to generalise beyond the specific feature space they were trained on.

These weaknesses are addressed by multimodal detection frameworks by combining diverse features and jointly learning representations that cross-validate and fuse evidence from multiple perspectives as approaches of integrating visual features with dynamic behaviours and graph-based static have demonstrated superior generalization against obfuscation and adversarial perturbations, capturing structural and semantic relationships that remain invariant under attacks [3, 8]. Adversarial resilient strategies which combine variational autoencoders (AEs) with MLP sharing feature extractors further revamp generalization to unseen malware [6] and these improvements point out to the limitations of unimodal or simply fused systems and highlight the importance of hybrid, multimodal frameworks for resilient malware detection.

Some existing multimodal approaches also struggle to effectively align diverse feature spaces and adaptively fuse cross-modal information under obfuscated and adversarial conditions [6, 9, 10] as single-modal detectors or simple fusion strategies also fail to capture the fine-grained semantic alignments and correlations necessary to maintain resilience when malware authors manipulate features or exploit modality-specific weaknesses.

Contrastive learning strengthens the model's capacity to discriminate between malware patterns and adversarial perturbed variants by using the agreement between semantically related samples across modalities while minimizing false correlations [10, 11]. As well, adaptive cross-modal fusion mechanisms dynamically weight and integrate modality-specific and cross-modal features to improve resistance against obfuscation-induced distortions, noise, and inconsistencies as this fusion may allow multimodal malware detectors to maintain high detection accuracy while offering improved adaptability and generalization against evolving evasion strategies.

This study offers an improved multimodal malware detection system that fills these limitations by making the following contributions:

- Fusing ResNet-50 for image features and MLP encoders for tabular information for strong multimodal representation that captures discriminative patterns.

- Dynamically integrating features, adaptive fusion and contrastive cross-modal learning by aligning multiple modalities to improve detection against obfuscation and adversarial attacks.

This study enhances malware detection and offers a practical defense against adversarial and obfuscation perturbed threats by merging multimodal representation, robust cross-modal learning, and adaptive fusion.

2. Related Works

Malware detection has shifted towards multimodal detection systems to counter new-evasion techniques such as adversarial and obfuscation perturbation attacks that attempt to deceive individual unimodal systems. Different feature modalities such as tabular features, sequence representations, and image encodings are fused into one multimodal system to exploit invariant, fine-grained patterns, similarity intents and generalisation capabilities to maximise detection performance. Image-based detection usually employs deep convolutional neural networks (CNNs) to automatically learn useful structural, statistical, semantic and visual patterns in malware samples or extracted features (denoted as pixel encoded greyscale byteplot images).

For example, ResNet-50 achieved good performance on multiple malware variants [12]. Novel combinations of CNNs with vision transformers have broken through known performance barriers to achieve high generalisation behaviour on diverse malware variants [13, 14]. These models show that good visual feature encoding provides a resilient baseline for detection.

Tabular features that approximate from static features like system call sequences, metadata or dynamic features such as API call frequencies can be used with popular machine learning models with bagging or boosting techniques, such as Light Gradient Boosting machines (LGBM) with good performance records [15]. Multimodal contrastive learning setups show that training in a cross-modal fashion can improve the feature alignment to produce robust feature vectors even when the data is sparse or corrupted [16]. Sequence features like package names or permissions can be transformed with time sequence models trained with RNNs or CNN-GRU hybrids to learn obfuscated dynamic changes in the sequence, outperforming traditional n-gram models that simply learn fixed combinations of time sequence states for malware detection – thus supporting deployment in less powerful execution environments [17].

Multimodal pre-training with sequence level contrastive learning can reinforce the cross-modal relations and achieve more effective fusion of tabular, image and sequence modalities [18]. Obfuscation and adversarial perturbed attack are still long-standing challenges for malware detection. Recent researches on hybrid models such as mixed between MLP and VAE show their effectiveness to resist against unknown adversarial perturbations by disentangling the latent features [6], adaptive cross-modal fusion mechanisms achieve resiliency while flexibly weighting features of devices to cancel out any noise or disturbance from one modality [19, 20]. For extracting at the best of CNN-based image analysis, sequence learning, like tabular feature modelling, full integration or fusion of image, sequence and tabular data remain considerable unsolved issues [21, 22], simple concatenation has limited effectiveness in considering cross-domain features and correlations. Typical studies ignore performance evaluation in simulated obfuscated then adversarial environments. Our work that uses the ResNet-50 and MLP encoder along with contrastive learning and adaptive fusion to align the features from image, sequence and tabular modes while dynamic weighting of features brings more effectiveness [18] to minimize noise or disturbance in any modality even in adversarial environments than other methods.

3. Methods

3.1. Dataset Description

The Multimodal CIC-MalDroid2020 dataset, a publicly accessible benchmark for Android malware detection in clean, hostile, and obfuscated scenarios, is used in this study. An Android program (APK) is represented by each sample, which is classified as either benign or malicious. Only two of the three complimentary modalities offered by the dataset are utilized:

Bytecode from APKs is translated into RGB images using image modality in order to maintain the code's spatial and structural patterns. These RGB images extracted from .DEX files are used to extract visual characteristics that are resistant to typical adversarial and obfuscation attack techniques. While the static features extracted from .APK files includes static components that provide ordered tabular data for each sample, such as permissions, API call, and metadata. These features capture intents and patterns of activity that are crucial for detecting malware. The dataset contains clean, obfuscated, and adversarial splits, each of which has both tabular and image modalities to give balanced representation for training, evaluation and testing.

3.2. Adversarial and Obfuscation Scenarios

Obfuscated samples, malware is transformed using common techniques such as packing, code reordering, junk code insertion, or encryption to alter surface representation but retain functionality. All subsets (clean, adversarial, and obfuscated) are mutually exclusive, and stratified sampling ensures consistent class distributions across training and testing sets which are all part of the secondary used. The details of the obfuscation and adversarial perturbations are not made available with the dataset source.

3.3. Data Pre-Processing

Image modality: All images are resized, augmented, and normalised during training using rotations, random crops, colour jitter, cutout masking and flips. For evaluation, images are resized and centre-cropped deterministically.

Tabular modality: Features are standardised through z-score normalisation to ensure consistent scaling:

$$X^{(\tilde{t})} = \frac{x^{(t)} - \mu}{\sigma}$$

where: $x^{(t)}$ = original tabular feature vector and μ, and σ = mean and standard deviation from training data respectively.

3.4. Problem Formulation

The detection task is formulated as supervised multimodal classification where each sample i is represented as:

$$x_i = (x_i^{(v)}, x_i^{(t)}),$$

where $x_i^{(v)}$: image input of the i^{th} sample (visual byteplot or grey-scale) and $x_i^{(t)}$: tabular feature vector of the i^{th} sample. The objective is to predict a label $y_i \in \{1,2,...,C\}$, where y_i: true class label of sample i, and C: total number of malware and benign classes.

- Image Encoder (f_v) → ResNet-50 produces visual embeddings:

$$z_i^{(v)} = f_v(x_i^{(v)}),$$

where $z_i^{(v)}$: latent visual embedding from the image encoder and $f_v(\cdot)$: ResNet-50 image encoder function.

- Tabular encoder (f_t) → *MLP* produces tabular embeddings:

$$z_i^{(t)} = f_t(x_i^{(t)}),$$

where $z_i^{(t)}$: latent embedding from tabular encoder, and $f_t(\cdot)$: MLP tabular encoder function.

- Cross-modal fusion (g) → Combines embeddings:

$$z_i = g(z_{i(v)}, z_{i(t)}),$$

where z_i: joint representation after fusing visual and tabular embeddings and $g(.)$: adaptive cross-modal fusion function.

- Classifier (h) → outputs class probabilities:

$$\hat{y}_i = h(z_i)$$

where $\hat{y}_i$: predicted class probabilities, and $h(\cdot)$: classification head (fully connected layer + softmax). The model is trained using a combination of cross-entropy classification loss and contrastive loss to align modalities:

$$L_{contrastive} = -\sum_{i=1}^{N} \log \left(\frac{\exp(z_i^{img} \cdot z_i^{tab} / T)}{\sum_{j=1}^{N} \exp(z_i^{img} \cdot z_j^{tab} / T)} \right)$$

where T: temperature parameter, N: batch size, z_i^{img}, z_i^{tab}: normalised embeddings for sample i, and $\cdot$: dot product for similarity.

3.5. Model Architecture

The proposed framework integrates image and tabular modalities as shown by the detailed model architecture and flow diagram in Figure 1:

- ResNet-50 extracts visual embeddings from byteplot or greyscale images.
- An MLP with masked autoencoder pre-training captures tabular feature correlations.
- Contrastive learning aligns modalities in a shared latent space.
- Adaptive fusion weights each modality according to its relevance per sample.
- The fused representation is passed to a fully connected classification head.

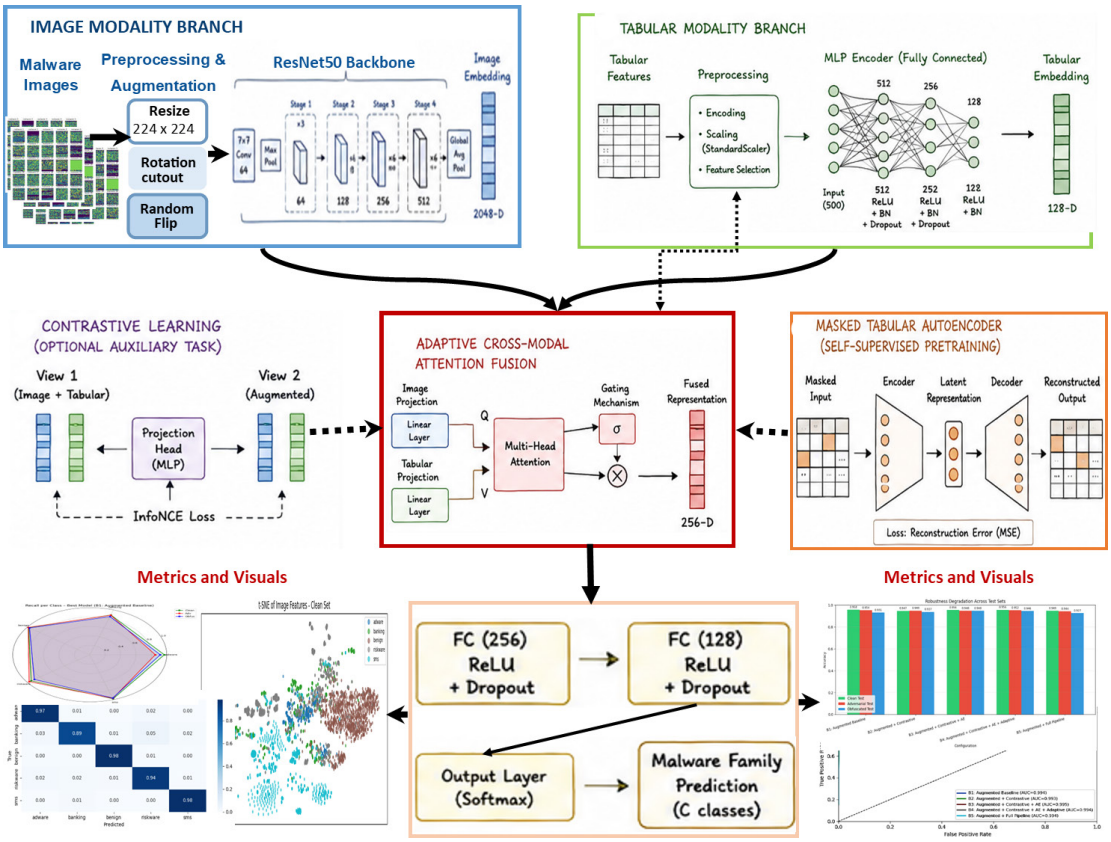

Figure 1. Detailed model architecture and flow Diagram.

3.6. Training and Implementation

The multimodal malware detection framework was implemented using the PyTorch deep learning library in the Kaggle Notebook optimized by using the GPU. The model integrates two modalities which are malware image representations and tabular behavioural features extracted from dynamic analysis. Image-based features were extracted using a pre-trained ResNet-50 network, where the final classification layer was finetuned to obtain a 2048-dimensional feature vector. The tabular modality was processed using a MLP consisting of three fully connected layers of sizes 512, 256, and 128, each followed by batch normalisation, ReLU activation, and dropout regularisation. To improve feature learning, a masked tabular autoencoder was first trained to reconstruct partially hidden tabular features using mean squared error loss. The learned encoder weights were then used to initialise the tabular encoder. Additionally, contrastive learning was applied by generating two augmented views of each image features and optimising a

contrastive loss with temperature τ = 0.1 to learn invariant representations and intents. During training, mix-up augmentation was used to combine and fuse samples from different classes. Model parameters were optimised using the AdamW optimiser with learning rates of 1×10^{-5} for the image encoder and 1×10^{-4} for other modules, with a cosine annealing learning rate scheduler. Training was performed with a batch size of 32 and early stopping with a patience of five epochs based on validation accuracy results.

3.7. Evaluation Metrics

Performance is measured using:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Macro-ROC-AUC} = \frac{1}{C} \sum_{c=1}^{C} \text{ROC-AUC}_c$$

where C is the total number of classes.

$$McNemar's\,Test = \chi^2 = \frac{\left(|b - c| - 1\right)^2}{b + c}$$

where:
b = misclassified by model 1 but correct in model 2
c = correct in model 1 but misclassified by model 2
Reject null hypothesis if $\chi^2 > 3.841$ for $\alpha = 0.05$
Metrics reported per-class and per-test configuration (clean, adversarial, and obfuscated).

4. Results

We evaluated five multimodal configurations (B1–B5) on clean, adversarial, and obfuscated test sets. The configurations

progressively integrate contrastive pre-training, a tabular masked autoencoder, adaptive cross-modal attention, and tabular noise injection during fine-tuning. All models were trained on an augmented dataset comprising clean, adversarial, and obfuscated samples under a stratified train-test-val split. The distribution of samples across training, validation, and test sets is presented in Table 1.

The augmented baseline (B1) has already demonstrated strong resilience, achieving 95.6% of accuracy on normal malware sample while maintaining a high performance on obfuscated samples of (93.6%) and adversarial samples of (94.6%). These findings suggest that incorporating the adversarial perturbed and obfuscated data during training offers a strong basis for distributional generalization and manipulation resistance. When the classification model is previously trained on a variety of augmented data, adding contrastive pre-training (B2) results in about comparable performance across all test sets, indicating that the contrastive objective does not offer a significant additional benefit.

The addition of the tabular masked autoencoder (B3) results in a significant improvement. The accuracy of detecting regular malware samples (clean) climbs to 96.2%, which is the greatest of all configurations; on obfuscated samples, the accuracy also rises to 94.5%, an absolute gain of over one percentage point over B2. However, when adaptive cross-modal attention (B4) is used instead of simple concatenation, the accuracy of obfuscated samples falls to 93.1%. Every test set exhibits the same decline. This decrease suggests that the acquired gating mechanism may lead to instability in the event of a distribution shift. The whole pipeline (B5), which incorporates tabular noise injection during fine-tuning, achieves the best detection accuracy of 94.8% on adversarial samples and 92.5% on obfuscated samples. This suggests a resilience trade-off

Table 1. Dataset split after leakage-free stratification.

Split	Clean	Adversarial	Obfuscated	Total
Training (augmented)	10,521	3872	3629	18,022
Validation (normal)	1169	-	-	1169
Test (normal)	5008	-	-	5008
Test (adversarial)	-	968	-	968
Test (obfuscated)	-	-	910	910

between systemic corruption and targeted perturbations, and the complete accuracy comparison is shown in Table 2.

The resilience degradation across clean, adversarial, and obfuscated test sets is visualised in Figure 2, which highlights that B3 exhibits the smallest performance drop under perturbation and manipulation, whereas B5 shows the largest degradation on obfuscated samples.

To assess whether the observed performance differences are statistically meaningful, McNemar's tests were conducted between consecutive configurations. No significant differences were observed between B1 and B2 ($p > 0.86$) which confirms that contrastive pre-training alone does not meaningfully alter predictions. The improvement from B2 to B3 is statistically significant on the clean test set ($p = 0.0042$), reinforcing the contribution of the masked autoencoder to improved representation learning. However, the addition of adaptive attention (B3 vs. B4) results in significant performance drops on both normal malware (clean) samples ($p = 0.0012$) and obfuscated data ($p = 0.0485$), as this confirms that this fusion strategy is detrimental in the present setting. It is clear that tabular noise injection does not result in a significant improvement because the differences between B4 and B5 are not statistically significant ($p > 0.18$). Using macro-average receiver operating characteristic (ROC) curves, the models' discriminative ability is further investigated. According to Figure 3, B5 slightly outperforms B3 on adversarial samples (0.992), but B3 obtains the best area under the curve (AUC) on typical malware (clean) samples (0.997) and obfuscated samples (0.993).

The distribution of resilience gains among malware categories is not consistent, according to a more thorough per-class analysis. On the basic malware (clean) test set, every setup consistently performs well in the Benign, Adware, Riskware, Banking, and SMS

Table 2. Accuracy comparison across configurations.

Configuration	Clean	Adversarial	Obfuscated
B1: Augmented Baseline	0.956	0.946	0.936
B2: + Contrastive	0.955	0.946	0.937
B3: + AE	0.962	0.947	0.945
B4: + Adaptive Attention	0.954	0.945	0.931
B5: Full Pipeline	0.951	0.948	0.925

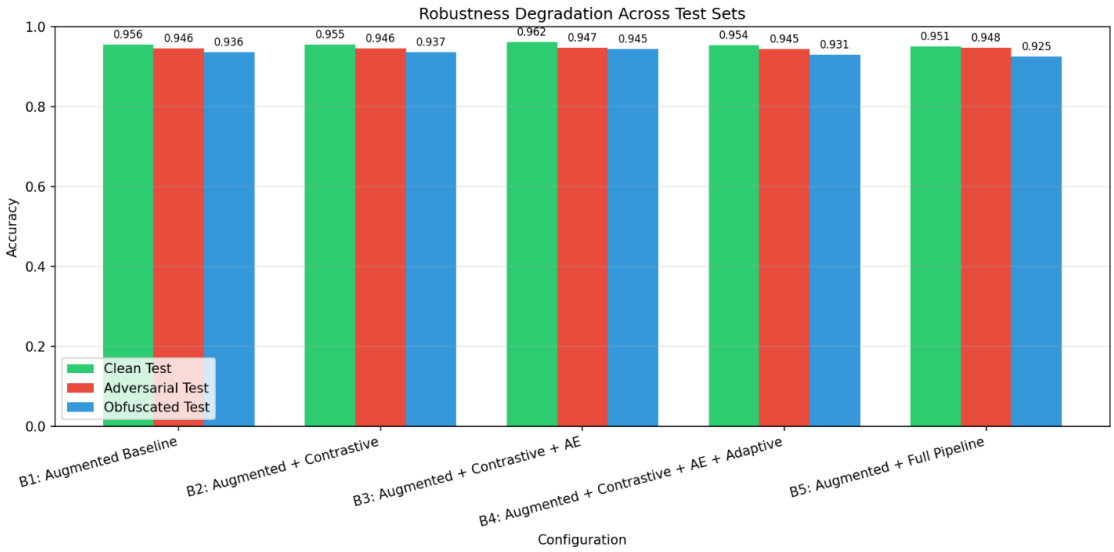

Figure 2. Resilience degradation across test sets.

classes. With a clear high recall for benign and SMS samples, B3 offers the most balanced F1 performance across categories. Table 3 shows each configuration's per-class F1 scores.

Riskware and banking are the most difficult classifications under adversarial settings, and B5 achieves the highest recall for these categories, which explains its slight overall adversarial sample advantage. On the other hand, adware turns into the most vulnerable class on the obfuscation sample. B3's improved obfuscated detection accuracy is explained by a significant improvement in adware recall when compared to other settings. The best adversarial model (B5) recall patterns across test sets are shown in the radar visualization in Figure 4, which highlights the model's advantages and disadvantages across classes.

Image features extracted by B5 were subjected to t-SNE in order to gain a better understanding of feature space behavior and intents. Normal malware (clean) samples form well-separated clusters that correspond to the five classes, as illustrated in Figure 5. On the other hand, adversarial and obfuscated samples show more overlap, especially across riskware and adware, which is consistent with the recall decreases found in the per-class analysis.

Figure 6 displays normalized confusion matrices for B5. Misclassifications are rare and mostly limited to semantically comparable classes on typical malware (clean) data. Confusion develops

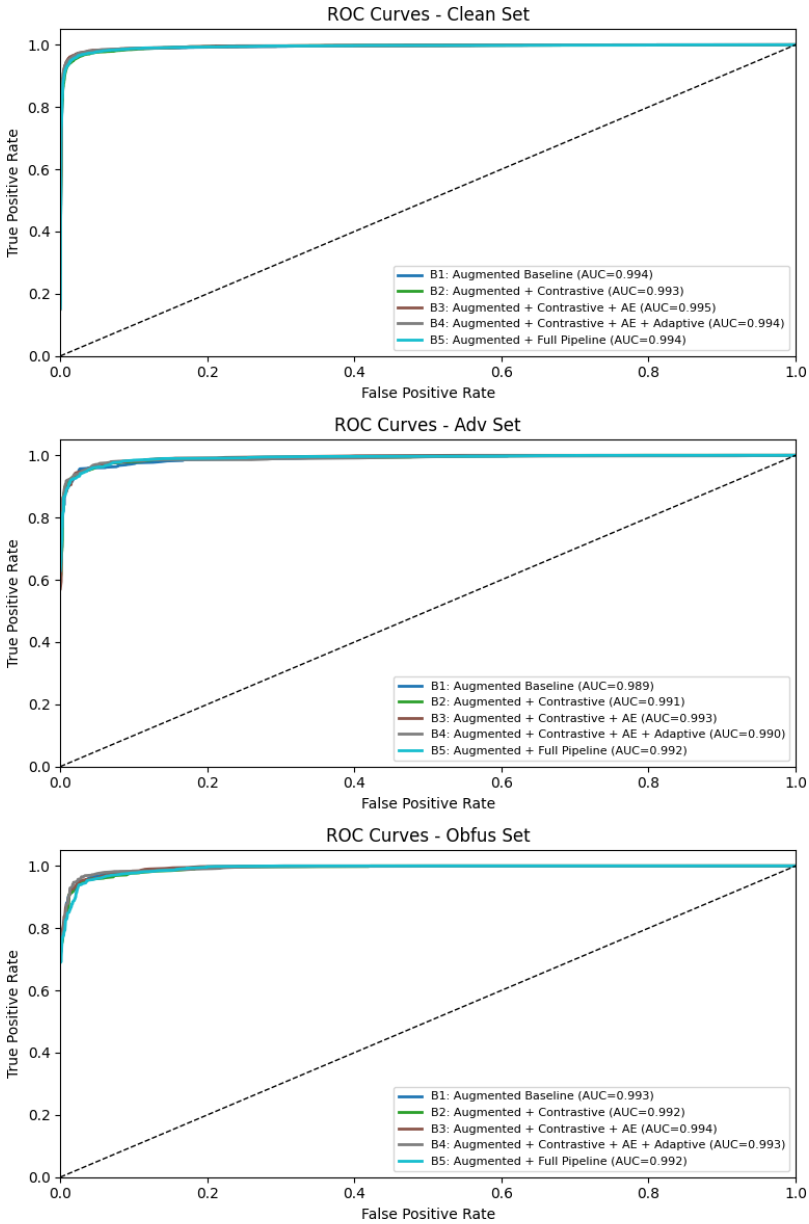

Figure 3. Macro-average ROC curves for all test sets.

under adversarial and obfuscated situations, especially when riskware and adware are involved, demonstrating the susceptibility of certain classes to disruptions.

Figures 7–10 give confusion matrices for each configuration (B1–B5) across all test sets for completeness, showing how misclassification patterns or intents changed when new components were added.

Table 3. Per-class F1 scores across configurations and test sets.

Config	Test set	Adware	Banking	Benign	Riskware	SMS
B1	Clean	0.920	0.916	0.968	0.944	0.985
	Adv	0.880	0.902	0.992	0.920	0.967
	Obfus	0.831	0.884	0.983	0.903	0.969
B2	Clean	0.928	0.912	0.973	0.940	0.983
	Adv	0.876	0.895	0.992	0.919	0.974
	Obfus	0.810	0.907	0.987	0.890	0.977
B3	Clean	0.933	0.925	0.980	0.947	0.986
	Adv	0.859	0.886	0.994	0.928	0.978
	Obfus	0.854	0.903	0.985	0.907	0.981
B4	Clean	0.909	0.915	0.974	0.939	0.984
	Adv	0.862	0.884	0.990	0.928	0.976
	Obfus	0.850	0.878	0.972	0.885	0.973
B5	Clean	0.916	0.905	0.970	0.936	0.981
	Adv	0.875	0.917	0.992	0.912	0.976
	Obfus	0.821	0.872	0.972	0.879	0.968

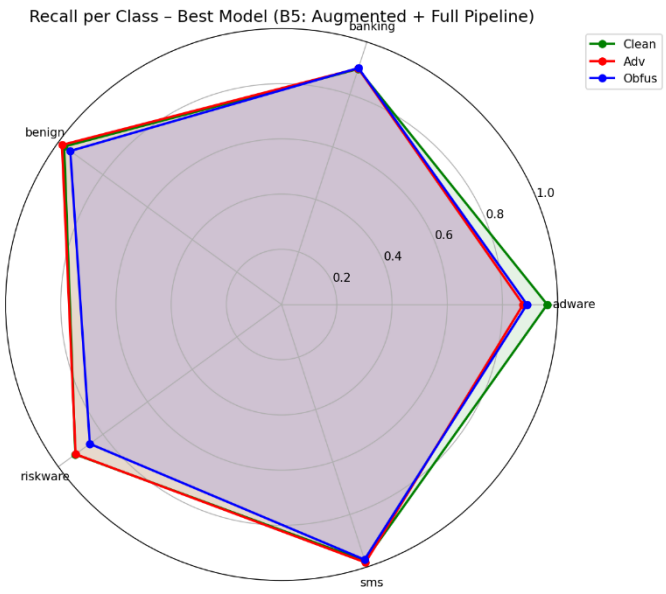

Figure 4. Radar chart of recall per class for B5.

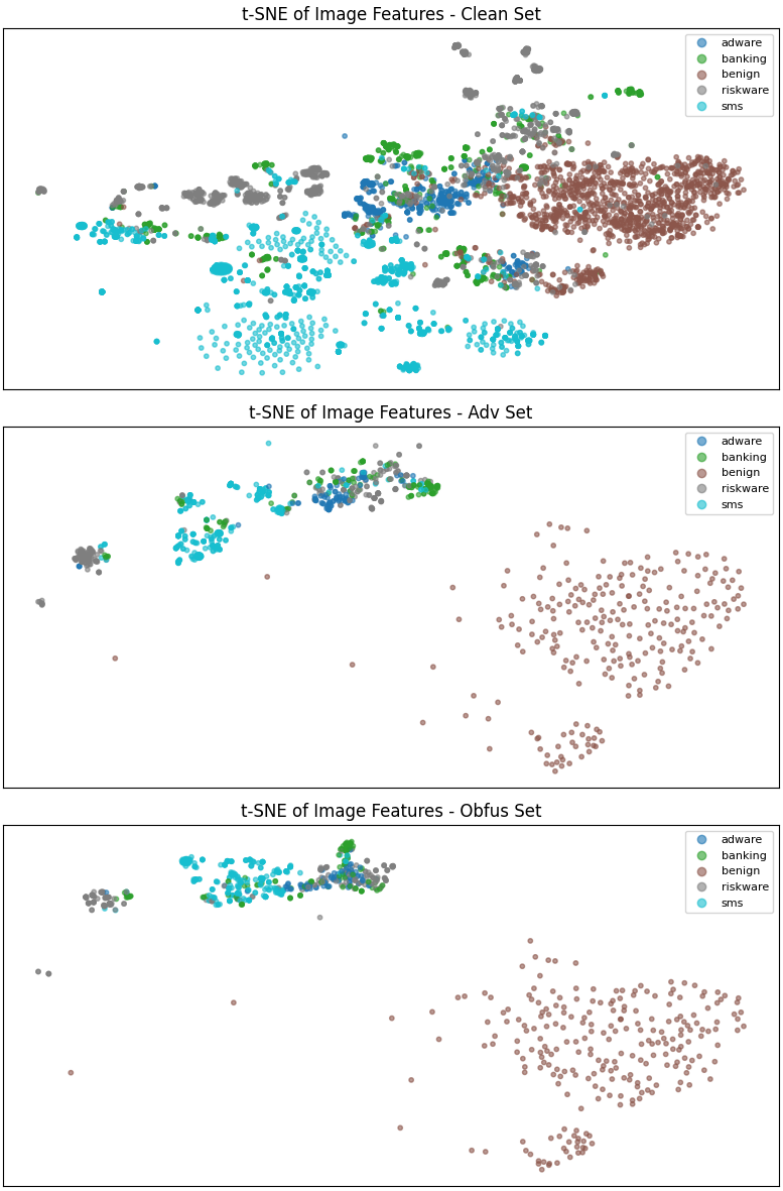

Figure 5. t-SNE visualisation of feature representations.

Ablation research (B1–B5) was carried out to separate the contribution of each architectural element. Contrastive pre-training by itself produces very little change in the presence of more training data, indicating its modest added advantage. The masked autoencoder, which significantly improves obfuscation resilience and clean performance, is the most beneficial and important component.

Conversely, adaptive attention results in statistically significant degradation, most likely due to unstable modality weighting during distribution shift while by increasing hidden manipulations and providing a minor adversarial advantage, tabular noise injection

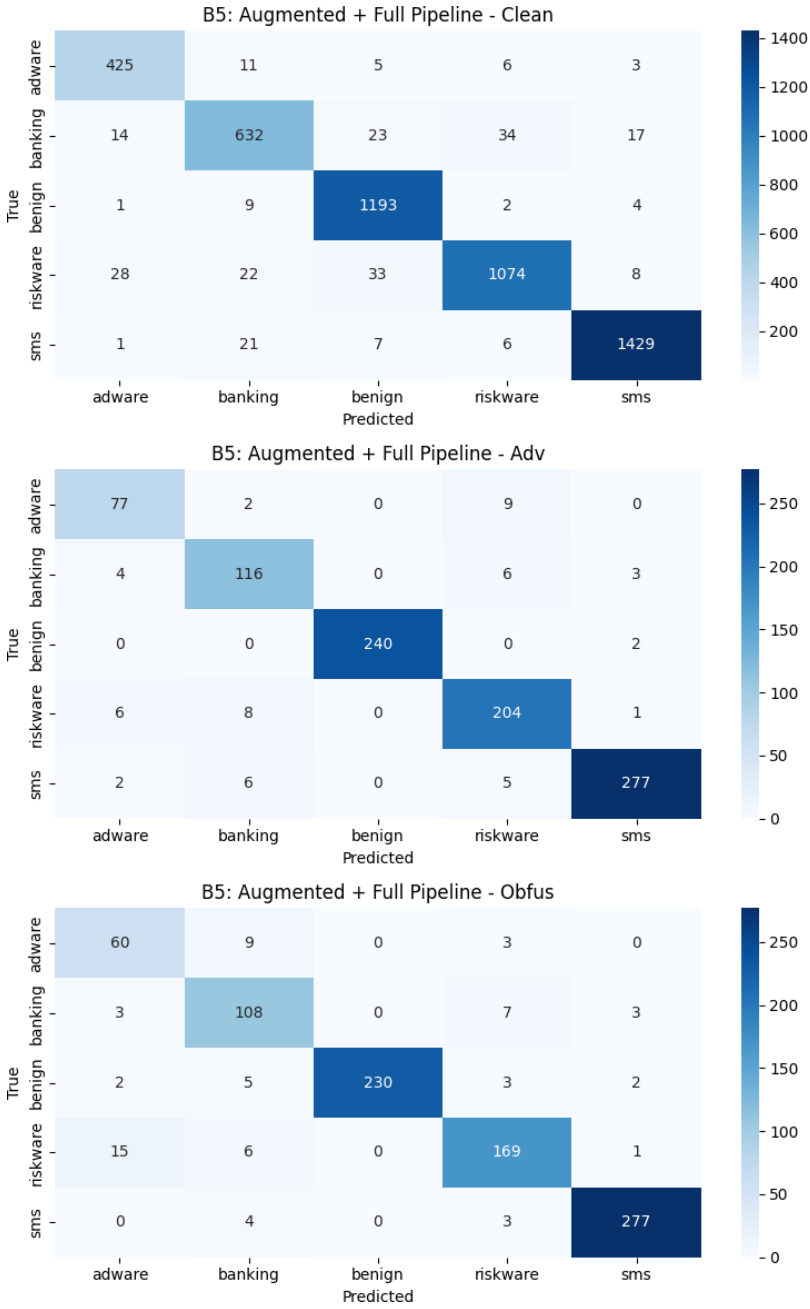

Figure 6. Normalised confusion matrices for B5.

supports the observed resilience performance. Recall per class-wise behavior across configurations are shown in Figures 11–13.

These results establish that B3 (augmented baseline + contrastive + masked autoencoder) is the most resilient and balanced

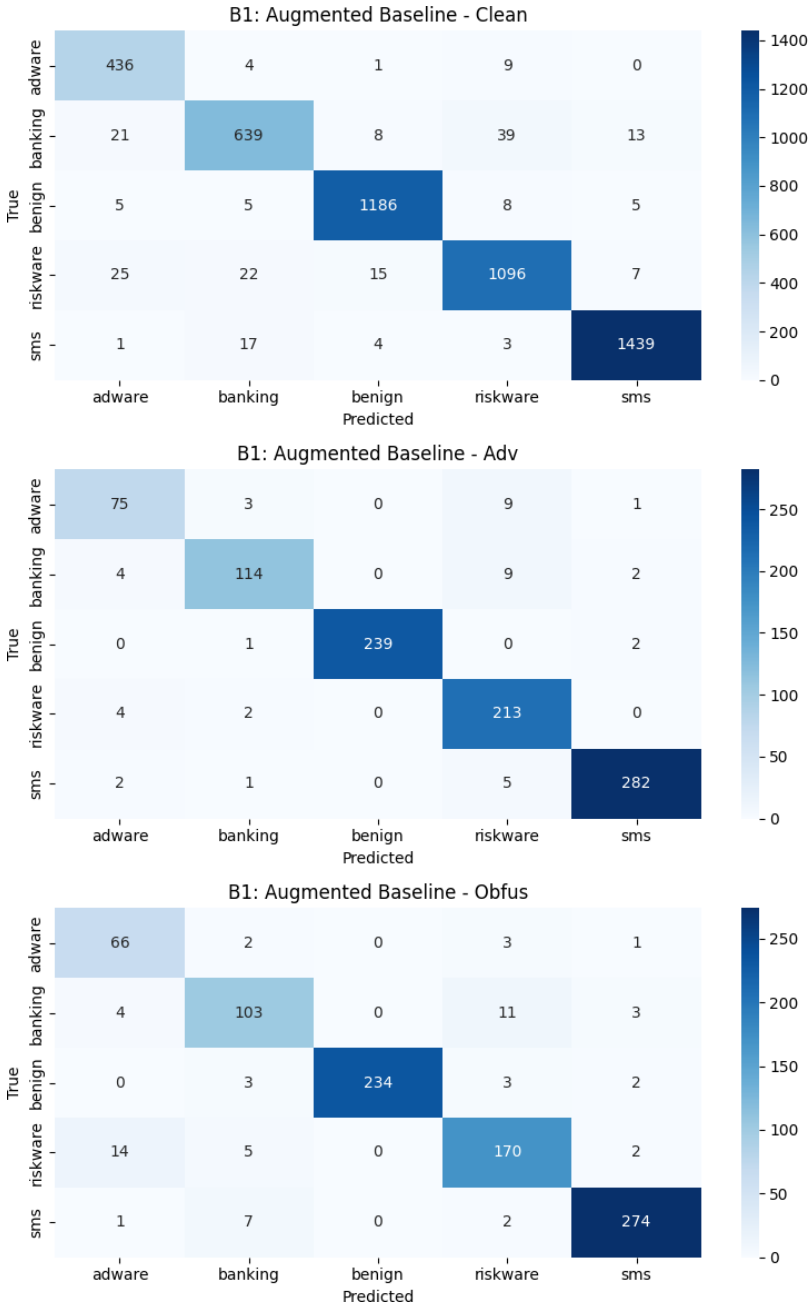

Figure 7. Confusion matrices for B1.

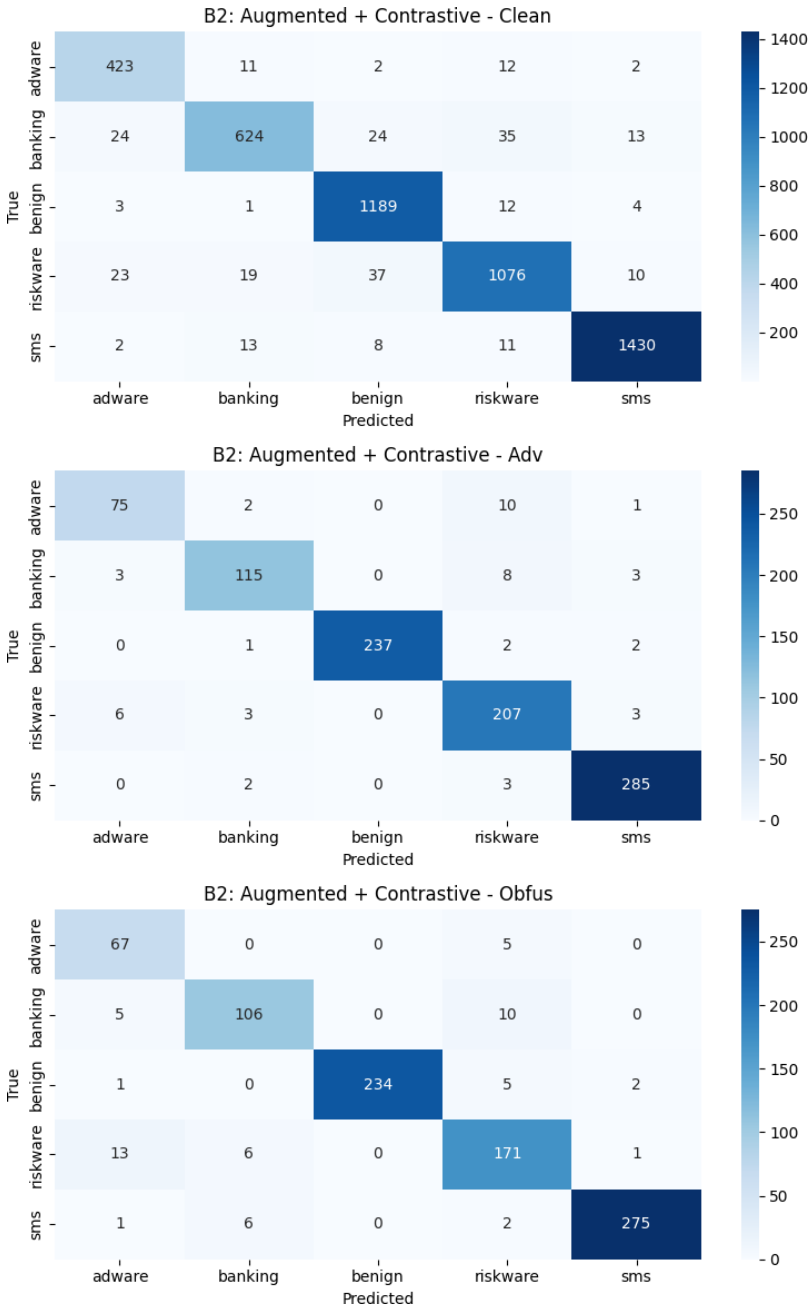

Figure 8. Confusion matrices for B2.

configuration, whereas B5 offers a minor advantage against adversarial perturbations at the expense of the obfuscation sample's resistance. These findings prove that resilience is not a monolith as architectural choices that strengthen resilient to one type of perturbation may concurrently weaken resistance to another.

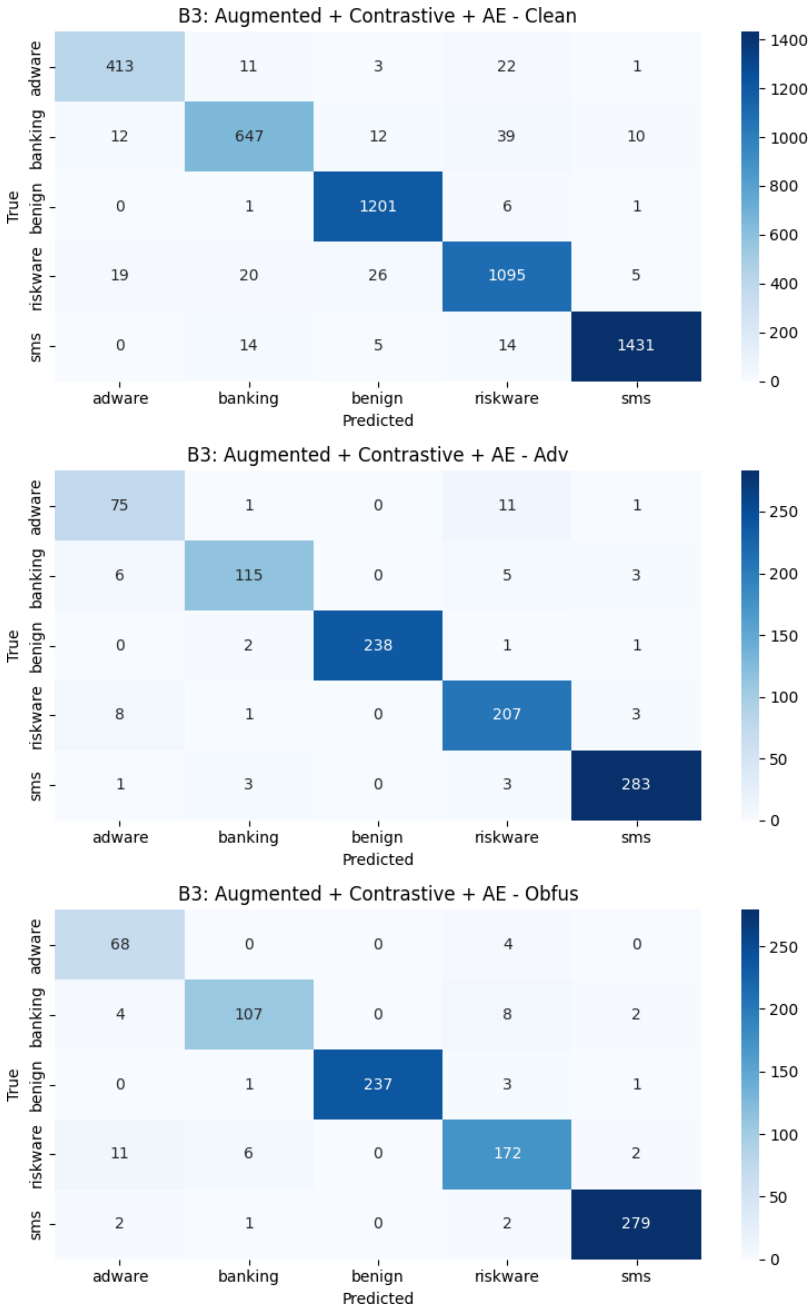

Figure 9. Confusion matrices for B3.

5. Conclusions

This study investigated the longevity of a multimodal Android malware detection system in normal, adversarial, and obfuscated environments. A comprehensive study of five

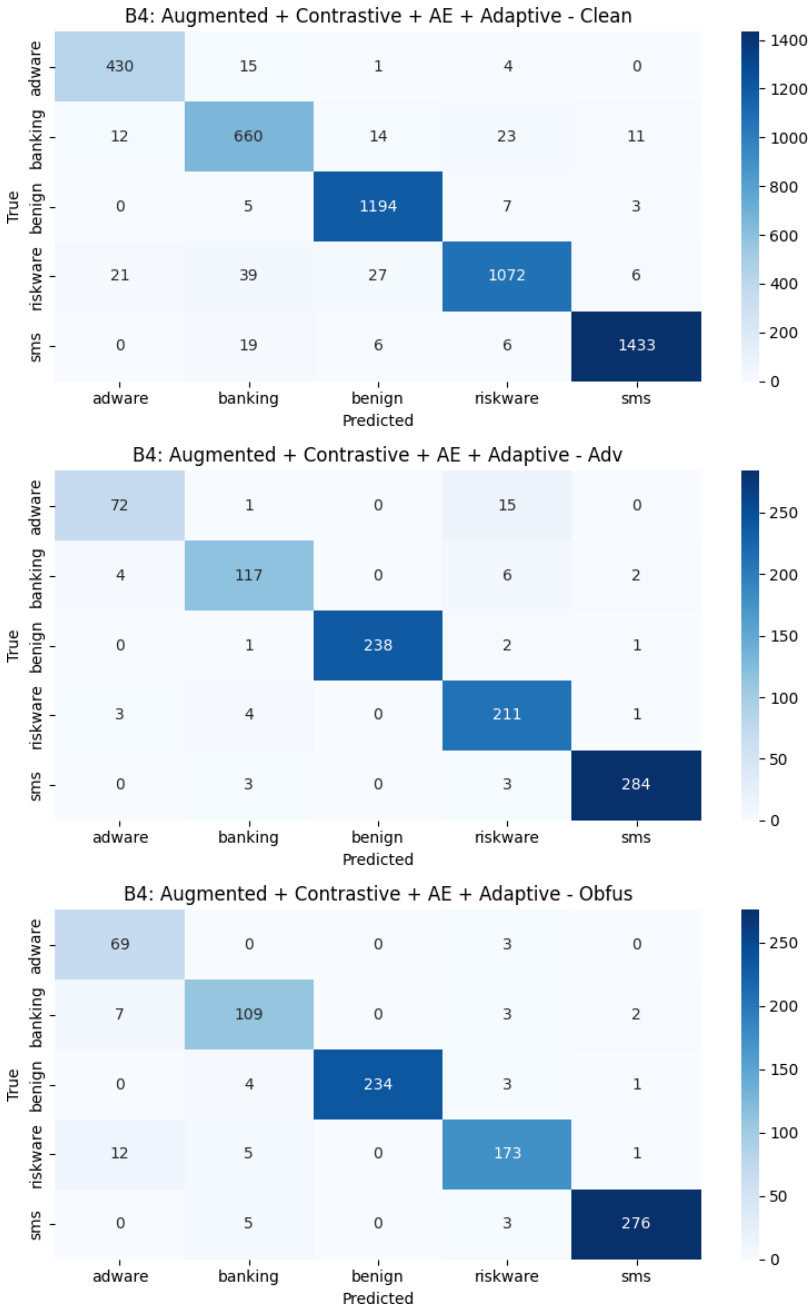

Figure 10. Confusion matrices for B4.

configurations revealed that resilience is more dependent on representation quality than architectural complexity. Training with enhanced data established a solid foundation for all test distributions. However, tabular masked autoencoder pre-training showed

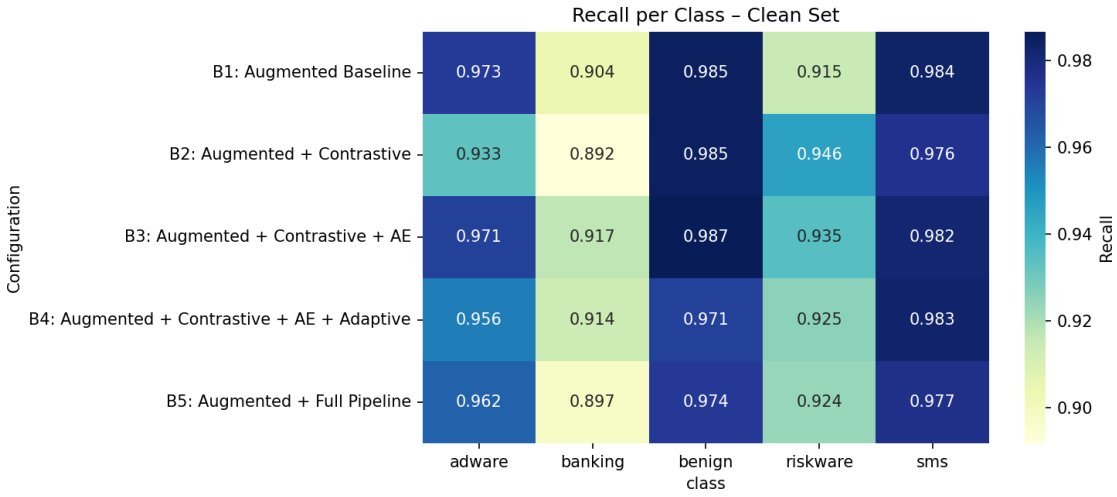

Figure 11. Recall per class on the clean set.

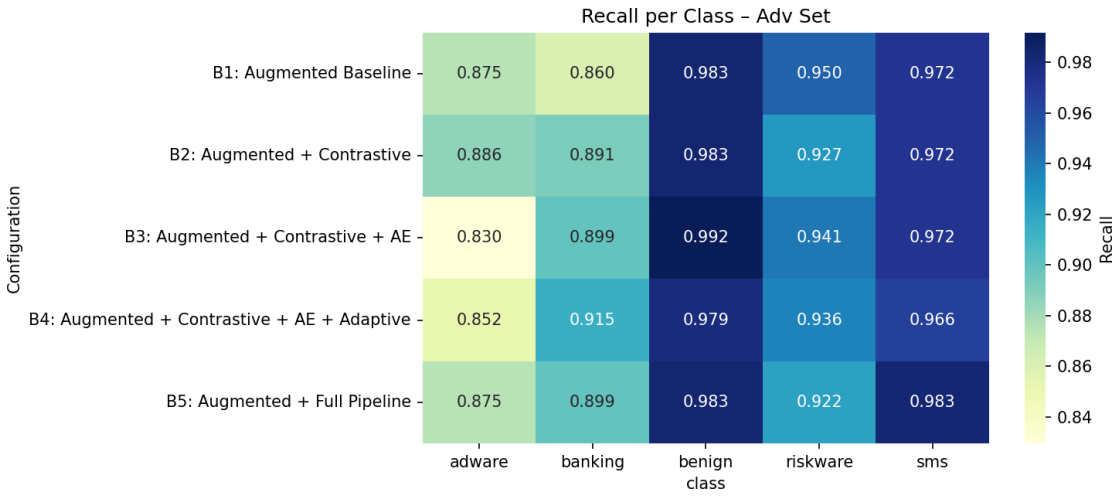

Figure 12. Recall per class on the adversarial set.

the greatest improvement in clean and obfuscated sample detection performance. In contrast, adaptive cross-modal attention decreased performance during distribution shift, demonstrating that greater fusion complexity does not always improve resilience. Tabular noise injection provided a modest advantage against adversarial perturbations while decreasing susceptibility to obfuscation samples, indicating a trade-off between modification types.

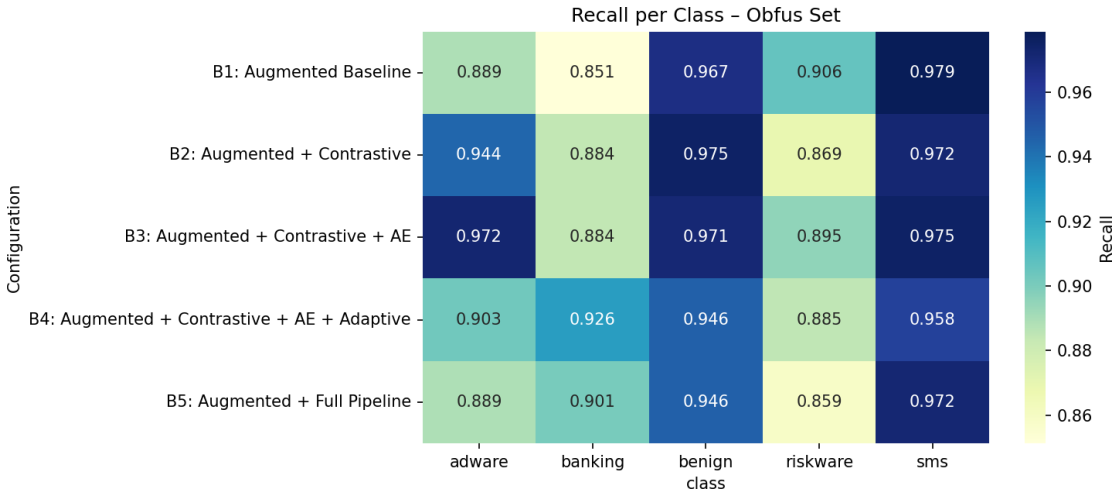

Figure 13. Recall per class on the obfuscated set.

This combination of enhanced training and masked autoencoder pre-training resulted in the most balanced and consistent performance. These findings highlight the importance of rigorous module selection and assessment across multiple threat scenarios for resilient multimodal malware detection, rather than relying solely on clean sample detection accuracy. To improve detection performance and robustness even further, future research should look into supervised contrastive aims, more reliable cross-modal fusion processes, and adaptive noise scheduling techniques. Long-term generalization under adversarial drift can be better understood by comparing different datasets and malware streams.

References

[1] K. Kancherla, S. Mukkamala, "Image visualization based malware detection," in *IEEE INFOCOM 2019 – IEEE Conference on Computer Communications,* 2013, pp. 40–44, doi: 10.1109/cicybs.2013.6597204.

[2] S. O'Shaughnessy, S. Sheridan, "Image-based malware classification hybrid framework based on space-filling curves," *Computers & Security*, vol. 116, Art. no. 102660, 2022, doi: 10.1016/j.cose.2022.102660.

[3] T.S. John, T. Thomas, S. Emmanuel, "Graph convolutional networks for Android malware detection with system call graphs," in *IEEE INFOCOM 2020 – IEEE Conference on Computer Communications,* 2020, pp. 162–170, doi: 10.1109/isea-isap49340.2020.235015.

[4] K. Xu, Y. Li, R.H. Deng, K. Chen, "DeepRefiner: Multi-layer Android malware detection system applying deep neural networks," in *IEEE INFOCOM 2018 – IEEE Conference on Computer Communications.* London: Institute of Electrical Electronics Engineers (IEEE), 2018, pp. 473–487, doi: 10.1109/eurosp.2018.00040.

[5] S. Aurangzeb, M. Aleem, "Evaluation and classification of obfuscated Android malware through deep learning using ensemble voting mechanism," *Science Reporter*, vol. 13, no. 1, 2023, doi: 10.1038/s41598-023-30028-w.

[6] H. Li, S. Zhou, W. Yuan, X. Luo, C. Gao, S. Chen, "Resilient Android malware detection against adversarial example attacks," in *WWW '21: Proceedings of the Web Conference 2021*, 2021, pp. 3603–3612, doi: 10.1145/3442381.3450044.

[7] D. Park, H. Khan, B. Yener, "Generation & evaluation of adversarial examples for malware obfuscation," in *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2019, pp. 1283–1290, doi: 10.1109/ICMLA.2019.00210.

[8] X. Ling, L. Wu, W. Deng, Z. Qu, J. Zhang, et al., "MalGraph: Hierarchical graph neural networks for resilient Windows malware detection," in *IEEE INFOCOM 2022 – IEEE Conference on Computer Communications,* 2022, pp. 1998–2007, doi: 10.1109/infocom48880.2022.9796786.

[9] M. Shahpasand, L. Hamey, D. Vatsalan, M. Xue, "Adversarial attacks on mobile malware detection," in *IEEE INFOCOM 2019 – IEEE Conference on Computer Communications,* 2019, pp. 17–20, doi: 10.1109/ai4mobile.2019.8672711.

[10] L. Wang, C. Zhang, H. Xu, Y. Xu, X. Xu, S. Wang, "Cross-modal contrastive learning for multimodal fake news detection," in *WWW '23: Proceedings of the Web Conference 2023*, 2023, pp. 5696–5704, doi: 10.1145/3581783.3613850.

[11] H. Peng, X. Gu, J. Li, Z. Wang, H. Xu, "Text-centric multimodal contrastive learning for sentiment analysis," *Electronics*, vol. 13, no. 6, Art. no. 1149, 2024, doi: 10.3390/electronics13061149.

[12] E. Rezende, G. Ruppert, T. Carvalho, F. Ramos, P. De Geus, "Malicious software classification using transfer learning of ResNet-50 deep neural network," in *IEEE INFOCOM 2017 – IEEE Conference on Computer Communications,* 2017, pp. 1011–1014, doi: 10.1109/icmla.2017.00-19.

[13] M. Ashawa, N. Owoh, S. Hosseinzadeh, J. Osamor, "Enhanced image-based malware classification using transformer-based convolutional neural networks (CNNs)," *Electronics*, vol. 13, no. 20, Art. no. 4081, 2024, doi: 10.3390/electronics13204081.

[14] W. El-Shafai, I. Almomani, A. Alkhayer, "Visualized malware multi-classification framework using fine-tuned CNN-based transfer learning models," *Applied Sciences*, vol. 11, no. 14, Art. no. 6446, 2021, doi: 10.3390/app11146446.

[15] R. Hasan, B. Biswas, Md. Samiun, M.A. Saleh, M. Prabha, et al., "Enhancing malware detection with feature selection and scaling techniques using machine learning models," *Science Reporter*, vol. 15, no. 1, Art. no. 9122, 2025, doi: 10.1038/s41598-025-93447-x.

[16] P. Hager, M.J. Menten, D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," in *IEEE INFOCOM 2023 – IEEE Conference on Computer Communications,* 2023, pp. 23924–23935, doi: 10.1109/cvpr52729.2023.02291.

[17] W.Y. Lee, J. Saxe, R. Harang, "SeqDroid: Obfuscated Android malware detection using stacked convolutional and recurrent neural networks," in *Deep Learning Applications for Cyber Security. Advanced Sciences and Technologies for Security Applications*, M. Alazab and M. Tang, Eds. Cham: Springer, 2019, pp. 197–210, doi: 10.1007/978-3-030-13057-2_9.

[18] L. Zhang, X. Zhou, Z. Zeng, Z. Shen, "Multimodal pre-training for sequential recommendation via contrastive learning," *ACM Transactions on Recommender Systems*, vol. 3, no. 1, pp. 1–23, 2024, doi: 10.1145/3682075.

[19] P. Prajapati, M. Stamp, "An empirical analysis of image-based learning techniques for malware classification," in *Malware Analysis Using Artificial Intelligence and Deep Learning*, M. Stamp, M. Alazab, and A. Shalaginov, Eds. Cham: Springer, 2021, pp. 411–435, doi: 10.1007/978-3-030-62582-5_16.

[20] X. Li, K. Qiu, C. Qian, G. Zhao, "An adversarial machine learning method based on OpCode N-grams feature in malware detection," in *IEEE INFOCOM 2020 – IEEE Conference on Computer Communications,* 2020, pp. 380–387, doi: 10.1109/dsc50466.2020.00066.

[21] T. Bilot, N. El Madhoun, K. Al Agha, A. Zouaoui, "A survey on malware detection with graph representation learning," *ACM Computing Surveys*, vol. 56, no. 11, pp. 1–36, 2024, doi: 10.1145/3664649.

[22] S. Han, H. Yun, Y. Park, "Deep learning for cybersecurity classification: Utilizing depth-wise CNN and attention mechanism on VM-obfuscated data," *Electronics*, vol. 13, no. 17, Art. no. 3393, 2024, doi: 10.3390/electronics13173393.