Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077697

ARTICLE

MSA-ViT: A Multi-Scale Vision Transformer for Robust Malware Image Classification

Bofan Yang, Bingbing Li and Chuanping Hu*

Key Laboratory of Cyberspace Security, School of Cyber Science and Engineering, Zhengzhou University, Ministry of Education, Zhengzhou, China

*Corresponding Author: Chuanping Hu. Email: cphu@zzu.edu.cn

Received: 15 December 2025; Accepted: 02 February 2026; Published: 09 April 2026

ABSTRACT: The rapid evolution of malware obfuscation and packing techniques significantly undermines the effectiveness of traditional static detection approaches. Transforming malware binaries into grayscale or RGB images enables learning-based classification, yet existing CNN- and ViT-based models depend heavily on fixed-resolution inputs and exhibit poor robustness under cross-resolution distortions. This study proposes a lightweight and sample-adaptive Multi-Scale Vision Transformer (MSA-ViT) for efficient and robust malware image classification. MSA-ViT leverages a fixed set of input scales and integrates them using a Scale-Attention Fusion (SAF) module, where the largest-scale CLS token serves as the query to dynamically aggregate cross-scale representations. To mitigate scale bias and improve generalization, SimCLR self-supervised pre-training and KL-divergence-based cross-scale consistency regularization are incorporated. Experiments on the Malimg and MaleVis datasets demonstrate that MSA-ViT achieves accuracies of 98.5% and 96.0%, respectively, outperforming existing baselines. Robustness evaluations further show that performance degradation remains below 1.8% under scaling, padding, and FGSM perturbations. Attention-based visualizations confirm the interpretability of the fusion mechanism. Overall, MSA-ViT provides an accurate, robust, and computationally efficient solution for image-based malware classification.

KEYWORDS: Malware classification; vision transformers; multi-scale fusion; robustness; self-supervised learning

1 Introduction

With the rapid escalation of cyber threats, malware remains a critical challenge to modern information security. Recent reports indicate that over 450,000 new malware samples emerge daily, with more than 1.2 billion variants accumulated worldwide. This growth is accelerated by AI-assisted metamorphism, obfuscation, and cross-platform propagation, which substantially increases detection difficulty. Conventional static analysis methods, including signature matching and heuristic rules, are computationally efficient but vulnerable to evasion by encrypted, packed, or polymorphic malware [1,2]. Consequently, their effectiveness degrades in continuously evolving threat environments [3,4].

Image-based malware analysis has been explored to mitigate these limitations. By transforming binaries into grayscale or texture images, learning-based models can capture discriminative spatial patterns [4,5], achieving strong performance (e.g., near 97% accuracy on Malimg) [6]. Nevertheless, CNNs primarily model local receptive fields and may fail to capture global structure and multi-scale dependencies (e.g., opcode-level textures vs. section-level layouts) [7,8]. Prior studies also report notable performance drops in highly variable settings, especially under resource constraints [2,9].

Vision Transformers (ViTs) leverage self-attention to model long-range dependencies [10], motivating Transformer-based malware detection [11,12]. For example, SHERLOCK [11] adopts self-supervised ViTs for Android malware representation, while ViTDroid [12] improves interpretability through attention visualization. Recent ViT-based methods (2023–2025) further demonstrate consistent gains over CNN baselines [1,13–17]. However, standard ViTs typically employ single-scale patch embeddings and fixed-resolution inputs, which limits robustness to scale variations and distortions common in malware images [10,12].

To reduce scale sensitivity, multi-scale Transformer architectures such as Swin Transformer [18] and MViT [19] adopt hierarchical or parallel designs. While effective, these approaches often introduce non-trivial computational overhead and architectural complexity [20,21]. Existing multi-scale fusion methods for malware detection usually combine features via static concatenation or uniform pooling [2,22], which ignores sample-specific scale relevance and can amplify scale bias, leading to unstable performance in practice [1,23]. Lightweight variants (e.g., LeViT-MC) improve efficiency but may still be costly for mobile or embedded deployment [2,13].

To address these challenges, we propose a lightweight and sample-adaptive **Multi-Scale Attention Vision Transformer (MSA-ViT)** for robust malware image classification. Our contributions are:

1. We design a **Scale-Attention Fusion (SAF)** module that adaptively aggregates multi-scale features using the largest-scale CLS token as the query, enabling dynamic cross-scale interaction without altering the internal ViT backbone (3.1 GFLOPs).
2. We introduce **KL-divergence-based cross-scale consistency regularization** to reduce dependence on any single scale and improve robustness to scaling, padding, and noise.
3. We employ a **dual-head, staged training scheme** that combines SimCLR self-supervised pre-training with supervised fine-tuning to improve generalization, particularly with limited labeled data.

Experiments on Malimg [4] and MaleVis [24] show that MSA-ViT outperforms ViT-B/16, Swin-T, and MAFormer [25] by 1.0%–2.5% across accuracy, F1, and MCC. Robustness tests further report less than 1.8% degradation under scaling, padding, and adversarial perturbations. Attention visualizations also reveal interpretable scale-dependent patterns, indicating effective adaptive fusion.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes MSA-ViT. Section 4 presents experimental results. Section 5 concludes the paper and discusses future directions.

2 Related Work

Transformer-based models have recently gained significant attention in both computer vision and cybersecurity. This section reviews research most relevant to this work, including Vision Transformers and their limitations, multiscale Vision Transformer architectures, deep learning-based malware detection, and attention-based fusion mechanisms.

2.1 Vision Transformer and Its Limitations

The Vision Transformer (ViT) [10] pioneered the application of Transformer architectures to computer vision by modeling images as sequences of fixed-size patch tokens processed via multi-head self-attention [26]. On large-scale datasets, ViT achieves performance comparable to or exceeding that of convolutional neural networks (CNNs). However, its reliance on a single input resolution and fixed patch size constrains its ability to simultaneously capture fine-grained local details and coarse global structures, particularly under pronounced scale variations.

Despite this limitation, ViT's global attention mechanism enables effective modeling of long-range dependencies and provides interpretable attention maps [27]. Moreover, self-supervised frameworks such as Masked Autoencoders (MAE) [28] mitigate data dependency by learning representations through masked reconstruction. These advances motivate the adoption of Transformer-based and self-supervised approaches in malware image analysis, where labeled data are often scarce and visual patterns exhibit high diversity.

2.2 Multiscale Vision Transformer Methods

To address scale sensitivity, several multiscale Vision Transformer variants have been proposed. Swin Transformer [18] employs a hierarchical design with shifted-window attention to progressively aggregate local and global information. CrossViT [20] introduces parallel branches with different patch sizes and cross-scale attention modules. MViT [19] and its extension MViTv2 [21] gradually trade spatial resolution for channel capacity to balance efficiency and representation power, while Transformer-in-Transformer [29] adopts nested attention for implicit multiscale modeling.

Although effective, these methods often require substantial architectural changes or incur increased computational costs, which limits their applicability in resource-constrained settings. This motivates the development of lightweight multiscale fusion strategies that preserve standard ViT backbones while improving scale robustness.

2.3 Deep Learning Methods for Malware Detection

Deep learning techniques have become increasingly important for malware detection due to the limitations of handcrafted features under obfuscation and polymorphism. Image-based approaches convert binary executables into grayscale or texture images and apply CNNs to learn spatial representations, achieving strong performance on benchmarks such as Malimg and MaleVis. However, CNNs are inherently limited by local receptive fields and struggle to jointly model global dependencies and multi-scale structures, resulting in reduced robustness under high intra-class variability and dataset shifts. It should be noted that image-based malware detection remains a form of static analysis and is vulnerable to obfuscation techniques that disrupt visual textures while preserving functionality, as discussed in recent studies on adversarial selective obfuscation [30].

2.4 Applications of Vision Transformers in Malware Detection

Vision Transformers provide a promising alternative for malware detection by enabling global dependency modeling through self-attention. SHERLOCK [11] employs self-supervised ViTs for Android malware representation learning and demonstrates strong performance in unlabeled settings. LeViT-MC [13] combines a lightweight ViT with DenseNet to achieve efficient inference on the MaleVis dataset, while Ashawa et al. [22] integrate ResNet and ViT for improved image-based malware classification.

Interpretability has also received increasing attention. ViTDroid [12] visualizes attention maps to identify critical opcode regions, and other studies combine Transformer models with multimodal visual representations to enhance explainability [23]. Despite these advances, existing ViT-based methods often suffer from high computational overhead and limited robustness to scale variations and input distortions. Representative multiscale Transformer-based models are summarized in Table 1.

Table 1: Comparison of representative multiscale Transformer-based models in terms of design and robustness.

Method	Scales	Fusion Type	FLOPs (G)	Malware Robustness
Swin-T	4	Hierarchical	4.5	Moderate
CrossViT	2	Cross-attention	4.8	Moderate
MAFormer	3	Local-Global	3.9	Moderate
MSA-ViT (Ours)	3	Scale-Attention	3.1	High

2.5 Attention Fusion Mechanisms

Attention-based fusion has been explored as an alternative to redesigning Transformer backbones. MAFormer [25], for example, introduces a Multi-scale Attention Fusion module that integrates local window attention with global context within each block, enhancing single-scale representations. While effective, such methods typically operate within a single-resolution feature space.

In contrast, applications such as malware image analysis inherently involve multi-resolution inputs and cross-scale dependencies. Our work follows the attention-based fusion paradigm but focuses on lightweight cross-scale aggregation at the backbone output level, aiming to improve robustness and efficiency while maintaining compatibility with standard ViT architectures.

3 Methodology

3.1 Overall Architecture

We propose a **Multi-Scale Attention Vision Transformer (MSA-ViT)**, whose overall architecture is illustrated in Fig. 1. Given an input malware image, multiple resolution variants are first generated to capture complementary semantic information at different scales. Each scale is processed by an independent Vision Transformer (ViT) backbone with separate parameters, producing a set of scale-specific global representations. To efficiently integrate these representations, we introduce the Scale-Attention Fusion (SAF) module at the backbone output stage, as illustrated in Fig. 2. In particular, the CLS token from the largest-scale branch is used as the query, while CLS tokens from all scales serve as keys and values, enabling adaptive cross-scale aggregation into a single fused token.

The fused representation is task-flexible: it can be fed into a classification head for supervised learning or into a projection head for self-supervised contrastive learning. This design allows MSA-ViT to improve generalization and robustness without modifying the internal structure of standard ViT backbones. The multi-scale ViT backbones are instantiated as independent networks without weight sharing, allowing each branch to learn scale-specific representations.

3.2 Data Preprocessing and Augmentation

We employ pre-generated malware image representations from the Malimg and MaleVis datasets to capture opcode textures and structural patterns efficiently. Malimg grayscale images are replicated across three channels to satisfy ViT input requirements, while MaleVis samples retain their native RGB format. All images are normalized to the $[0, 1]$ range. To mitigate dataset imbalance and redundancy, correlation-based feature selection (CFS) [2] is applied to reduce redundant pixel-level correlations and improve generalization.

For multi-scale modeling [18–21], each image is resized to a fixed set of resolutions $\{S_1, \ldots, S_n\}$ using bilinear interpolation, producing aligned multi-resolution views that preserve cross-scale semantic

consistency. During training, scale-specific random transformations are applied, while deterministic center cropping is used at test time.

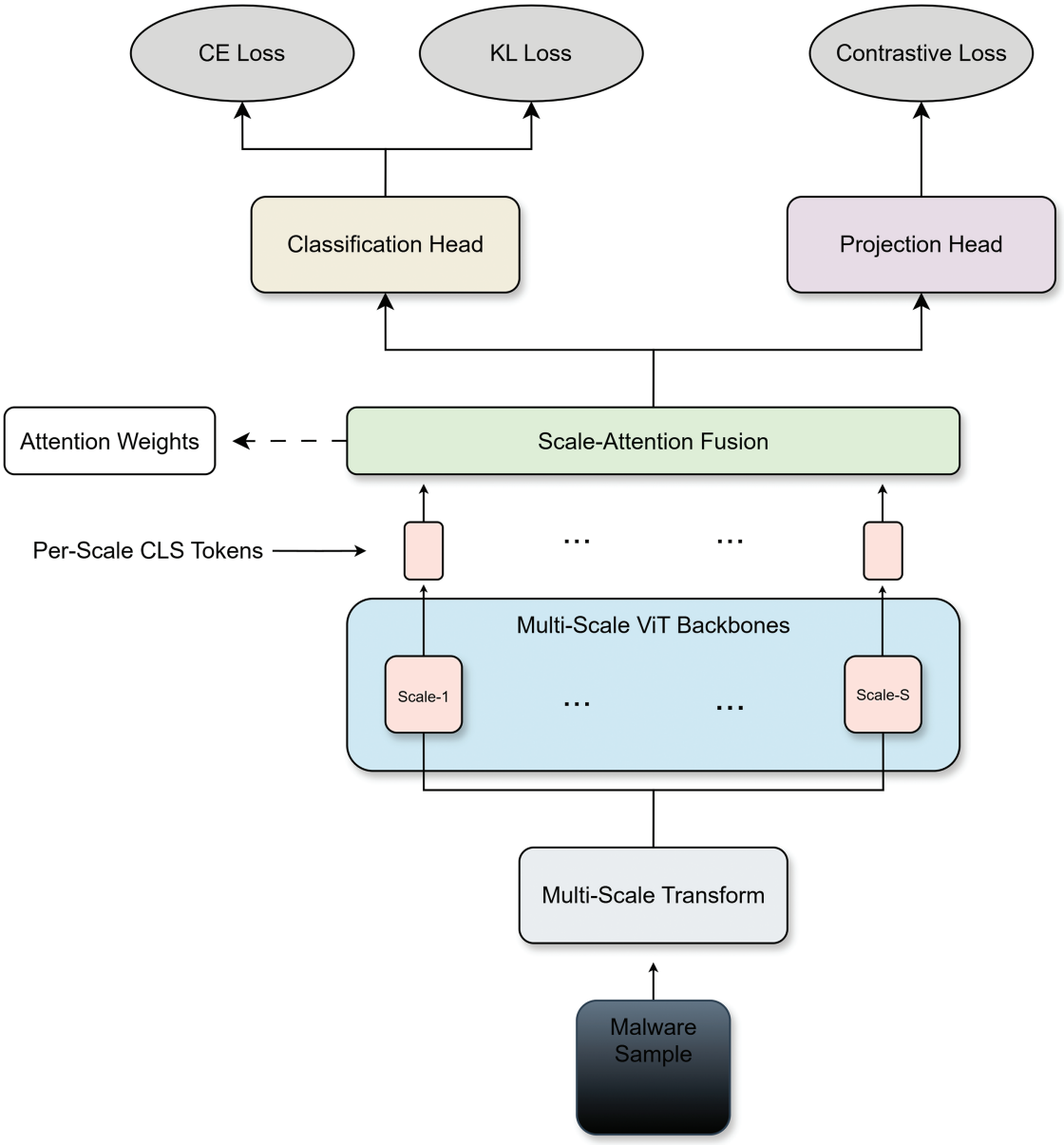


Figure 1: Overview of the proposed MSA-ViT architecture.

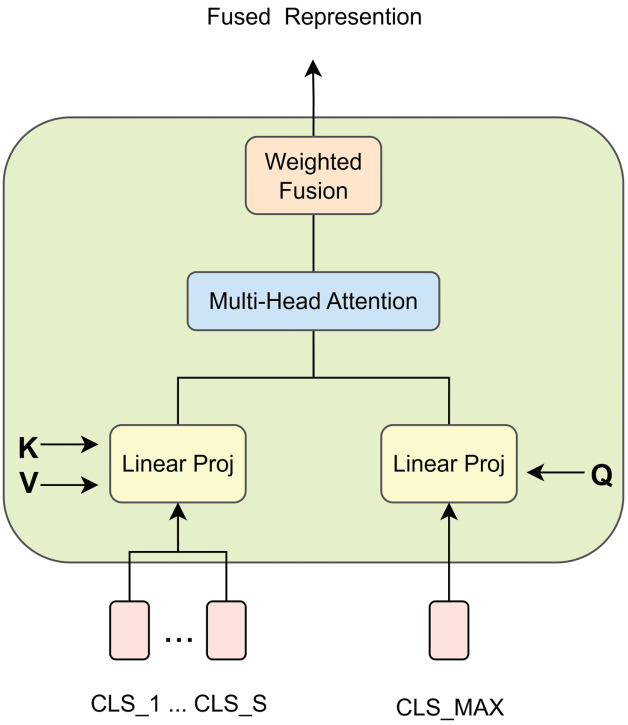


Figure 2: Internal structure of the Scale-Attention Fusion (SAF) module.

Given the structural sensitivity of malware images, only lightweight augmentations are employed to avoid semantic distortion:

- **Random Crop:** Cropping with at least 85% coverage.
- **Horizontal Flip:** Applied with a low probability (10%–15%).
- **Mild Color Jitter:** Brightness and contrast variations within ±15%.

Strong augmentations such as random erasing or large rotations are avoided to preserve intrinsic malware semantics.

3.3 Scale-Attention Fusion

Different malware families exhibit varying reliance on fine-grained textures or coarse structural layouts. Static fusion strategies (e.g., concatenation or averaging) fail to capture such scale-dependent relevance. To address this, we introduce a **Scale-Attention Fusion** mechanism that adaptively weights multi-scale representations.

Given CLS tokens $\{t^{(1)}, \ldots, t^{(S)}\}$ from S scales, where $t^{(s)} \in \mathbb{R}^D$, the CLS token of the largest scale is selected as the query, while tokens from all scales form the keys and values:

$$Q = W_Q t^{(S)}, \quad K = W_K[t^{(1)}, \ldots, t^{(S)}], \quad V = W_V[t^{(1)}, \ldots, t^{(S)}]. \tag{1}$$

Multi-head scaled dot-product attention computes scale-wise importance:

$$\alpha = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right), \quad \alpha \in \mathbb{R}^{1\times S}. \tag{2}$$

The fused representation is obtained as:

$$z_{\text{fuse}} = \alpha V = \sum_{s=1}^{S} \alpha_s W_V t^{(s)}. \tag{3}$$

The attention weights α_s quantify the relative contribution of each scale, enabling interpretable and adaptive multi-scale aggregation. Compared with prior multiscale ViT designs, SAF is lightweight, scale-flexible, and requires no modification to internal Transformer blocks.

3.4 Cross-Scale Consistency Regularization

Without explicit constraints, multi-scale models may over-rely on a single dominant scale, reducing robustness to resolution perturbations. To alleviate this issue, we introduce a **cross-scale consistency regularization** during supervised training.

Let the fused prediction be

$$p = \text{softmax}(h(z_{\text{fuse}})), \tag{4}$$

and the prediction from the s-th scale branch be

$$p_s = \text{softmax}(h(z_s)), \quad s = 1, \ldots, S. \tag{5}$$

We define the consistency loss as the average Kullback–Leibler divergence:

$$L_{\text{con}} = \frac{1}{S} \sum_{s=1}^{S} \text{KL}(p \parallel p_s). \tag{6}$$

The final objective is

$$L = L_{\text{CE}} + \lambda L_{\text{con}}, \quad \lambda \in [0.1, 0.5], \tag{7}$$

where L_{CE} denotes the cross-entropy loss. This regularization encourages all scales to learn discriminative representations, improving robustness under scaling, padding, and cropping perturbations.

3.5 Dual-Head Structure and Training Strategy

We adopt a dual-head design on the fused representation, consisting of a classification head and a projection head, trained using a staged strategy.

Classification Head: During supervised learning, predictions are obtained as

$$p = \text{softmax}(W_{\text{cls}} \cdot \text{LN}(z_{\text{fuse}}) + b_{\text{cls}}), \tag{8}$$

with the standard cross-entropy loss

$$L_{\text{CE}} = -\sum_{c=1}^{C} y_c \log p_c. \tag{9}$$

Projection Head: For self-supervised pre-training, a SimCLR-style projection head [30] maps z_{fuse} to a normalized embedding:

$$z' = \text{ReLU}(W_1 z_{\text{fuse}} + b_1), \quad \hat{z} = \frac{W_2 z' + b_2}{\|W_2 z' + b_2\|}. \tag{10}$$

The NT-Xent loss is defined as

$$L_{\text{NT-Xent}} = -\log \frac{\exp(\hat{z}_1 \cdot \hat{z}_2/\tau)}{\sum_{k=1}^{N} \exp(\hat{z}_1 \cdot \hat{z}_k/\tau)}. \tag{11}$$

Training Procedure:

- **Stage 1:** Self-supervised pre-training using the projection head and NT-Xent loss.
- **Stage 2:** Supervised fine-tuning with the classification head and cross-scale consistency regularization.

This staged learning paradigm leverages unlabeled data to learn robust multi-scale representations and subsequently enhances discriminative performance under supervised training.

4 Experiments and Results

4.1 Dataset Description

We evaluate MSA-ViT on two public malware image benchmarks: Malimg and MaleVis. Malimg contains 9339 grayscale images spanning 25 malware families. MaleVis consists of 14,226 RGB images converted from Windows executables and covers 26 classes, including 25 malicious families (adware, trojans, viruses, worms, and backdoors) and one benign class. Both datasets exhibit inter-class imbalance, motivating the use of metrics beyond accuracy.

4.2 Evaluation Metrics

We report Accuracy (Acc), Precision (Pr), Recall (Re), F1-score, and Matthews Correlation Coefficient (MCC). We further report micro- and macro-averaged AUC for multi-class ROC analysis and visualize confusion matrices to examine class-wise errors.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{12}$$

$$\text{Pr} = \frac{TP}{TP + FP}. \tag{13}$$

$$\text{Re} = \frac{TP}{TP + FN}. \tag{14}$$

$$\text{F1} = 2 \times \frac{\text{Pr} \times \text{Re}}{\text{Pr} + \text{Re}}. \tag{15}$$

4.3 Experimental Setup

All experiments are implemented in PyTorch 2.5.0 with CUDA 13.0 on an NVIDIA A100 GPU (40 GB). We use a stratified 70/30 train/test split and mixed-precision training (FP16).

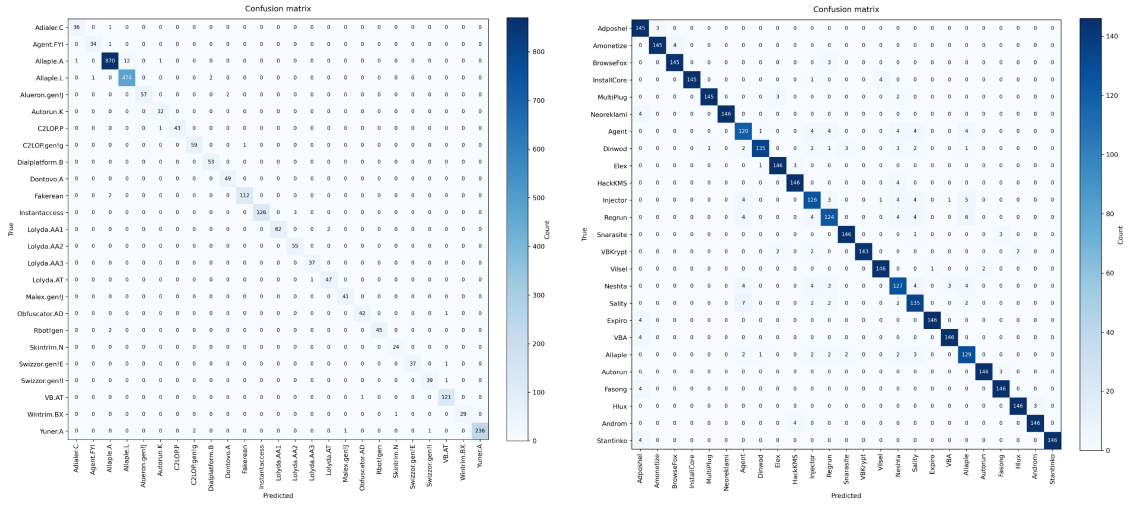
MSA-ViT adopts ViT-B/16 as the backbone architecture and operates on three input scales (128 × 128, 256 × 256, 384 × 384) and trained in two stages. First, each scale branch is pre-trained using SimCLR for 200 epochs (batch size 256, projection dimension 128, temperature $\tau = 0.5$). Second, supervised fine-tuning is performed with an initial 10-epoch warm-up for SAF and classifier training, followed by 150 epochs of end-to-end optimization.

We use AdamW with a learning rate of 5×10^{-4}, cosine annealing with 10% warm-up, weight decay 0.05, batch size 256, DropPath rate 0.1–0.3, gradient clipping (norm = 1.0), and early stopping (patience 20). Hyperparameters are selected via grid search using 10% of the training set. Lightweight augmentations are applied during training, including random cropping (≥85%), horizontal flipping ($p = 0.1$), brightness and contrast jitter (±15%), and normalization to $[0,1]$. We further apply KL-based scale-consistency regularization ($\lambda = 0.5$) and label smoothing (0.1).

The final model requires 3.1 GFLOPs per inference. For a fair comparison of computational efficiency, all FLOPs reported in this paper correspond to the effective inference cost measured under the same profiling setting using a standard FLOPs profiling tool. Specifically, FLOPs for MSA-ViT and all Transformer-based baselines (ViT-B/16, Swin-T, CrossViT, and MAFormer) are measured with identical input resolutions and batch size. Therefore, the reported FLOPs reflect practical inference cost rather than the theoretical computational complexity reported in original papers.

4.4 Results and Discussion

Fig. 3 reports confusion matrices on Malimg and MaleVis. Both matrices exhibit strong diagonal dominance, indicating that most samples are correctly classified. Errors are sparse and mainly occur among a small number of visually similar families, suggesting stable decision boundaries across classes.



(a) Malimg (b) MaleVis

Figure 3: Confusion matrices of MSA-ViT on (**a**) Malimg and (**b**) MaleVis.

We acknowledge that Malimg is a widely used and relatively saturated benchmark, where many approaches have already achieved high accuracy. In this work, Malimg is primarily used for robustness analysis, ablation studies, and interpretability evaluation, while the more challenging MaleVis dataset is emphasized for assessing generalization and practical effectiveness.

We further evaluate the threshold-independent discriminative capability of the proposed model using ROC curves. As illustrated in Fig. 4, on the Malimg dataset, MSA-ViT achieves micro- and macro-averaged AUC values of 0.99, indicating an excellent separability among malware families even under severe class imbalance. As depicted in Fig. 5, on the more challenging MaleVis dataset, the model attains micro- and

macro-averaged AUC values of 0.98, demonstrating robust multi-class discrimination performance across varying decision thresholds.

Table 2 summarizes comparisons with representative baselines. MSA-ViT achieves 98.5% accuracy on Malimg and 96.0% on MaleVis, outperforming prior methods across ACC/Pr/Re/F1.

We conduct ablation studies to quantify the contribution of key components (multi-scale inputs, SAF, SimCLR pre-training, and consistency regularization). Table 3 shows that each component provides consistent gains, with multi-scale inputs and SAF contributing the largest improvements. **Scale Sensitivity Analysis.** To further justify the selection of the fixed three-scale configuration (128×128, 256 × 256, 384 × 384), we conduct a scale sensitivity analysis by progressively removing either the smallest or the largest scale. As shown in Table 3, reducing the model to a single-scale input (256 × 256) results in an accuracy drop of approximately 2.4% on Malimg and 2.4% on MaleVis, confirming the limitation of fixed-resolution modeling. Using two scales partially recovers performance, yet still underperforms the full tri-scale setting by 1.0%–1.2% across ACC, F1, and MCC.

These results indicate that the smallest scale contributes robust global structural cues, while the largest scale provides complementary fine-grained texture information. Removing either scale disrupts this balance and degrades performance. Therefore, the adopted three-scale design offers a stable trade-off between representation diversity and computational efficiency, and the performance gains are not sensitive to a specific single resolution but instead stem from effective multi-scale collaboration.

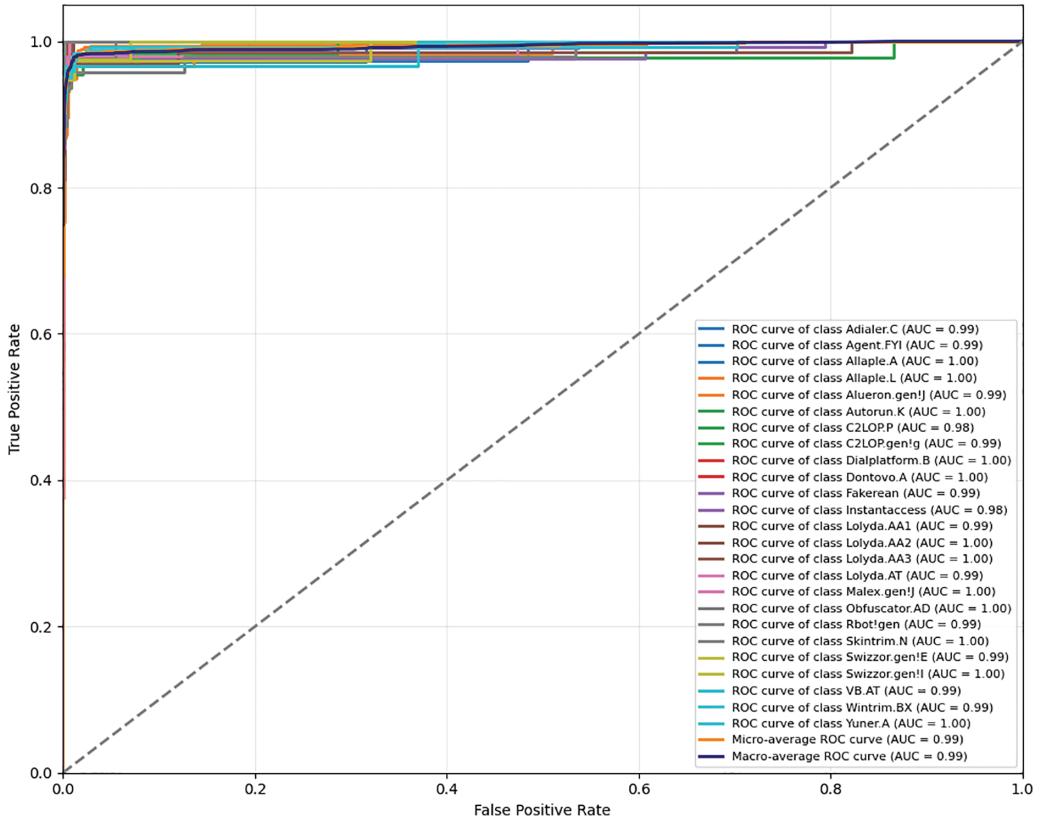


Figure 4: ROC curves of MSA-ViT on the Malimg dataset.

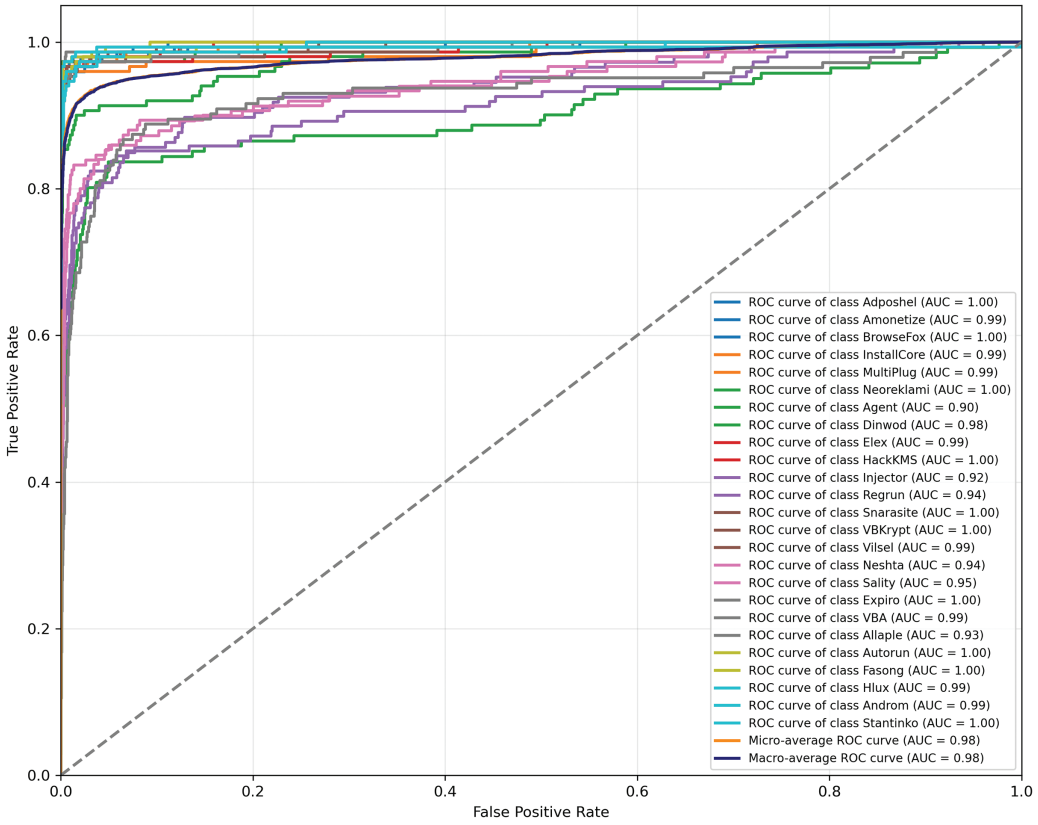


Figure 5: ROC curves of MSA-ViT on the MaleVis dataset.

Table 2: Comparison with representative baselines on Malimg and MaleVis.

Method	Malimg				MaleVis			
	ACC	Pr	Re	F1	ACC	Pr	Re	F1
Singh et al. [31]	0.9780	0.9760	0.9750	0.9740	–	–	–	–
Sruthi et al. [32]	0.9630	0.9630	0.9582	0.9606	0.8629	0.8685	0.8628	0.8656
Belal and Sundaram [9]	0.9828	0.9620	0.9637	0.9622	–	–	–	–
Cui et al. [33]	0.9450	0.9464	0.9431	0.9447	0.9213	0.9209	0.9189	0.9199
Nkrumah et al. [1]	–	–	–	–	0.9400	0.9400	0.9400	0.9400
Paik and Jin [34]	–	–	–	–	0.9349	0.9290	0.9349	0.9282
Nivaashini et al. [2]	0.9760	0.9800	0.9700	0.9700	0.9070	0.9100	0.9000	0.9000
MSA-ViT (Ours)	**0.9850**	**0.9792**	**0.9832**	**0.9812**	**0.9600**	**0.9606**	**0.9598**	**0.9598**

Table 3: Ablation results on Malimg and MaleVis.

Configuration	Malimg			MaleVis		
	ACC	F1	MCC	ACC	F1	MCC
Single Scale (256 × 256)	0.9600	0.9565	0.9540	0.9360	0.9358	0.9345
Dual Scales (128 + 256)	0.9735	0.9708	0.9695	0.9486	0.9454	0.9441
w/o SAF (Avg. Pooling)	0.9670	0.9635	0.9620	0.9389	0.9387	0.9374
w/o SimCLR Pre-training	0.9700	0.9668	0.9655	0.9456	0.9454	0.9441
w/o Consistency Reg.	0.9730	0.9698	0.9685	0.9485	0.9483	0.9470
Full MSA-ViT (Ours)	0.9850	0.9812	0.9824	0.9600	0.9598	0.9585

Table 4 reports robustness under perturbations relevant to malware visualization. Across scaling, padding, and FGSM attacks, MSA-ViT limits accuracy loss to below 1.8% on both datasets, supporting the robustness benefits of multi-scale inputs, SAF, and consistency regularization.

Table 4: Robustness of MSA-ViT under perturbations.

Perturbation	Malimg			MaleVis		
	ACC	F1	MCC	ACC	F1	MCC
Baseline	0.9850	0.9812	0.9824	0.9600	0.9598	0.9585
Scaling 0.4×	0.9695	0.9657	0.9670	0.9427	0.9425	0.9412
Scaling 2.5×	0.9738	0.9700	0.9712	0.9456	0.9454	0.9441
Padding 10%	0.9820	0.9784	0.9796	0.9581	0.9579	0.9566
Padding 20%	0.9790	0.9754	0.9765	0.9523	0.9521	0.9508
Padding 30%	0.9740	0.9704	0.9716	0.9466	0.9464	0.9451
FGSM (ε = 0.0039)	0.9700	0.9664	0.9676	0.9421	0.9419	0.9409

To interpret SAF, Fig. 6 visualizes the average scale-attention weights. On Malimg, low-resolution cues dominate, while on MaleVis the mid-resolution branch receives the highest weight, indicating that SAF adapts scale contributions to dataset characteristics.

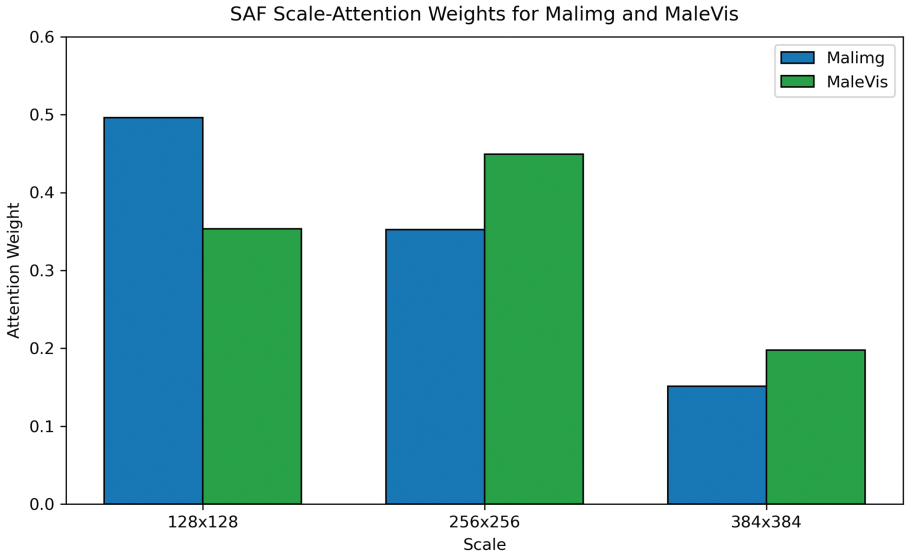


Figure 6: Scale-attention weight distributions on Malimg and MaleVis.

5 Conclusions and Future Work

This paper presents **MSA-ViT**, a lightweight multi-scale Vision Transformer framework for malware image classification. MSA-ViT integrates parallel multi-scale ViT backbones with a **Scale-Attention Fusion (SAF)** module to adaptively aggregate cross-scale representations. In addition, SimCLR-based self-supervised pre-training and KL-divergence-based cross-scale consistency regularization are incorporated to improve generalization and provide insights into scale-aware feature utilization. Experimental results on Malimg and MaleVis demonstrate that MSA-ViT achieves competitive performance, with accuracies of 98.5% and 96.0%, respectively, while maintaining a low inference cost of 3.1 GFLOPs, making it suitable for deployment in resource-constrained environments.

Ablation studies confirm the complementary contributions of multi-scale inputs and the SAF module, while robustness evaluations indicate that performance degradation under scaling, padding, and FGSM-based perturbations remains below 1.8%. It is important to note, however, that adversarial robustness is evaluated at the image-representation level rather than at the executable binary level. Pixel-level perturbations introduced by FGSM may render binaries non-executable; therefore, the reported robustness reflects the stability of the visual classifier rather than executable-preserving adversarial resistance. Moreover, as image-based malware analysis constitutes a form of static analysis, it remains susceptible to advanced obfuscation techniques that disrupt visual texture while preserving malicious functionality.

Despite these advantages, several limitations remain. First, the use of a fixed set of input scales may reduce adaptability to malware samples with extreme resolution or length variability. Second, although self-supervised pre-training alleviates label scarcity, robustness under severe noise, distribution shifts, or executable-preserving adversarial manipulations warrants further investigation. Third, the current study focuses on static image representations and does not incorporate dynamic execution behaviors, which may provide complementary information for real-world malware detection.

Future work will explore several promising directions. These include: (i) adaptive or data-driven scale selection mechanisms to better accommodate highly variable malware samples; (ii) robustness enhancements under more realistic threat models, including executable-preserving attacks and distribution shifts; and (iii) extensions to broader deployment settings, such as online or real-time detection, privacy-preserving

and federated learning frameworks, and cross-platform evaluations (e.g., Android malware), to address evolving and heterogeneous security threats.

Acknowledgement: I would like to express their heartfelt gratitude to their supervisor, Professor Chuanping Hu, for his invaluable guidance and support throughout the study and research. His insightful advice and encouragement in academic exploration, manuscript writing, and critical thinking have been a constant source of inspiration and motivation.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Bofan Yang; data collection: Bofan Yang; analysis and interpretation of results: Bofan Yang, Bingbing Li; draft manuscript preparation: Bofan Yang; supervision: Bofan Yang, Chuanping Hu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: Not applicable. This study did not involve human participants or animal subjects.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nkrumah B, Asante M, Adbdul-Salam G, Adu-Gyamfi WK. Data-efficient image transformers for robust malware family classification. J Cyber Secur. 2024;6(1):131–53 doi:10.32604/jcs.2024.053954.
2. Nivaashini M, Aarthi S, Ramya S. MalNet: detection of malwares using ensemble learning techniques. In: Proceedings of the 7th International Conference on Electronics, Communication and Aerospace Technology (ICECA). Piscataway, NJ, USA: IEEE; 2023. p. 1469–77.
3. Moser A, Krügel C, Kirda E. Limits of static analysis for malware detection. In: Proceedings of the 23rd Annual Computer Security Applications Conference (ACSAC); 2007 Dec 10–14; Miami Beach, FL, USA. p. 421–30.
4. Nataraj L, Karthikeyan S, Jacob G, Manjunath BS. Malware images: visualization and automatic classification. In: VizSec'11: Proceedings of the 8th International Symposium on Visualization for Cyber Security. New York, NY, USA: ACM; 2011. p. 1–7.
5. Makandar A, Patrot A. Malware class recognition using image processing techniques. In: Proceedings of the International Conference on Data Management, Analytics and Innovation (ICDMAI); 2017 Feb 24–26; Pune, India. p. 76–80.
6. Roseline SA, Geetha S, Kadry S, Nam Y. Intelligent vision-based malware detection and classification using deep random forest paradigm. IEEE Access. 2020;8:206303–24 doi:10.1109/access.2020.3036491.
7. Lin TY, Dollár P, Girshick RB, He K, Hariharan B, Belongie SJ. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44.
8. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2012;60:84–90 doi:10.1145/3065386.
9. Belal MM, Sundaram DMS. Global-local attention-based butterfly vision transformer for visualization-based malware classification. IEEE Access. 2023;11:69337–55 doi:10.1109/access.2023.3293530.
10. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
11. Seneviratne S, Shariffdeen R, Rasnayaka S, Kasthuriarachchi N. Self-supervised vision transformers for malware detection. IEEE Access. 2022;10:103121–35 doi:10.1109/access.2022.3206445.
12. Syed TA, Nauman M, Khan S, Jan S, Zuhairi MFA. ViTDroid: vision transformers for efficient, explainable attention to malicious behavior in Android binaries. Sensors. 2024;24(20):6690. doi:10.3390/s24206690.

13. Bavishi S, Modi S. Accelerating malware classification: a vision transformer solution. arXiv:2409.19461. 2024.
14. Yang D, Ding Y, Zhang H, Li Y. PVitNet: an effective approach for Android malware detection using pyramid feature processing and vision transformer. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2024 Apr 14–19; Seoul, Republic of Korea. p. 2440–4.
15. Jeong IW, Lee HJ, Kim GN, Choi SH. MalFormer: a novel vision transformer model for robust malware analysis. IEEE Access. 2025;13:122671–83 doi:10.1109/access.2025.3588232.
16. Kunwar P, Aryal K, Gupta M, Abdelsalam M, Bertino E. SoK: leveraging transformers for malware analysis. arXiv:2405.17190. 2024.
17. Nair SJ, Syam SR. Comparing transformers and CNN approaches for malware detection: a comprehensive analysis. In: Proceedings of the 15th International Conference on Computing, Communication and Networking Technologies (ICCCNT); 2024 Jul 1–3; Mandi, India.
18. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada. p. 9992–10002.
19. Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, et al. Multiscale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada. p. 6804–15.
20. Chen CF, Fan Q, Panda R. CrossViT: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 11–17; Montreal, QC, Canada. p. 347–56.
21. Li Y, Wu C, Fan H, Mangalam K, Xiong B, Malik J, et al. MViTv2: improved multiscale vision transformers for classification and detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 19–24; New Orleans, LA, USA. p. 4794–804.
22. Ashawa M, Owoh NP, Hosseinzadeh S, Osamor J. Enhanced image-based malware classification using transformer-based convolutional neural networks. Electronics. 2024;13(20):4081 doi:10.3390/electronics13204081.
23. Ullah F, Alsirhani A, Alshahrani MM, Alomari A, Naeem H, Shah SA. Explainable malware detection system using transformers-based transfer learning and multi-model visual representation. Sensors. 2022;22(18):6766 doi:10.3390/s22186766.
24. Bozkir AS, Cankaya AO, Aydos M. Utilization and comparison of convolutional neural networks in malware recognition. In: 2019 27th Signal Processing and Communications Applications Conference (SIU); 2019 Apr 24–26; Sivas, Turkey. p. 1–4.
25. Wang Y, Sun HX, Wang X, Zhang B, Li C, Xin Y, et al. MAFormer: a transformer network with multi-scale attention fusion for visual recognition. Neurocomputing. 2022;595:127828 doi:10.1016/j.neucom.2024.127828.
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS); 2017 Dec 4–9; Long Beach, CA, USA.
27. Naseer M, Ranasinghe K, Khan SH, Hayat M, Khan FS, Yang MH. Intriguing properties of vision transformers. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS); 2021 Dec 6–14; Online. Red Hook, NY, USA: Curran Associates, Inc.; 2021.
28. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 19–24; New Orleans, LA, USA. p. 15979–88.
29. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y. Transformer in transformer. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS); 2021 Dec 6–14; Online. Red Hook, NY, USA: Curran Associates, Inc.; 2021.
30. Chen T, Kornblith S, Norouzi M, Hinton GE. A simple framework for contrastive learning of visual representations. arXiv:2002.05709. 2020.
31. Singh S, Krishnan D, Vazirani V, Vinayakumar R, Alsuhibany SA. Deep hybrid approach with sequential feature extraction and classification for robust malware detection. Egypt Inform J. 2024;27:100539 doi:10.1016/j.eij.2024.100539.

32. Sruthi P, Singh S, Shoiab M, Kumar D, Reddy LV, Shnain AH, et al. Robust intelligent malware detection using deep learning. IEEE Access. 2019;7:46717–38 doi:10.1063/5.0261642.
33. Cui Z, Xue F, Cai X, Cao Y, Wang GG, Chen J. Detection of malicious code variants based on deep learning. IEEE Trans Ind Inform. 2018;14:3187–96 doi:10.1109/tii.2018.2822680.
34. Paik JY, Jin R. Malware family prediction with an awareness of label uncertainty. Comput J. 2024;67:376–90. doi:10.1093/comjnl/bxac181.