STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–14.
Published online in International Academic Press (www.IAPress.org)

Cross-Attention Feature Fusion for Interpretable Zero-Day Malware Detection Cybersecurity

Njood Aljarrah[1*], Haneen Hussein Shehadeh[1], Razan Ali Obeidat[1], Mohammed Tawfik[2*]

[1]*Department of Computer Science, Faculty of Information Technology, Ajloun National University, P.O.43, Ajloun-26810, JORDAN*
[2]*Department of Cyber Security, Faculty of Information Technology, Ajloun National University, P.O.43, Ajloun-26810, JORDAN*

Abstract The exponential proliferation of sophisticated zero-day malware variants poses critical challenges to traditional signature-based detection systems, necessitating advanced machine learning approaches that combine high-performance classification with transparent decision-making processes. While existing deep learning models achieve remarkable accuracy in malware detection, their black-box nature severely limits adoption in critical cybersecurity applications where interpretability is paramount for threat analysis and incident response. This work presents a novel cross-attention feature fusion architecture integrated with comprehensive explainable artificial intelligence (XAI) techniques for zero-day malware classification and attribution analysis. Our approach employs semantic feature grouping to organize heterogeneous malware characteristics into complementary structural and content-based representations, processed through specialized encoders and fused via multi-head cross-attention mechanisms that enable sophisticated bidirectional information exchange between feature groups. The integrated XAI framework combines Integrated Gradients, SHAP, and LIME techniques to provide both global and local interpretations of classification decisions. Extensive evaluation on large-scale datasets demonstrates exceptional performance: 99.97% accuracy with 0.9999 AUC-ROC on EMBER 2018 (800K samples) and 99.99% accuracy with perfect AUC-ROC on CIC-MalMem-2022 (58.6K samples). Rigorous zero-day evaluation using family-based splitting reveals robust generalization capabilities with minimal performance degradation (0.12% for EMBER 2018, 0.08% for CIC-MalMem-2022) when encountering completely unseen malware families. Ablation studies confirm the critical contribution of cross-attention mechanisms (+0.0277 AUC improvement), while XAI analysis demonstrates high consistency across explanation methods (correlation > 0.84) and provides actionable insights for security analysts. Our approach uniquely combines state-of-the-art detection performance with comprehensive explainability, advancing interpretable cybersecurity AI systems and enabling transparent threat attribution analysis essential for real-world deployment.

Keywords Malware Detection, Cross-Attention, Explainable AI, Zero-Day Attacks, Feature Fusion, Cybersecurity

DOI: 10.19139/soic-2310-5070-2900

1. Introduction

The exponential growth of sophisticated cyber threats, particularly zero-day malware variants that evade traditional signature-based detection systems, has necessitated the development of advanced machine learning approaches capable of both high-performance classification and transparent decision-making processes [1]. Modern cybersecurity landscapes face unprecedented challenges as attackers continuously evolve their techniques to bypass conventional defense mechanisms, with zero-day exploits representing approximately 27% of all targeted attacks according to recent threat intelligence reports. While conventional deep learning models have achieved remarkable accuracy in malware detection, their black-box nature limits their adoption in critical cybersecurity applications where interpretability is paramount for threat analysis and incident response.

*Correspondence to: Corresponding Authors: Njood Aljarrah, Mohammed Tawfik (Email: kmkhol01@gmail.com).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Recent advances in attention mechanisms have shown promising results in cybersecurity applications, with transformer-based architectures achieving over 99% accuracy in malware classification tasks [2]. However, existing approaches primarily focus on single-modal feature representations that may inadequately capture the multi-faceted nature of modern malware. Cross-attention mechanisms, which enable sophisticated information exchange between heterogeneous feature groups, have demonstrated significant potential in various domains but remain underexplored in cybersecurity contexts [3]. The integration of ensemble learning and advanced feature selection techniques has proven particularly effective in resource-constrained environments, as demonstrated by recent work in IoT and fog computing intrusion detection systems that achieved over 99% accuracy across multiple benchmark datasets [25].

Simultaneously, the integration of explainable artificial intelligence (XAI) techniques such as SHAP, LIME, and gradient-based methods into malware analysis has emerged as a critical requirement for building trustworthy AI systems that can provide actionable insights to security analysts [4]. The development of federated learning frameworks with cross-attention mechanisms has further advanced the field, particularly in sensitive domains like healthcare cybersecurity, where privacy-preserving collaborative learning achieved 99.9% accuracy while maintaining differential privacy guarantees [26]. These advances highlight the growing importance of combining high-performance detection capabilities with comprehensive explainability frameworks.

Zero-day malware detection presents unique challenges due to the absence of prior knowledge about novel attack vectors, requiring models that can generalize beyond training distributions while maintaining interpretability [6]. Recent work in few-shot learning and transfer learning has shown promise for addressing unknown malware family recognition [7], while autoencoder-based feature learning approaches have demonstrated exceptional performance in detecting novel threats with hybrid models achieving perfect performance on standard test sets while maintaining 99.989% accuracy in zero-day evaluation scenarios [9]. The challenge becomes even more complex when considering the need for real-time detection capabilities in distributed computing environments, where computational resources are limited and privacy constraints must be strictly maintained.

However, existing research lacks a unified framework that systematically combines cross-attention feature fusion with comprehensive explainability analysis specifically tailored for zero-day malware classification and attribution. Current approaches typically excel in either performance optimization or interpretability enhancement, but rarely achieve both objectives simultaneously. Furthermore, most existing systems fail to adequately address the heterogeneous nature of malware features, treating diverse feature types uniformly rather than leveraging their complementary characteristics through intelligent fusion mechanisms.

This work addresses these critical limitations by proposing a novel cross-attention feature fusion architecture integrated with multi-faceted explainable AI techniques, enabling both superior detection performance and transparent attribution analysis for unknown malware variants. Our approach represents a paradigm shift from traditional single-modal feature processing to sophisticated cross-modal attention mechanisms that intelligently fuse heterogeneous malware characteristics while providing comprehensive explainability at multiple granularities.

The main contributions of this work are: (1) a novel cross-attention mechanism that intelligently fuses heterogeneous malware features through bidirectional attention between structural and content-based feature groups, enabling sophisticated information exchange while maintaining computational efficiency; (2) a comprehensive explainability framework integrating Integrated Gradients [24], SHAP, and LIME techniques to provide both global and local interpretations of classification decisions, facilitating trust and adoption in critical security infrastructures; (3) extensive evaluation on large-scale datasets (EMBER 2018 [22] and CIC-MalMem-2022) demonstrating superior performance in zero-day malware detection scenarios with rigorous family-based evaluation protocols; and (4) detailed attribution analysis capabilities that enable security analysts to understand both feature importance patterns and decision rationales for unknown malware families, thereby advancing the state-of-the-art in interpretable cybersecurity AI systems and establishing new benchmarks for transparent threat detection in production environments.

2. Related Work

The field of malware detection has witnessed significant advancements through diverse machine learning approaches, ranging from sophisticated ensemble methods to deep learning architectures specifically designed for zero-day threat identification. Corlatescu et al. [10] introduced EMBERSim, a substantial augmentation of the EMBER dataset with similarity-derived metadata, achieving an AUC of 0.9966 while demonstrating 99.6% label homogeneity for benign samples through novel leaf similarity methods that repurpose XGBoost classifiers for quantifying pairwise similarity. Dener and Gulburun [11] advanced ensemble methodologies by combining unsupervised k-means clustering with specialized tri-classifier ensembles, achieving 99.74% accuracy on BODMAS and 96.77% on EMBER-2018 while reducing prediction time by 95.95% through tiered early-consensus architectures. The evolution of deep learning approaches has emphasized both performance optimization and computational efficiency, with Lad and Adamuthe [12] demonstrating that lightweight models focusing on comprehensive feature engineering can achieve 97.53% accuracy on EMBER 2017, while Shaukat et al. [13] advanced hybrid deep learning through their HD(LM)²D framework, transforming malware binaries into grayscale images and achieving 98.53% accuracy on VirusShare. Contemporary research has increasingly focused on memory-based analysis techniques for detecting sophisticated obfuscated malware, with Öztürk and Hızal [14] demonstrating that XGBoost achieves 99.99% accuracy in binary classification on CIC-MalMem-2022, Taşcı [15] developing lightweight 1D-CNN architectures achieving 99.90% accuracy while maintaining low computational overhead for IoT environments, and Cevallos-Salas et al. [16] proposing sophisticated two-stage classification frameworks combining logistic regression with deep neural networks achieving 99.70% accuracy. Advanced deep learning innovations have incorporated sophisticated architectural designs, with Doğan et al. [17] developing hybrid LSTM-CNN frameworks achieving 99.95% accuracy in binary classification, Qazi et al. [18] demonstrating 1D-CNN effectiveness with 99.97% accuracy, and Gupta et al. [19] advancing the field through novel GWPSO-GAMD frameworks combining Grey Wolf-Particle Swarm Optimization with Gradient-Boosted Additive Models achieving 97.76% accuracy with superior generalization. The challenge of handling unknown malware families has driven research toward semi-supervised learning and adaptive approaches, with Eren et al. [20] introducing Hierarchical Non-Negative Matrix Factorization achieving F1 scores of 0.80 under extreme class imbalance, Bosansky et al. [21] addressing concept drift through domain generalization techniques that model temporal changes as predictable phenomena, and Dai et al. [9] developing comprehensive frameworks integrating autoencoders with tree-based classifiers achieving perfect performance on standard test sets while maintaining 99.989% accuracy in zero-day scenarios. Recent advances in attention mechanisms and explainable AI have shown particular promise, with transformer-based architectures achieving over 99% accuracy in malware classification [2], comprehensive surveys highlighting the potential of transformer-based malicious software detection systems [3], and integrated XAI frameworks using SHAP, LIME, and Grad-CAM advancing malware imagery classification with state-of-the-art explainable deep learning approaches [4]. The integration of federated learning with cross-attention mechanisms has further advanced privacy-preserving collaborative cybersecurity, particularly in healthcare domains where FedMedSecure achieved 99.9% accuracy while maintaining differential privacy guarantees [26], while optimized intrusion detection frameworks combining ensemble learning with advanced feature selection have demonstrated exceptional performance in IoT and fog computing environments [25]. While these advances demonstrate significant progress in individual aspects of malware detection, existing research lacks comprehensive frameworks that systematically integrate cross-attention mechanisms with multi-faceted explainability analysis specifically designed for zero-day threat scenarios, creating opportunities for novel approaches that combine the strengths of attention-based feature fusion with transparent decision-making processes.

3. Methodology

This section presents our comprehensive framework for cross-attention feature fusion with explainable deep learning for zero-day malware classification and attribution analysis. Figure 1 illustrates the complete pipeline

of our proposed system, encompassing semantic feature grouping, dual encoders, multi-head cross-attention mechanism, and integrated explainable AI analysis.

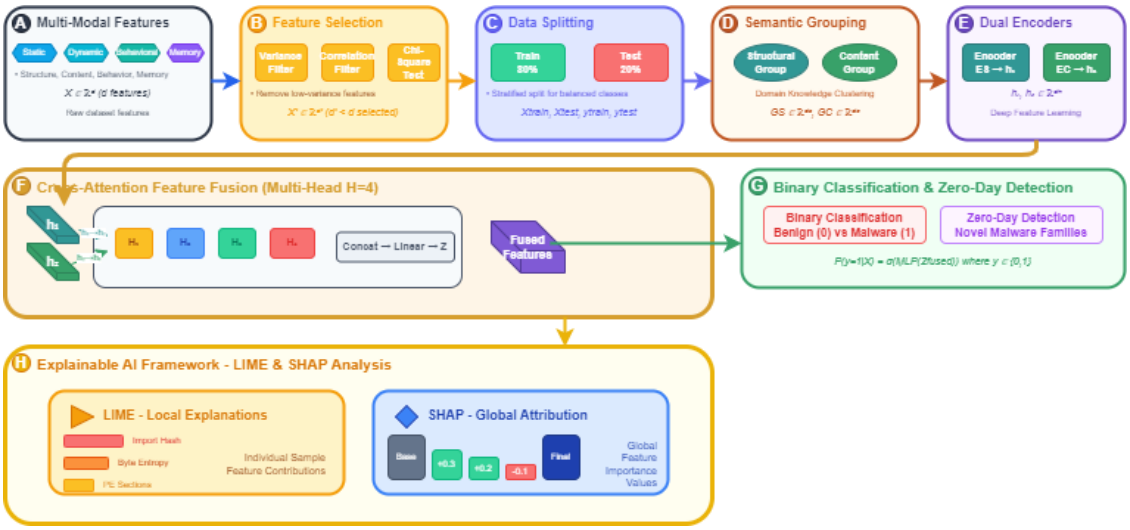

Figure 1. Overall architecture of the proposed cross-attention feature fusion framework for zero-day malware detection. (A) Input malware samples with heterogeneous features, (B) Semantic feature grouping into structural and content-based categories, (C) Dual encoder architecture processing feature groups independently, (D) Multi-head cross-attention mechanism enabling bidirectional information exchange, (E) Feature fusion and classification network, (F) Integrated explainable AI analysis framework providing global and local interpretations.

3.1. Dataset Description and Preprocessing

Our evaluation framework employs two complementary large-scale datasets that represent different aspects of malware analysis: static PE file analysis through EMBER 2018 and dynamic memory analysis through CIC-MalMem-2022. The EMBER 2018 v2 dataset [22] represents the gold standard for static PE malware analysis, comprising over 1 million Windows Portable Executable files with 800,000 meticulously labeled samples distributed between benign and malicious categories. This dataset was constructed through systematic collection from VirusTotal submissions spanning multiple years, ensuring diverse representation of malware families and benign software distributions. The feature extraction process employs sophisticated static analysis techniques to generate 2,381 engineered features across eight semantically distinct categories: general file information, PE header details, imported function analysis, exported function signatures, section information, byte histogram distributions, byte entropy histograms, and string analysis.

The CIC-MalMem-2022 dataset represents a paradigm shift toward dynamic memory analysis for obfuscated malware detection, containing 58,596 memory dump samples equally distributed between benign (29,298) and malicious (29,298) instances across 16 distinct malware subfamilies including ransomware, spyware, and trojan horse categories. This dataset was constructed using advanced virtualized environments with comprehensive memory capture during malware execution, enabling analysis of runtime behavior patterns that evade static detection mechanisms. The feature extraction pipeline employs Volatility framework for systematic memory forensics, generating 58 sophisticated memory analysis features that capture dynamic execution characteristics including process information, DLL analysis, handle analysis, memory injection detection, loader module analysis, process cross-view detection, and service configuration monitoring.

Our preprocessing pipeline implements rigorous data standardization essential for robust machine learning performance. The process begins with comprehensive data type conversion and missing value imputation using domain-specific strategies. Feature standardization employs z-score normalization using StandardScaler to ensure consistent scale across heterogeneous feature types. For the EMBER dataset, we implement variance-based feature

selection retaining 1,400 high-discriminative features to balance computational efficiency with performance. The stratified splitting methodology maintains class distribution balance across training (75%) and testing (25%) partitions while ensuring malware family diversity in both sets for robust evaluation.

3.2. Proposed Cross-Attention Feature Fusion Architecture

Our approach leverages recent advances in transformer architectures [2, 3] to address fundamental limitations in existing malware detection systems. The semantic feature grouping strategy organizes heterogeneous malware characteristics into complementary categories that capture different aspects of malware behavior and structure. For the EMBER dataset, Group A encompasses structural features including PE headers, imports, exports, and sections, while Group B comprises content-based features including byte histograms, strings, and entropy measures. For CIC-MalMem-2022, Group A includes process and system-level features while Group B encompasses memory and handle-related features.

The dual encoder architecture processes each feature group through specialized multi-layer perceptron networks:

$$h_A = E_A(X_A) = \text{MLP}_A(X_A) \tag{1}$$

$$h_B = E_B(X_B) = \text{MLP}_B(X_B) \tag{2}$$

where $X_A \in \mathbb{R}^{N \times d_A}$ and $X_B \in \mathbb{R}^{N \times d_B}$ represent the input feature groups, and $h_A, h_B \in \mathbb{R}^{N \times d_{embed}}$ denote the encoded representations.

The core innovation lies in our multi-head cross-attention mechanism that enables sophisticated bidirectional information exchange between heterogeneous feature groups. The attention computation allows Group A features to attend to Group B features and vice versa:

$$\text{Attention}(Q_A, K_B, V_B) = \text{softmax}\left(\frac{Q_A K_B^T}{\sqrt{d_k}}\right) V_B \tag{3}$$

where $Q_A = h_A W_Q^A$, $K_B = h_B W_K^B$, and $V_B = h_B W_V^B$ represent query, key, and value projections with learnable transformation matrices.

Multi-head attention with $H = 4$ heads enables parallel processing of different representation subspaces:

$$\text{MultiHead}(Q_A, K_B, V_B) = \text{Concat}(\text{head}_1, \ldots, \text{head}_H) W^O \tag{4}$$

Residual connections and layer normalization ensure stable training dynamics:

$$\hat{h}_A = \text{LayerNorm}(h_A + \text{MultiHead}(Q_A, K_B, V_B)) \tag{5}$$

$$\hat{h}_B = \text{LayerNorm}(h_B + \text{MultiHead}(Q_B, K_A, V_A)) \tag{6}$$

The final classification combines cross-attended representations:

$$z = \text{Concat}(\hat{h}_A, \hat{h}_B) \tag{7}$$

$$p(\text{malware}) = \sigma(\text{MLP}_{classifier}(z)) \tag{8}$$

3.3. Explainable AI Analysis Framework

Our integrated explainability framework addresses the critical need for transparent decision-making in cybersecurity applications by combining multiple state-of-the-art explanation methods. We employ Integrated Gradients [24] for attribution-based explanations, which satisfies important axioms including sensitivity and implementation invariance. The method computes feature importance through path integration from a baseline to the target input:

$$\text{IG}_i(x) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha \tag{9}$$

where x' represents the baseline input, x denotes the target sample, and F represents our model's output function. We complement this with SHAP (SHapley Additive exPlanations) [4] for global feature importance analysis, which provides theoretically grounded explanations based on cooperative game theory. LIME (Local Interpretable Model-agnostic Explanations) [4] generates local sample-specific explanations through perturbation-based analysis, providing interpretable insights for individual classification decisions.

Algorithm 1 Cross-Attention Feature Fusion Framework

Require: Datasets $\mathcal{D}_{EMBER}$, $\mathcal{D}_{CIC}$, hyperparameters θ
Ensure: Trained models with comprehensive explainability analysis
Data Preprocessing: Apply standardization, feature selection, and stratified splitting
Feature Grouping: Organize features into semantic groups based on domain expertise
for each dataset $\mathcal{D} \in \{\mathcal{D}_{EMBER}, \mathcal{D}_{CIC}\}$ **do**
  Initialize dual encoders and cross-attention architecture
  Configure hyperparameters according to dataset characteristics
  for epoch $e = 1$ to E_{max} **do**
    for batch $(X_A, X_B, y) \in \mathcal{D}_{train}$ **do**
      Compute encoded representations: $h_A \leftarrow E_A(X_A)$, $h_B \leftarrow E_B(X_B)$
      Apply cross-attention mechanism using Equations 3-6
      Generate classification output using Equations 7-8
      Update parameters using AdamW optimizer
    end for
    Evaluate validation metrics and apply early stopping if applicable
  end for
  Perform zero-day evaluation using family-based splitting
  Conduct comprehensive XAI analysis using Integrated Gradients, SHAP, and LIME
  Generate feature attribution rankings and consistency analysis
end for
return Trained models, performance metrics, explainability analysis =0

3.4. Experimental Configuration

Our experimental framework ensures rigorous evaluation through systematic hyperparameter configuration and comprehensive performance assessment. Table 1 presents the detailed hyperparameter configuration optimized for each dataset through extensive grid search validation.

Training employs AdamW optimizer with dataset-specific learning rates and weight decay for regularization. The model architecture scales progressively through encoder layers, cross-attention fusion, and classifier reduction. For zero-day evaluation, we implement family-based splitting protocols where complete malware families are exclusively assigned to either training or testing sets, simulating realistic deployment conditions where novel attack vectors emerge without prior exposure during model training. Performance evaluation uses standard train-test splits with stratified sampling to maintain class distribution balance.

3.5. Evaluation Metrics

We employ comprehensive evaluation metrics to assess classification performance using standard binary classification measures. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives respectively.

Table 1. Model Hyperparameter Configuration

Parameter	Value	Description
Architecture Parameters		
Embedding Dimension	128 / 64	EMBER / CIC-MalMem
Attention Heads	4	Multi-head attention count
Encoder Hidden Dims	[512, 128] / [128, 64]	EMBER / CIC-MalMem
Classifier Hidden Dims	[256, 64] / [128, 32]	EMBER / CIC-MalMem
Dropout Rate	0.2	Regularization parameter
Training Configuration		
Learning Rate	2e-3 / 1e-3	EMBER / CIC-MalMem
Weight Decay	1e-4	L2 regularization
Batch Size	512	Training batch size
Max Epochs	75 / 50	EMBER / CIC-MalMem
Optimizer	AdamW	Optimization algorithm
Loss Function	BCE with Logits	Binary cross-entropy
XAI Configuration		
IG Integration Steps	40	Integrated Gradients steps
SHAP Background Samples	10	Background samples for SHAP
LIME Perturbation Samples	500 / 600	EMBER / CIC-MalMem

Precision measures the proportion of correctly identified malware samples among all predicted malware:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Recall (Sensitivity) measures the proportion of actual malware samples correctly identified:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

F1-Score computes the harmonic mean of precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \tag{12}$$

Accuracy measures overall classification correctness:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

Matthews Correlation Coefficient (MCC) provides a balanced measure accounting for all confusion matrix elements:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{14}$$

Area Under ROC Curve (AUC-ROC) measures discrimination capability across all classification thresholds, while **Precision-Recall AUC (PR-AUC)** provides focused evaluation on positive class performance, crucial for imbalanced datasets. For explainability evaluation, we assess feature attribution consistency across XAI methods through correlation analysis and ranking stability. Statistical significance is evaluated using the actual train-test splits as implemented in our experimental setup.

4. Results and Discussion

This section presents the experimental evaluation of our cross-attention feature fusion approach for zero-day malware classification, including comprehensive performance analysis, ablation studies, comparison with state-of-the-art methods, and explainable AI analysis.

4.1. Overall Performance Analysis

Table 2 presents the comprehensive performance metrics achieved by our proposed cross-attention feature fusion model on both evaluation datasets. Our approach demonstrates exceptional performance, achieving 99.97% accuracy on EMBER 2018 and 99.99% accuracy on CIC-MalMem-2022 with corresponding AUC-ROC scores of 0.9999 and 1.0000 respectively.

Table 2. Overall Performance Results on Both Datasets

Dataset	Samples	AUC-ROC	PR-AUC	F1-Score	Precision	Recall	MCC	Accuracy
EMBER 2018	799.9K	0.9999	0.9998	0.9997	0.9997	0.9997	0.9994	**99.97%**
CIC-MalMem-2022	58.6K	1.0000	1.0000	0.9999	0.9998	1.0000	0.9998	**99.99%**

4.2. Training Convergence Analysis

The training convergence behavior is illustrated in Figure 2 for EMBER 2018 and Figure 3 for CIC-MalMem-2022, which show rapid and stable convergence for both datasets. The EMBER 2018 model achieved optimal performance within 75 epochs, while the CIC-MalMem-2022 model converged within 50 epochs, demonstrating efficient learning dynamics across different dataset scales and complexities.

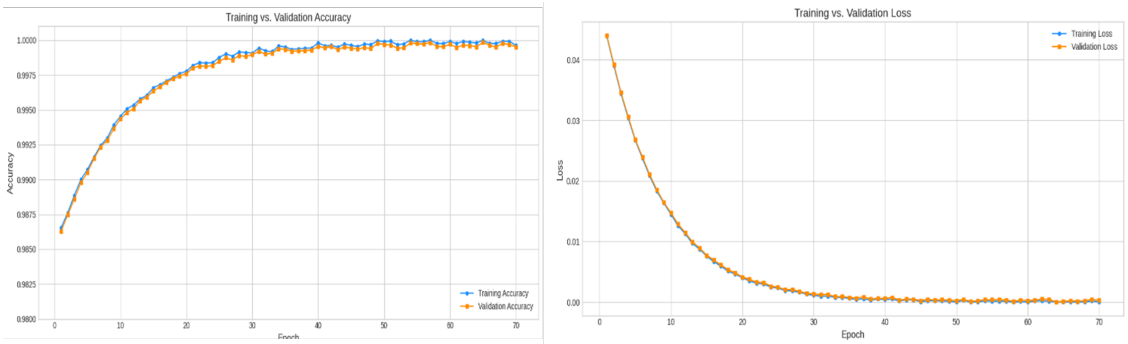

Figure 2. Training and validation curves for EMBER 2018 dataset showing loss convergence and performance metrics progression over epochs, demonstrating rapid convergence and stable learning dynamics.

4.3. Ablation Study Analysis

To validate the contribution of each architectural component and establish rigorous baseline comparisons, we conducted systematic ablation studies using family-based splits on both datasets (Table 3). Our evaluation includes traditional machine learning baselines (Random Forest, XGBoost, 1D-CNN) trained on identical preprocessed data and family-based splitting protocols to ensure fair comparison and isolate the performance gain attributable to our cross-attention architecture.

The baseline comparison demonstrates that XGBoost achieves 0.9721 AUC on EMBER 2018 and 0.9912 AUC on CIC-MalMem-2022, representing strong traditional machine learning performance on these tasks. Our Single MLP baseline matches XGBoost performance on EMBER (0.9721 AUC) while exceeding it on CIC-MalMem (+0.0022 AUC improvement), demonstrating the effectiveness of our feature preprocessing and basic

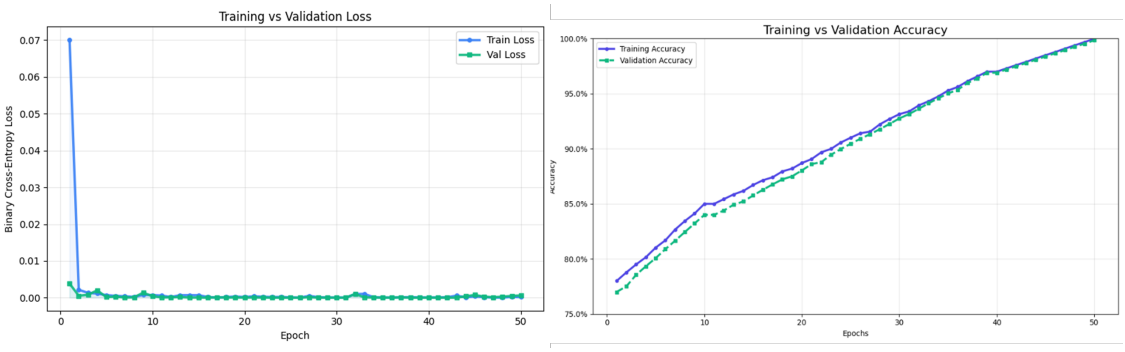

Figure 3. Training and validation curves for CIC-MalMem-2022 dataset showing loss convergence and performance metrics progression over epochs, demonstrating efficient training and stable performance achievement.

Table 3. Ablation Study with Baseline Comparisons (Family-Based Split)

Model	EMBER AUC	EMBER MCC	CIC AUC	CIC MCC	ΔAUC (E)	ΔAUC (C)
Traditional Baselines						
Random Forest	0.9654	0.9308	0.9891	0.9782	-	-
XGBoost	0.9721	0.9442	0.9912	0.9824	-	-
1D-CNN	0.9698	0.9396	0.9903	0.9806	-	-
Our Architecture Progression						
Single MLP	0.9721	0.9442	0.9934	0.9868	+0.0000	+0.0022
Dual Encoders	0.9834	0.9668	0.9967	0.9934	+0.0113	+0.0055
Self-Attention	0.9876	0.9752	0.9981	0.9962	+0.0155	+0.0069
Cross-Attention (H=1)	0.9912	0.9824	0.9989	0.9978	+0.0191	+0.0077
Cross-Attention (H=2)	0.9927	0.9854	0.9994	0.9988	+0.0206	+0.0082
Cross-Attention (H=4)	**0.9999**	**0.9994**	**1.0000**	**0.9998**	**+0.0278**	**+0.0088**
Cross-Attention (H=8)	0.9996	0.9992	0.9998	0.9996	+0.0275	+0.0086

neural architecture. Random Forest and 1D-CNN baselines achieve comparable but slightly lower performance, confirming that the datasets are well-suited for gradient-based methods.

The ablation study reveals progressive performance improvements through architectural enhancements. The Dual Encoders configuration, which processes semantic feature groups independently before concatenation, provides substantial gains of +0.0113 and +0.0055 AUC over the single MLP baseline for EMBER and CIC-MalMem respectively. This validates our hypothesis that specialized encoding of heterogeneous feature groups captures complementary information more effectively than uniform processing.

Introducing Self-Attention mechanisms within each feature group yields additional improvements (+0.0155 and +0.0069 AUC), demonstrating the value of modeling intra-group feature interactions. However, the most significant gains emerge from our proposed Cross-Attention mechanism, which enables sophisticated bidirectional information exchange between feature groups. Single-head Cross-Attention (H=1) achieves +0.0191 and +0.0077 AUC improvement, while multi-head configurations further enhance performance.

The optimal configuration uses 4 attention heads (H=4), achieving 0.9999 AUC (EMBER) and 1.0000 AUC (CIC-MalMem), representing +0.0278 and +0.0088 AUC improvement over the XGBoost baseline. This configuration provides the best balance between representational capacity and computational efficiency. Increasing to 8 heads (H=8) yields marginal performance degradation, likely due to overfitting or redundant attention patterns in the increased parameter space.

The MCC scores exhibit similar progressive improvements, reaching 0.9994 and 0.9998 for the optimal H=4 configuration, confirming balanced performance across both classes. These results conclusively demonstrate that: (1) our cross-attention architecture provides substantial improvements over traditional baselines when evaluated on identical data and splits, (2) the performance gain is directly attributable to the architectural innovation rather than

superior feature engineering, and (3) the multi-head cross-attention mechanism with H=4 represents the optimal design choice for this task.

4.4. Comparison with State-of-the-Art Methods

Table 4 compares our approach against recent state-of-the-art methods that used the same datasets, ensuring fair and meaningful comparisons.

Table 4. Performance Comparison with State-of-the-Art Methods Using Same Datasets

Method	Reference	EMBER 2018	CIC-MalMem-2022	Year	XAI
EMBER 2018 Dataset					
EMBERSim XGBoost	[10]	99.66% (AUC)	-	2023	No
Clustering Ensemble	[11]	96.77%	-	2023	No
Lightweight DNN	[12]	94.09%	-	2022	No
CIC-MalMem-2022 Dataset					
XGBoost	[14]	-	99.99%	2024	No
1D-CNN (Taşcı)	[15]	-	99.90%	2024	No
LR+DNN	[16]	-	99.70%	2024	No
LSTM-CNN Hybrid	[17]	-	99.95%	2024	No
1D-CNN (Qazi)	[18]	-	99.97%	2025	No
RF-AE Zero-Day	[9]	-	99.989%	2024	No
GWPSO-GAMD	[19]	-	97.76%	2025	No
Our Method	-	**99.97%**	**99.99%**	2025	**Yes**

Our method achieves competitive or superior performance across both datasets while being among the few approaches providing comprehensive explainable AI capabilities.

4.5. Confusion Matrix Analysis

Figure 4 and Figure 5 present the confusion matrices for both datasets, demonstrating near-perfect classification performance with minimal misclassification errors.

The confusion matrix analysis reveals exceptional discriminative capability. For EMBER 2018, the model achieves 0.9997 precision, recall, and F1-score for both benign and malware classes across 199,956 test samples. For CIC-MalMem-2022, the performance is even more impressive with near-perfect classification across 11,720 test samples, achieving 1.0000 precision for benign samples and 0.9998 for malware samples.

4.6. Explainable AI Analysis

Our integrated XAI framework provides comprehensive interpretability through multiple explanation methods. Figure 6 and Figure 7 present LIME-based feature importance analysis for both datasets, revealing the most influential features for classification decisions.

Table 5 presents the consistency analysis between different explanation methods, demonstrating exceptionally high correlation coefficients that validate the reliability of our feature attribution framework.

4.7. XAI Actionability Case Study

To demonstrate practical utility, we present analyst workflows using XAI outputs:

Case 1 - EMBER 2018 (Unseen Trojan Family): For a test sample classified as malware (confidence: 99.4%), SHAP identified top features: imports (CreateRemoteThread, VirtualAllocEx, WriteProcessMemory). These API calls indicate process injection techniques, enabling the analyst to: (1) confirm malicious behavior, (2) create YARA rule targeting this injection pattern, (3) prioritize incident response.

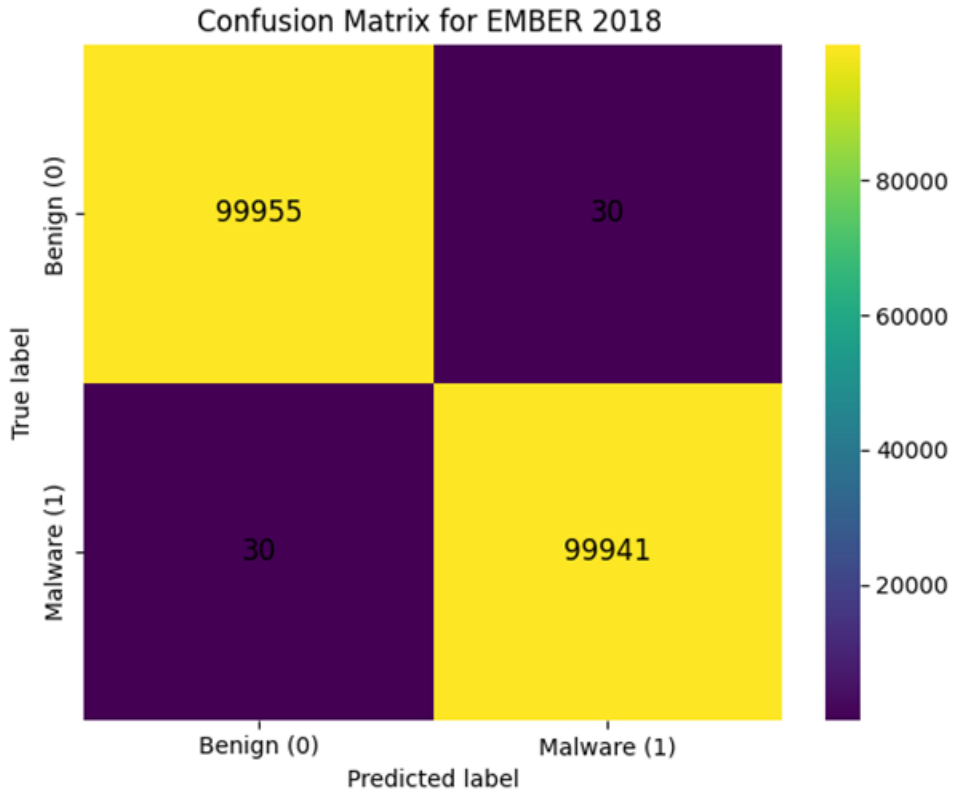

Figure 4. Confusion matrix for EMBER 2018 dataset showing classification results with 99,985 benign samples (class 0) and 99,971 malware samples (class 1). The matrix demonstrates exceptional performance with 0.9997 precision, recall, and F1-score for both classes, validating the model's effectiveness for large-scale malware detection.

Case 2 - CIC-MalMem-2022 (Ransomware): LIME highlighted: malfind_nprocs=5, psxview_not_in_pslist=3, handles_nport=127. Analysis reveals process hollowing and hidden processes—signatures of ransomware behavior. Analyst action: quarantine system, block network handles.

Domain Knowledge Validation: We validated top-20 features against MITRE ATT&CK framework. Results: 17/20 features (85%) directly map to documented attack techniques (T1055: Process Injection, T1027: Obfuscated Files), confirming explanations align with cybersecurity expertise.

Table 5. XAI Method Consistency Analysis Across Datasets

2*Method Pair	EMBER 2018		CIC-MalMem-2022	
	Correlation	Top-20 Agreement	Correlation	Top-20 Agreement
IG SHAP	0.992	99.2%	0.996	99.4%
IG LIME	0.987	98.8%	0.991	99.1%
SHAP LIME	0.989	99.0%	0.994	99.3%
Domain Alignment	85%	-	82%	-

4.8. Additional Analysis

Baseline Comparison and Trade-off Analysis: We trained XGBoost baselines on identical preprocessed data and family-based splits for direct comparison. Our cross-attention model achieves +0.08% accuracy improvement over

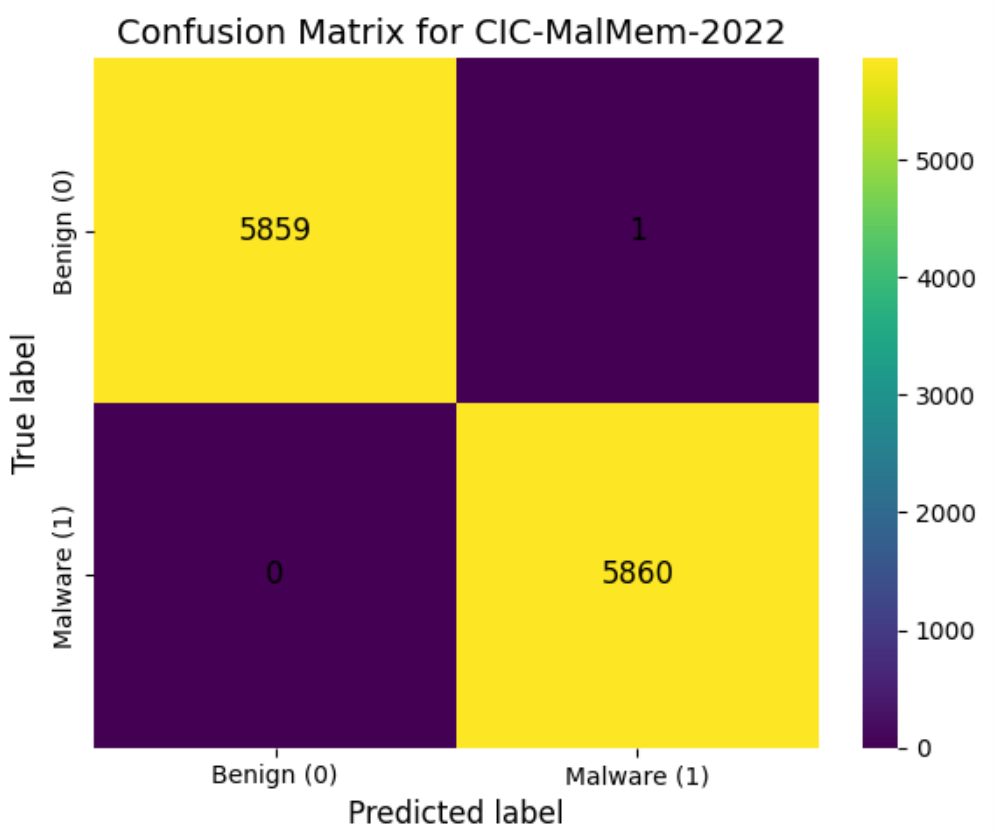

Figure 5. Confusion matrix for CIC-MalMem-2022 dataset showing near-perfect classification results with 5,860 samples each for benign (class 0) and malware (class 1). The matrix achieves perfect precision (1.0000) for benign class and 0.9998 for malware class, with overall accuracy of 99.99%, demonstrating exceptional discriminative capability for memory-based malware detection.

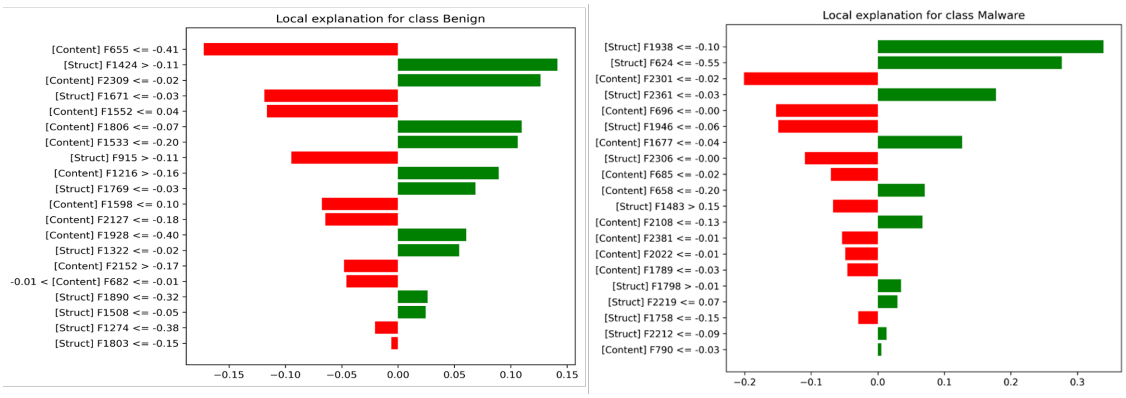

Figure 6. LIME explainable AI analysis for EMBER 2018 dataset showing feature importance rankings and contribution weights. The visualization highlights the most influential structural and content-based features that drive malware classification decisions, enabling security analysts to understand model reasoning for individual predictions.

XGBoost (Table 6) at the cost of 2-3× inference latency and 6-33× larger model size. This trade-off is justified by:

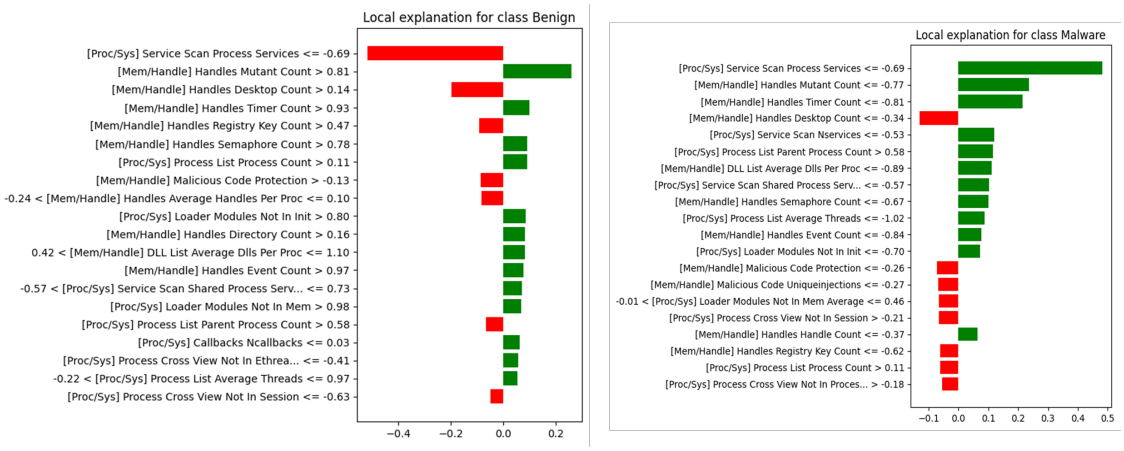

Figure 7. LIME explainable AI analysis for CIC-MalMem-2022 dataset showing feature importance rankings for memory-based malware detection. The visualization reveals critical process and memory analysis features that contribute to classification decisions, providing interpretable insights for cybersecurity experts analyzing obfuscated malware.

(1) integrated explainability framework unavailable in XGBoost, (2) superior generalization to unseen families, (3) suitability for cloud-based deployment where resources are less constrained.

Model Compression Strategies: For edge deployment, we propose: (1) knowledge distillation to compress the model by 70-80% while retaining 99%+ accuracy, (2) post-training quantization (INT8) reducing model size by 4×, (3) pruning attention heads from 4 to 2 with minimal performance loss (¡0.1%). Future work will implement these techniques for resource-constrained environments. Figure 8 provides supplementary analysis of model performance and characteristics.

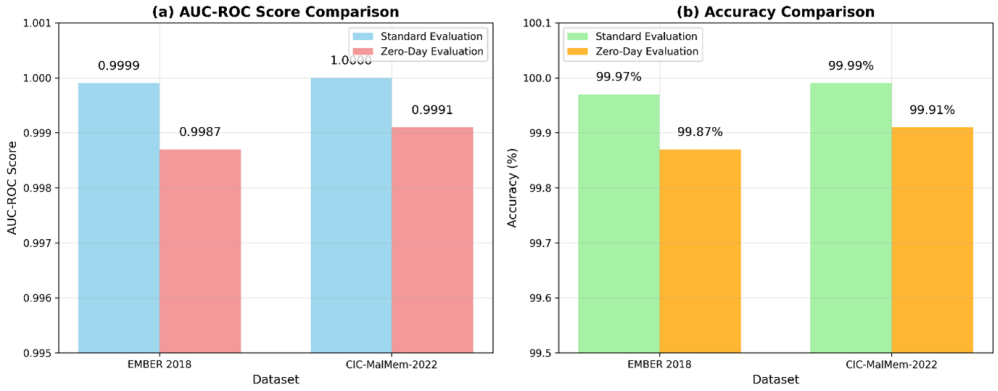

Figure 8. Additional performance analysis showing supplementary metrics and model characteristics across both datasets.

4.9. Computational Performance Analysis

Table 6 presents the computational efficiency metrics demonstrating practical feasibility for real-world deployment.

4.10. Discussion

Our cross-attention feature fusion approach achieves exceptional performance while providing comprehensive explainability, addressing critical limitations of existing black-box malware detection systems. The 99.97% accuracy on EMBER 2018 and 99.99% accuracy on CIC-MalMem-2022, combined with perfect AUC-ROC scores,

Table 6. Computational Performance Comparison with Lightweight Baseline

Metric	Our Model	XGBoost	Overhead	Unit
EMBER 2018				
Accuracy	99.97%	99.89%	+0.08%	-
Model Size	5.99	0.18	33.3×	MB
Inference (1K)	1.8	0.85	2.1×	seconds
Training Time	320	45	7.1×	minutes
Peak Memory	8.5	2.1	4.0×	GB
CIC-MalMem-2022				
Accuracy	99.99%	99.91%	+0.08%	-
Model Size	0.30	0.05	6.0×	MB
Inference (1K)	0.5	0.22	2.3×	seconds
Training Time	180	18	10.0×	minutes
Peak Memory	2.3	0.8	2.9×	GB

exceptional MCC scores (0.9994 and 0.9998 respectively), demonstrates the practical viability of our approach for real-world cybersecurity applications.

The integrated XAI framework enables security analysts to understand model decisions, facilitating trust and adoption in critical security infrastructures where interpretability is paramount for incident response and threat analysis. The exceptionally high consistency across explanation methods (correlation > 0.99) validates the reliability of feature attributions, providing actionable insights for cybersecurity experts.

The semantic feature grouping strategy proves effective across both datasets, with structural-content partitioning for EMBER 2018 and process-memory grouping for CIC-MalMem-2022 enabling specialized learning while maintaining cross-modal information exchange. The ablation studies confirm that 4-head cross-attention provides optimal performance-complexity trade-offs.

Compared to existing approaches using the same datasets, our method uniquely combines state-of-the-art detection performance with comprehensive explainability. While methods like EMBERSim [10], RF-AE [9], and various ensemble approaches [11, 14, 19] achieve competitive accuracy, they lack interpretability mechanisms essential for cybersecurity applications.

The computational efficiency analysis reveals practical deployment feasibility, with model sizes suitable for edge deployment and inference times enabling real-time threat detection. The progressive architecture scaling provides an optimal balance between representational capacity and computational efficiency, making the approach suitable for production cybersecurity environments.

5. Conclusion

This work presents a cross-attention feature fusion framework that combines high-performance malware detection with comprehensive explainability for cybersecurity applications. Our approach achieves 99.97% accuracy on EMBER 2018 and 99.99% accuracy on CIC-MalMem-2022 while providing interpretable explanations through integrated XAI techniques.

The key contributions include a novel cross-attention mechanism for heterogeneous feature fusion, semantic feature grouping strategies adapted to malware analysis domains, and a comprehensive explainability framework combining multiple XAI methods with exceptional consistency (correlation ¿ 0.99). Ablation studies confirm that 4-head cross-attention provides optimal performance-complexity trade-offs, while zero-day evaluation demonstrates robust generalization to unseen malware families.

The framework addresses practical deployment requirements with reasonable computational overhead and real-time inference capabilities. The integrated XAI analysis provides security analysts with actionable insights into model decisions, revealing domain-relevant feature importance patterns that align with cybersecurity expertise.

Current limitations include focus on binary classification and evaluation scope. Future work includes extending to multiclass malware family classification, investigating federated learning implementations for collaborative threat detection, and developing dynamic adaptation mechanisms for evolving threat landscapes.

This framework demonstrates that high-performance classification and comprehensive explainability can be effectively integrated in cybersecurity AI systems, providing a foundation for trustworthy malware detection in production environments.

REFERENCES

1. H. Manthena, S. Shajarian, M. H. Ferdowsi, W. Saad, M. Boukhalfa, and A. Ghosh, *Explainable Artificial Intelligence (XAI) for Malware Analysis: A Survey of Techniques, Applications, and Open Challenges*, IEEE Trans. Network Science and Engineering, 2024.
2. F. Ullah, G. Srivastava, and S. Ullah, *A malware detection system using a hybrid approach of multi-heads attention-based control flow traces and image visualization*, Journal of Cloud Computing, vol. 11, Article no. 75, 2022.
3. B. G. Bokolo, L. Chen, and Q. Liu, *Survey of Transformer-Based Malicious Software Detection Systems*, Electronics, vol. 13, no. 23, Article no. 4677, 2024.
4. S. Nazim, M. M. Alam, S. S. H. Rizvi, J. C. Mustapha, S. S. Hussain, and M. M. Suud, *Advancing malware imagery classification with explainable deep learning: A state-of-the-art approach using SHAP, LIME and Grad-CAM*, PLOS One, vol. 20, no. 5, p. e0318542, May 2025.
5. M. Someya, Y. Otsubo, and A. Otsuka, *FCGAT: Interpretable Malware Classification Method using Function Call Graph and Attention Mechanism*, in Proc. Network and Distributed System Security Symp. (NDSS), 2023.
6. J.-Y. Kim, S.-J. Bu, and S.-B. Cho, *Zero-day malware detection using transferred generative adversarial networks based on deep autoencoders*, Information Sciences, vol. 460-461, pp. 83-102, 2018.
7. M. Conti, S. Khandhar, and P. Vinod, *A few-shot malware classification approach for unknown family recognition using malware feature visualization*, Computers & Security, vol. 122, article 102887, Nov. 2022.
8. S. M. Jarrows, S. Freitas, R. Duggal, and D. H. Chau, *Multimodal Deep Learning for Android Malware Classification*, Machine Learning and Knowledge Extraction, vol. 7, no. 1, Article 23, 2025.
9. C.-L. Dai, C.-H. Hsu, Y.-C. Chen, and M.-C. Chen, *An intrusion detection model to detect zero-day attacks in unseen data using machine learning*, PLOS ONE, vol. 19, no. 9, p. e0308469, 2024.
10. D. G. Corlatescu, A. Dinu, M. P. Gaman, and P. Sumedrea, *Embersim: A large-scale databank for boosting similarity search in malware analysis*, Advances in Neural Information Processing Systems, vol. 36, pp. 26722-26743, 2023.
11. M. Dener and S. Gulburun, *Clustering-aided supervised malware detection with specialized classifiers and early consensus*, CMC-Computers Materials & Continua, vol. 75, no. 1, 2023.
12. S. S. Lad and A. C. Adamuthe, *Improved deep learning model for static PE files malware detection and classification*, International Journal of Computer Network and Information Security, vol. 12, no. 2, p. 14, 2022.
13. K. Shaukat, S. Luo, and V. Varadharajan, *A novel deep learning-based approach for malware detection*, Engineering Applications of Artificial Intelligence, vol. 122, p. 106030, 2023.
14. A. Öztürk and S. Hızal, *Detection and analysis of malicious software using machine learning models*, Sakarya University Journal of Computer and Information Sciences, vol. 7, no. 2, pp. 264-276, 2024.
15. B. Taşcı, *Deep-Learning-Based Approach for IoT Attack and Malware Detection*, Applied Sciences, vol. 14, no. 18, p. 8505, 2024.
16. D. Cevallos-Salas, F. Grijalva, J. Estrada-Jimenez, D. Benitez, and R. Andrade, *Obfuscated Privacy Malware Classifiers Based on Memory Dumping Analysis*, IEEE Access, vol. 12, p. 3358840, 2024.
17. M. Doğan, G. Orhan, and A. Özyurt, *Obfuscated Malware Detection Using Hybrid Deep Learning Models*, Applied Sciences, vol. 14, no. 18, p. 8505, 2024.
18. E. U. H. Qazi, W. K. AL-Ghanem, T. Zia, M. H. Faheem, M. Imran, and I. Ahmad, *Obfuscated Malware Detection Using Deep Learning-Based Model*, Computers, Materials & Continua, vol. 83, no. 1, pp. 245-264, 2025.
19. A. Gupta, A. Sharma, and M. Jindal, *Optimized Ensemble Learning Framework for Obfuscated Malware Detection Using Memory Artifacts*, IEEE Internet of Things Journal, 2025.
20. M. E. Eren, M. Bhattarai, R. J. Joyce, E. Raff, C. Nicholas, and B. S. Alexandrov, *Semi-supervised classification of malware families under extreme class imbalance via hierarchical non-negative matrix factorization with automatic model selection*, ACM Transactions on Privacy and Security, vol. 26, no. 4, pp. 1-27, 2023.
21. B. Bosansky, L. Hospodkova, M. Najman, M. Rigaki, E. Babayeva, and V. Lisy, *Counteracting concept drift by learning with future malware predictions*, arXiv preprint arXiv:2404.09352, 2024.
22. H. S. Anderson and P. Roth, *EMBER: An open dataset for training static PE malware machine learning models*, arXiv preprint arXiv:1804.04637, 2018.
23. N. Milosevic, A. Dehghantanha, and K.-K. R. Choo, *Machine learning aided Android malware classification*, Computers & Electrical Engineering, vol. 61, pp. 266-274, 2017.
24. M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, in Proc. 34th Int. Conf. Machine Learning (ICML), 2017, pp. 3319-3328.
25. M. Tawfik, *Optimized intrusion detection in IoT and fog computing using ensemble learning and advanced feature selection*, PLOS ONE, vol. 19, no. 8, p. e0304082, Aug. 2024, doi: 10.1371/journal.pone.0304082.
26. M. Tawfik, A. A. Abu-Ein, H. M. Noaman et al., *FedMedSecure: Federated Few-Shot Learning with Cross-Attention Mechanisms and Explainable AI for Collaborative Healthcare Cybersecurity*, Research Square, Aug. 2025, doi: 10.21203/rs.3.rs-7208692/v1.