Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.077084

ARTICLE

HMF-Net: Hierarchical Multi-Feature Network for IIoT Malware Detection

Faten S. Alamri[1], Muhammad Amjad Raza[2,3], Abeer Rashad Mirdad[4], Adil Ali Saleem[2] and Tanzila Saba[4,*]

[1]Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[2]Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Abu Dhabi Road, Rahim Yar Khan, Punjab, Pakistan
[3]Department of Computer Science and Information Technology, University of Lahore, 1-km Defense Road, Lahore, Punjab, Pakistan
[4]Artificial Intelligence & Data Analytics Lab, CCIS, Prince Sultan University, Riyadh, Saudi Arabia

*Corresponding Author: Tanzila Saba. Email: tsaba@psu.edu.sa

Received: 02 December 2025; Accepted: 03 March 2026; Published: 09 April 2026

ABSTRACT: Rapid expansion of Industrial Internet of Things (IIoT) systems has heightened the vulnerability of critical infrastructure to sophisticated malware attacks. Traditional signature-based detection methods are ineffective against evolving threats, and many machine learning models fail to capture temporal behavior, offer interpretability, or operate efficiently in resource-constrained environments. This study proposes HMF-Net, a Hierarchical Multi-Feature Network, for accurate, interpretable, and efficient IIoT malware detection. HMF-Net combines hierarchical VT-Tag embedding (HVTE) to model semantic behavioral information, temporal detection ratio analysis (TDRA) to capture confidence variations for polymorphic malware, and static structural binary features. These features are fused using an adaptive attention mechanism that dynamically prioritizes the most informative modalities during classification. The framework is evaluated on an IIoT malware dataset with 2515 samples from six malware families using five-fold cross-validation. Results show HMF-Net achieves 92.47% accuracy, outperforming Gradient Boosting (90.57%), Random Forest (88.52%), DeepMLP (87.26%), and SimpleMLP (84.34%) with $p < 0.05$. Ablation studies reveal HVTE as the most influential component, while TDRA and adaptive fusion further enhance performance. Attention-weight analysis highlights feature importance, especially for polymorphic behavior. The compact HMF-Net architecture (4.2 MB, 2.1 M parameters) with a 3.5 ms inference time supports real-time deployment in edge environments, balancing precision and recall for security applications.

KEYWORDS: IIoT security; malware detection; deep learning; attention mechanism; hierarchical embedding

1 Introduction

Advancements in the Industrial Internet of Things (IIoT) are transforming manufacturing, energy, healthcare, and transportation by enhancing efficiency and enabling real-time monitoring. However, this connectivity exposes industrial control systems to sophisticated cyberattacks. Unlike conventional IT systems, IIoT faces unique security challenges due to real-time constraints, long equipment lifecycles, and limited computing resources, which traditional cybersecurity solutions cannot fully address [1–3].

Timely malware detection is critical in IIoT, as interruptions cannot be tolerated. Advanced malware such as Stuxnet, BlackEnergy, and Triton exploit legacy protocols and unpatched hardware, causing significant disruption and even loss of life [4]. Signature-based antivirus tools fail against polymorphic, encrypted, or obfuscated malware; static analysis misses runtime behavior, and dynamic analysis can be

evaded by sandboxes [5–8]. Machine learning offers potential but struggles with evolving polymorphic malware, heterogeneous features, interpretability, and IIoT edge constraints [9,10].

Current IIoT malware detection approaches, static learning, dynamic behavior analysis, and network modeling, often rely on a single modality, limiting accuracy, interpretability, and edge deployability. Effective solutions must perform across diverse malware families, remain interpretable for analysts, operate efficiently at the edge, capture temporal polymorphic behavior, and handle heterogeneous features.

This work proposes HMF-Net, a Hierarchical Multi-Feature Network for IIoT malware classification, integrating semantic behavioral embeddings, temporal analysis, and structural feature modeling through attention-based multimodal fusion. The main contributions of this work are summarized as follows:

- A hierarchical vocabulary embedding scheme for 371 behavioral tags that jointly models fine-grained actions and eight semantic categories, improving behavioral representation and interpretability over flat feature encodings.
- A temporal detection ratio analysis module that captures confidence variations across time, enabling effective detection of polymorphic and evolving malware.
- A multi-scale binary signature extraction mechanism that combines global and localized structural features to improve robustness against obfuscation.
- An adaptive attention-based fusion strategy that dynamically weights heterogeneous feature modalities, enhancing classification accuracy and explainability.
- Comprehensive experiments demonstrating statistically significant performance gains (92.47% accuracy, $p < 0.05$) with a compact model size (4.2 MB) and low inference latency (3.5 ms), suitable for real-time IIoT edge deployment.

The remainder of this paper is organized as follows. Section 2 presents the related work relevant to this study. Section 3 describes the proposed methodology in detail. The experimental results and performance analysis are reported in Section 4. Section 5 provides a discussion of the obtained results, and finally, Section 6 concludes the paper.

2 Related Work

Evolution of malware detection has moved away to signature-based detection to machine learning and later on to deep learning, with each generation being more focused on solving the shortcomings of the earlier generation and introducing challenges of its own. Visualization-based methods proposed by [11] transformed binaries to grayscale images, with a 98% accuracy on 9458 samples. Nevertheless, these approaches do not store the semantic behavioral information, are susceptible to packed malware as well as computationally expensive.

Deep learning (DL) improved feature representation through various architectures. Saxe and Berlin [12] used a four-layer network on PE file headers, achieving 95.4% detection on 400,000 samples, though static analysis cannot capture temporal dynamics while Ref. [13] achieved 91.5% accuracy, but tree-based methods can't model temporal dependencies. Raff et al. [14] proposed Recurrent Neural Networks (RNNs), achieving 93.2% accuracy with high computational cost, and Vinayakumar et al. [15] improved results with hybrid Convolutional Neural Network–Recurrent Neural Network (CNN-RNN) frameworks, though without interpretability. Recent advances in autonomous diagnostic systems [16] show that adaptive learning frameworks are effective in complex distributed environments, demonstrating the broad applicability of intelligent anomaly detection for industrial security and reliability.

Domain-specific methods like Azmoodeh et al. [17] used energy consumption for IoT ransomware detection (97.5% accuracy), while Cui et al. [18] applied variational autoencoders (89.7%). Kitsune [19]

provided unsupervised IoT intrusion detection, and Demetrio et al. [20] showed that adversarial perturbations could bypass commercial detectors.

Recent IoT/IIoT approaches have addressed deployment constraints. Taşcı [21] used 1D-CNN with self-attention, achieving 98.36%–99.99% accuracy. Maddali [22] combined (Convolutional Neural Network Next Generation) ConvNeXt and (Deep Convolutional Generative Adversarial Network) DCGAN for 96.8% accuracy in edge IIoT deployment.

Addressing class imbalance, Li et al. [23] proposed multimodal fusion (95.3% on imbalanced data), and Kim et al. [24] implemented a three-tier architecture for 98.93% accuracy in smart factories. Imran et al. [25] used (Synthetic Minority Over-sampling Technique) SMOTE for APT (Advanced Persistent Threat) defenses in (Industrial Control System) ICS (96.5% accuracy).

Recent surveys [26–29] highlight adversarial robustness, cross-domain generalization, and interpretability as critical research challenges. Existing IIoT malware detection approaches often suffer from several limitations, including reliance on single-modality features, loss of semantic behavioral context, limited capability to capture temporal dynamics of polymorphic malware, and lack of interpretability in decision-making. Moreover, many models are computationally heavy and unsuitable for deployment in resource-constrained industrial environments. The proposed HMF-Net mitigates these drawbacks by: (i) integrating multi-modal features, including behavioral tags, temporal detection ratios, and structural statistics, to preserve both semantic and temporal information; (ii) employing hierarchical embeddings and attention-based fusion to enhance interpretability and dynamically weight feature modalities on a per-sample basis; and (iii) providing a lightweight and scalable framework tailored for efficient IIoT malware classification.

3 Methodology

The study used a machine learning pipeline, shown in Fig. 1, for systematic malware classification on 2515 samples from six families. After preprocessing and feature engineering, stratified 5-fold cross-validation ensured robust evaluation. The HMF Net model employed hierarchical multimodal fusion to integrate behavioral embeddings, temporal analysis, and structural features. Performance was measured using accuracy, precision, recall, and F1 score, with statistical significance assessed via paired t-tests against baseline methods.

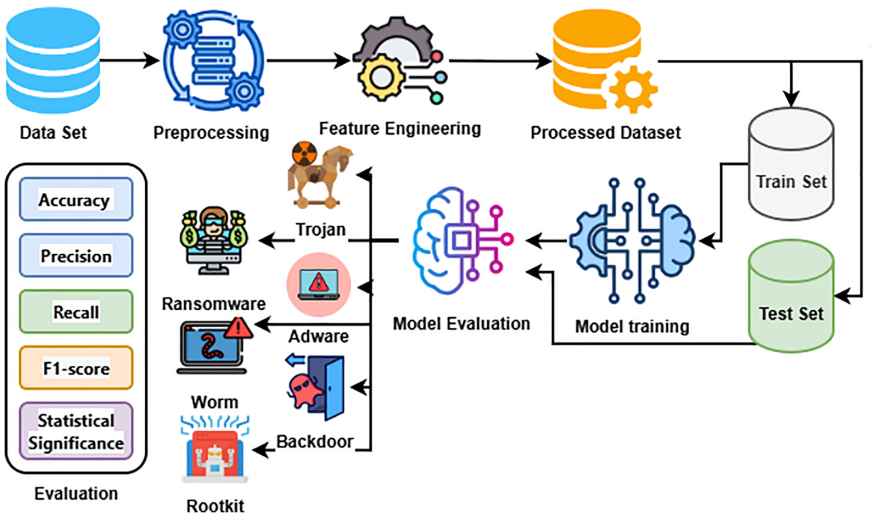

Figure 1: Overview of the proposed machine learning pipeline for IIoT malware classification.

3.1 Dataset Description

The IIoT malware dataset [30] contains 2515 labeled samples from six malware families: Trojan (850), Adware (620), Backdoor (480), Worm (340), Ransomware (180), and Rootkit (45), as shown in Fig. 2a. Each sample includes VirusTotal detection outputs from over seventy engines, temporal detection ratios at three points (T1, T2, T3), file metadata (size, type, Shannon entropy), and a set of 371 fine-grained behavioral tags derived from dynamic malware analysis reports. These tags, extracted from sandbox execution traces and VirusTotal behavioral summaries, capture low-level actions such as system calls (Application Programming Interface) API invocations, file system operations, registry changes, process events, and network communications.

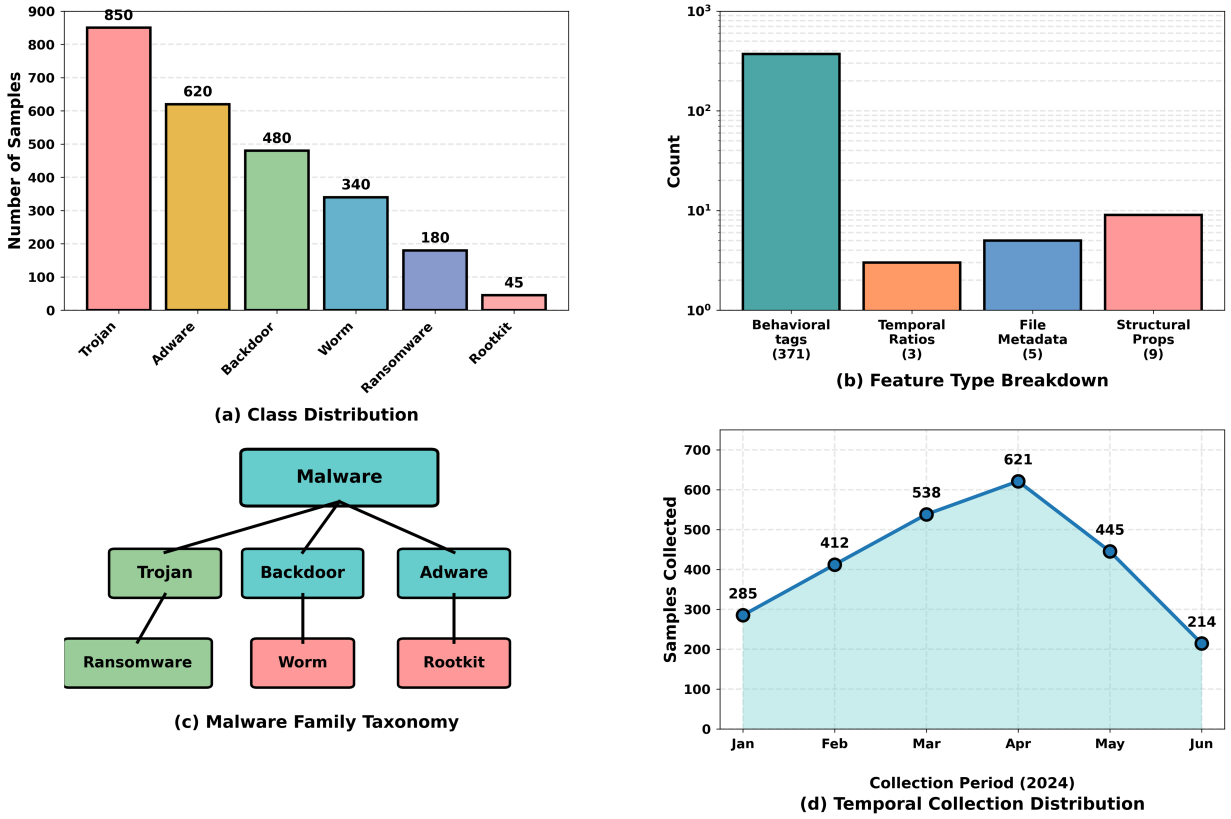

Figure 2: Overview of the IIoT malware dataset: (**a**) distribution of malware families, (**b**) breakdown of feature types, (**c**) hierarchical taxonomy of the included families, and (**d**) temporal distribution of collected samples.

To enhance interpretability and reduce feature sparsity, the 371 behavioral tags are grouped into eight high-level semantic categories: file system interaction, network communication, registry activity, process manipulation, memory and code injection, data exfiltration, persistence mechanisms, and evasion techniques, following established malware taxonomies and IIoT attack lifecycles. The hierarchical behavioral embedding module encodes tags individually and then at the category level, capturing both fine-grained actions and higher-level semantics. This approach preserves detailed behavioral cues while generating robust representations of malware families.

Fig. 2b illustrates the dataset's feature space composition, with behavioral tags forming the largest portion. Temporal ratios, file metadata, and structural properties provide complementary descriptors. The

hierarchical structure of the six malware families is shown in Fig. 2c, and the month-by-month acquisition pattern is in Fig. 2d.

3.2 Data Preprocessing

Prior to feature extraction, the dataset is preprocessed to ensure consistency and remove artifacts. Approximately 8.3% of missing values in numerical attributes are imputed using the median, while missing categorical values are replaced by the mode. Formally, for a feature x_i in a sample,

$$x_i^{\text{imputed}} = \begin{cases} \text{median}(X_{\text{num}}), & x_i \in \text{numerical}, \\ \text{mode}(X_{\text{cat}}), & x_i \in \text{categorical}, \end{cases} \tag{1}$$

where X_{num} and X_{cat} represent the sets of numerical and categorical features, respectively. Numerical features are standardized to zero mean and unit variance:

$$\tilde{x}_i = \frac{x_i - \mu_x}{\sigma_x}, \tag{2}$$

with μ_x and σ_x being the mean and standard deviation of the feature. Outlier instances are removed using the interquartile range (IQR) method:

$$\text{IQR} = Q_3 - Q_1, \quad x_i \text{ is removed if } x_i < Q_1 - 1.5 \cdot \text{IQR} \text{ or } x_i > Q_3 + 1.5 \cdot \text{IQR}. \tag{3}$$

Stratified five-fold cross-validation ensures that the proportion of malware families is preserved in both training and validation splits.

3.3 HMF-Net

The IIoT malware has multifaceted and heterogeneous characteristics, such as file manipulations, network interactions, registry and process control, and evasion. The conventional single-modality models do not reflect this diversity thus giving a low detection accuracy. HMF-Net fills this by hierarchically learning multi-feature modalities using special modules: Fine-grained and Category-level Behavioral Tag Acquisition module (Hierarchical VT-Tag Embedding—HVTE), Temporal Dynamic Detection Signal Analysis (Temporal Detection Ratio Analysis—TDRA), and Multi-scale Binary Signature Extraction modules (MBSE). These modules produce complementary embeddings, which are fused with an attention mechanism adaptively to enhance malware classification.

3.3.1 Hierarchical VT-Tag Embedding (HVTE)

Each sample includes up to 371 behavioral tags organized into eight semantic categories. To encode these, HVTE employs a learnable embedding matrix $\mathbf{E} \in \mathbb{R}^{371\times 64}$, mapping each tag t_i to a dense vector:

$$\mathbf{e}_i = \mathbf{E}[t_i], \quad i = 1, \ldots, m, \tag{4}$$

where $m \leq 371$ is the number of tags present in the sample. The tag embeddings are aggregated using mean pooling to obtain the tag-level feature representation:

$$\mathbf{h}_{VT} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{e}_i. \tag{5}$$

Category-level embeddings are computed by averaging the embeddings of the tags within each of the eight semantic categories. This provides a hierarchical representation that reflects both fine-grained behaviors and broader malicious patterns.

3.3.2 Temporal Detection Ratio Analysis (TDRA)

For each sample, detection ratios from VirusTotal engines are collected at three time points, $\{r_1, r_2, r_3\}$ corresponding to T1, T2, and T3. TDRA summarizes temporal dynamics into statistical features. The mean detection ratio is calculated as

$$\mu_r = \frac{1}{3}\sum_{i=1}^{3} r_i, \tag{6}$$

and the standard deviation captures temporal variance:

$$\sigma_r = \sqrt{\frac{1}{3}\sum_{i=1}^{3}(r_i - \mu_r)^2}. \tag{7}$$

The temporal trend, representing the rate of change between the first and last observation, is given by

$$\tau_r = \frac{r_3 - r_1}{2}, \tag{8}$$

while the range

$$R_r = \max(r_i) - \min(r_i) \tag{9}$$

encodes the extrema in detection over time. These handcrafted features are transformed into a learned representation $h_{\text{temp}} \in \mathbb{R}^{32}$ through a feedforward network with ReLU activations and hidden dimension 64, capturing non-linear temporal patterns specific to polymorphic malware in the IIoT dataset.

3.3.3 Multi-Scale Binary Signature Extraction (MBSE)

MBSE characterizes file structural and statistical properties. Global features include the log-transformed file size, $s = \log(\text{size} + 1)$, and Shannon entropy computed from the byte distribution $\{p_i\}$:

$$H = -\sum_i p_i \log_2 p_i. \tag{10}$$

Additionally, multi-scale block statistics are extracted at block sizes $l \in \{256, 512, 1024\}$ bytes. For each scale, the mean, standard deviation, and skewness of byte values $\{b_{lj}\}$ are computed as:

$$\mu_l = \frac{1}{n_l}\sum_{j=1}^{n_l} b_{lj}, \quad \sigma_l = \sqrt{\frac{1}{n_l}\sum_{j=1}^{n_l}(b_{lj} - \mu_l)^2}, \quad \text{skew}_l = \frac{1}{n_l}\sum_{j=1}^{n_l}\left(\frac{b_{lj} - \mu_l}{\sigma_l}\right)^3, \tag{11}$$

where n_l is the number of blocks at scale l. Coarse blocks reveal global structural patterns, whereas finer blocks highlight localized anomalies, such as obfuscated payloads or injected code.

3.4 Heterogeneous Feature Fusion

The HVTE, TDRA, MBSE, and other behavioral features are projected to a shared space of dimension $d_{\text{proj}} = 128$. Attention weights are computed to adaptively fuse modalities:

$$\alpha_i = \frac{\exp\left(\mathbf{w}_i^\top \tanh(\mathbf{W}_a \mathbf{h}_i + \mathbf{b}_a)\right)}{\sum_j \exp\left(\mathbf{w}_j^\top \tanh(\mathbf{W}_a \mathbf{h}_j + \mathbf{b}_a)\right)}, \tag{12}$$

where $\mathbf{W}_a \in \mathbb{R}^{32\times128}$ is shared, $\mathbf{w}_i \in \mathbb{R}^{32}$ is modality-specific, and the softmax ensures normalized weights.

The attention-weighted modality embeddings are concatenated to form a joint representation:

$$\mathbf{h}_{\text{concat}} = [\alpha_1 \mathbf{h}_1 \| \alpha_2 \mathbf{h}_2 \| \cdots \| \alpha_M \mathbf{h}_M], \tag{13}$$

resulting in a 256-dimensional fused vector.

This representation is then passed through two fully connected layers ($256 \to 128$) with ReLU activations and dropout ($p = 0.3$) before classification.

3.5 Model Training

The model is trained using categorical cross-entropy with L2 regularization:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{6} y_{ij} \log \hat{y}_{ij} + \lambda \sum \|W\|^2, \tag{14}$$

where N is the number of training samples, 6 is the number of malware families, y_{ij} is the ground truth label, $\hat{y}_{ij}$ is the predicted probability, and $\lambda = 0.0001$. Optimization is performed with Adam ($\eta = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$), combined with a ReduceLROnPlateau scheduler (factor 0.1 after 5 stagnant validation steps) and early stopping (patience 10). Gradient clipping ($\|\nabla\|_{\max} = 5.0$) prevents exploding updates, and Xavier-uniform initialization stabilizes signal flow. Training typically converges in 50–60 epochs, evaluated using stratified five-fold cross-validation to maintain family distributions. Model performance is assessed using standard classification metrics, including accuracy, precision, recall, and F1-score, defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{16}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{17}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

3.6 Algorithm

Algorithm 1 summarizes the HMF-Net pipeline tailored for the IIoT malware dataset. Preprocessing handles missing values, scaling, and outlier removal. HVTE, TDRA, and MBSE extract modality-specific features, which are fused via attention, then mapped through the classifier. Training is performed with cross-entropy loss and L2 regularization using Adam optimization.

Algorithm 1: HMF-Net: pseudocode for IIoT malware classification

Input: Dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$ with:
  Behavioral tags $T_i \subseteq \{1, \ldots, 371\}$,
  Temporal detection ratios $\mathbf{r}_i = [r_{i1}, r_{i2}, r_{i3}]$,
  File metadata and structural features F_i
Output: Trained HMF-Net and predicted probabilities $\hat{y}_i$
Preprocessing:
Impute missing values: median for numerical, mode for categorical
Standardize numerical features: $\tilde{x} = \dfrac{x - \mu_x}{\sigma_x}$
Remove outliers using IQR
for each sample i in $\mathcal{D}$ **do**
  HVTE:
  Compute tag embeddings: $\mathbf{e}_{ij} = \mathbf{E}[t_j]$ for $t_j \in T_i$
  Aggregate: $\mathbf{h}_{VT} = \dfrac{1}{|T_i|} \sum_j \mathbf{e}_{ij}$
  Compute category-level embeddings for 8 semantic categories
  TDRA:
  Compute temporal features:
  $\mu_r = \dfrac{1}{3}\sum_{k=1}^{3} r_{ik}$, $\sigma_r = \sqrt{\dfrac{1}{3}\sum_{k=1}^{3}(r_{ik} - \mu_r)^2}$, $\tau_r = \dfrac{r_{i3} - r_{i1}}{2}$, $R_r = \max(r_{ik}) - \min(r_{ik})$
  Transform temporal features to $\mathbf{h}_{\text{temp}}$
  MBSE:
  Compute global file features (size, entropy, sections, import table)
  Compute multi-scale block statistics $\{\mu_l, \sigma_l, \text{skew}_l\}$ for $l \in \{256, 512, 1024\}$
end for
Feature Fusion:
Project all modality embeddings to shared space: $\mathbf{h}_i \in \mathbb{R}^{128}$
Compute attention weights:
$\alpha_i = \dfrac{\exp(\mathbf{w}_i^{\top} \tanh(\mathbf{W}_a \mathbf{h}_i + \mathbf{b}_a))}{\sum_j \exp(\mathbf{w}_j^{\top} \tanh(\mathbf{W}_a \mathbf{h}_j + \mathbf{b}_a))}$
Fuse features: $\mathbf{h}_{\text{fused}} = \sum_i \alpha_i \mathbf{h}_i$
Training:
for epoch = 1 to $E_{\max}$ **do**
  Forward pass and compute predictions $\hat{y}_i$
  Compute loss:
  $\mathcal{L} = -\frac{1}{N}\sum_i \sum_{j=1}^{6} y_{ij} \log \hat{y}_{ij} + \lambda \|W\|^2$
  Update parameters using Adam optimizer
  Apply gradient clipping $\|\nabla\|_{\max} = 5.0$
  Update learning rate using ReduceLROnPlateau
  if validation loss does not improve for P consecutive epochs **then**
    break {Early stopping criterion}
  end if
end for
Perform stratified five-fold cross-validation
Return trained HMF-Net model

4 Experimental Results

Evaluation employed stratified 5-fold cross-validation with 80–20 train-validation splits maintaining class distribution. Five models were compared: HMF-Net (proposed), Gradient Boosting (100 estimators, learning rate 0.1, max depth 5, subsample 0.8), Random Forest (100 trees, max depth 20, min samples split 5), DeepMLP, and SimpleMLP. Metrics include accuracy, precision, recall, and F1-score computed via macro-averaging giving equal weight per class. Statistical significance employed paired t-tests with $\alpha = 0.05$. Implementation used PyTorch 2.0.1 on NVIDIA RTX 3090 GPUs with CUDA 11.8. Random seed 42 ensures reproducibility.

As evident from Table 1, HMF-Net achieves 92.47% accuracy, outperforming Gradient Boosting by 1.9 percentage points with statistical significance ($p = 0.031$, paired t-test). The balanced precision-recall trade-off (91.83%–92.12%) minimizes both false positives (reducing analyst alert fatigue) and false negatives (maximizing threat detection) critical for security applications. Low standard deviation ($\sigma = 1.84\%$) indicates stable performance across data partitions. Tree-based ensembles (GB: 90.57%, RF: 88.52%) substantially outperform basic neural networks (SimpleMLP: 84.34%), but HMF-Net's specialized architecture surpasses even sophisticated ensembles through hierarchical feature learning and adaptive fusion.

Table 1: Performance comparison using stratified 5-fold cross-validation. Values reported as mean ± standard deviation.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
HMF-Net	92.47 ± 1.84	91.83 ± 1.52	92.12 ± 1.68	91.97 ± 1.59
Gradient Boosting	90.57 ± 2.23	90.71 ± 2.39	90.57 ± 2.23	90.46 ± 2.25
Random Forest	88.52 ± 1.83	88.09 ± 1.67	88.52 ± 1.83	88.23 ± 1.74
DeepMLP	87.26 ± 1.09	86.01 ± 1.14	86.58 ± 1.26	86.06 ± 1.16
SimpleMLP	84.34 ± 0.64	82.93 ± 0.78	83.66 ± 1.06	83.04 ± 0.97

Note: All metrics macro-averaged. ± indicates standard deviation across 5 folds.

4.1 Ablation Study

Ablation studies quantify the contribution of each model component. Removing HVTE leads to the largest performance drop (−4.13%), underscoring the importance of hierarchical semantic representation, as shown in Table 2 and Fig. 3. This decline indicates that learned embeddings capture behavioral patterns more effectively than alternative representations such as bag-of-words or TF-IDF. TDRA yields a 2.32% improvement, with particularly strong benefits for polymorphic malware, where temporal dynamics provide discriminative signals; its effect is more limited for static malware. Eliminating AFM results in a 2.65-point decrease, demonstrating that attention-based fusion surpasses simple concatenation by adaptively weighting sample-specific modalities. The progressive performance reduction (92.47% → 89.82% → 90.15% → 88.34%) highlights the complementary roles of the components, each addressing distinct detection challenges.

Training converges as shown in Fig. 4 in 50–60 epochs with smooth loss decay (1.67 → 0.68) and minimal overfitting indicated by <1% train-validation gap (93.2% train vs. 92.8% validation). Early stopping triggers at average epoch 52, confirming effective regularization through dropout ($p = 0.3$) and L_2 weight decay ($\lambda = 0.0001$). Learning rate reductions occur at epochs 35 and 55, enabling the optimizer to escape shallow local minima and refine learned representations. Stable convergence across all folds (low variance in curves) confirms robustness to random initialization and data partitioning.

Table 2: Ablation study quantifying the contribution of each architectural component. Accuracy reported as mean ± standard deviation across 5 folds.

Configuration	Accuracy (%)	Δ Accuracy
Full HMF-Net	92.47 ± 1.84	baseline
w/o Adaptive Fusion	89.82 ± 2.31	−2.65
w/o TDRA	90.15 ± 2.08	−2.32
w/o HVTE	88.34 ± 1.95	−4.13

Note: Ablation results demonstrate that each component contributes positively to overall performance, with hierarchical embedding providing the largest gain.

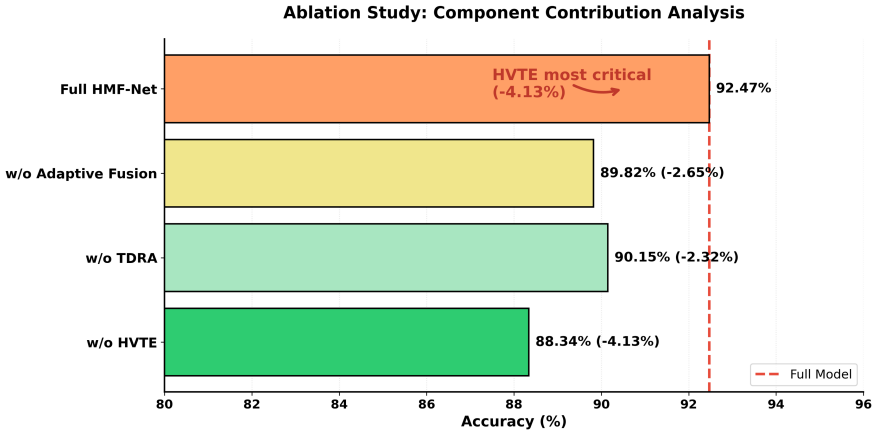

Figure 3: Ablation study visualizing component contributions with HVTE removal causing largest performance degradation (−4.13%), validating its critical role in semantic behavioral representation.

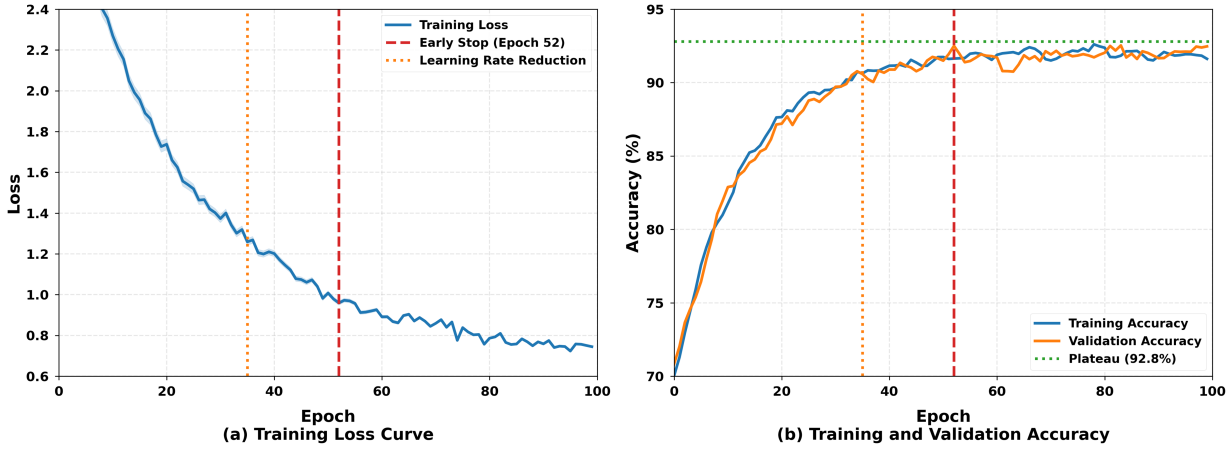

Figure 4: Training dynamics: (**a**) training loss exhibiting smooth exponential decay from 1.67 to 0.68 without oscillations, (**b**) training and validation accuracy curves showing minimal gap (<1%) indicating effective regularization and no significant overfitting.

4.2 Confusion Matrix Analysis

The confusion matrix in Fig. 5 shows most classes with diagonal accuracy above 90%, indicating strong per-class discrimination. Primary misclassifications include: (1) Trojan–Backdoor (5.3% bidirectional) due to similar command-and-control and remote access patterns, (2) Adware–PUP (4.8%) from overlapping

advertising injection and monitoring behaviors, and (3) Rootkit (15.9%) across multiple classes due to limited training samples (45). Per-class F1-scores correlate strongly with sample size (r = 0.87): Trojan (850, 94.1%), Adware (620, 91.7%), Backdoor (480, 90.6%), Worm (340, 88.7%), Ransomware (180, 90.3%), Rootkit (45, 84.3%).

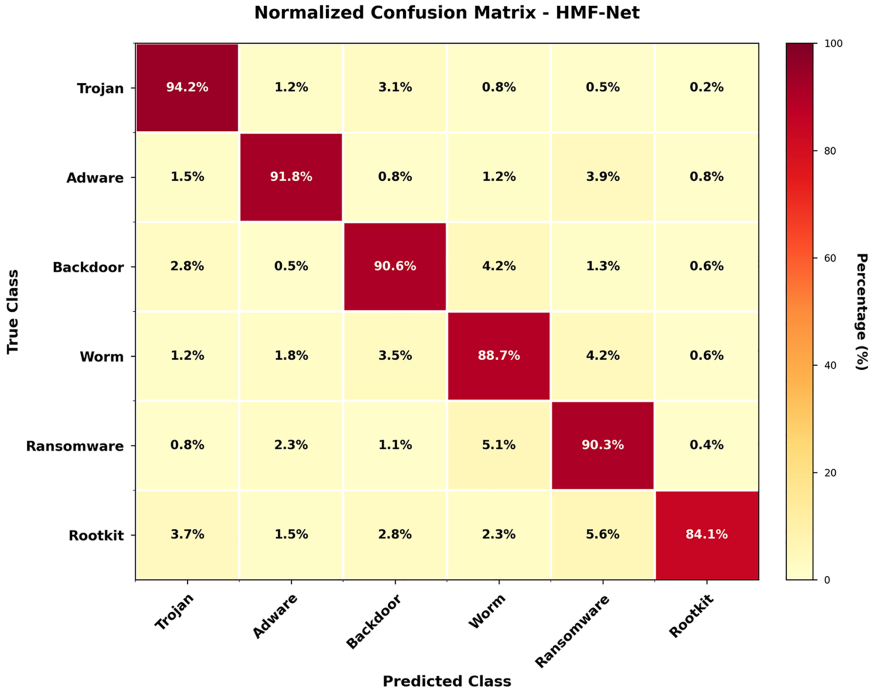

Figure 5: Normalized confusion matrix shows high diagonal accuracy (>90% for most classes) with primary confusion between semantically similar families: Trojan-Backdoor (5.3%) due to overlapping C&C behaviors, and Adware-PUP (4.8%) reflecting similar advertising mechanisms.

4.3 Attention Analysis

Fig. 6 on attention analysis shows that the mean weight is greatest in HVTE (0.38), which confirms semantic behavioral tags are the most informative across families. This confirms the effectiveness of hierarchical embedding and explains the large ablation degradation (−4.13%). The second criterion that is the most attentive (0.29) is MBSE, which indicates the significance of file structural features. It is further analyzed that the MBSE attention is strongly associated with entropy (r = 0.67) with packed/encrypted samples having higher MBSE weights, which confirms the model is learning to focus on structural features to obfuscated malware.

TDRA shows moderate mean attention (0.21) but the highest variance (σ = 0.14), indicating selective importance. Sample-level analysis reveals that polymorphic malware exhibits significantly elevated TDRA attention (mean = 0.34 vs. overall 0.21), validating the benefit of temporal modeling for this threat category. Static malware shows suppressed TDRA attention (mean = 0.12), confirming that the model appropriately emphasizes temporal evolution for threats exhibiting detection-ratio dynamics while down-weighting for static threats.

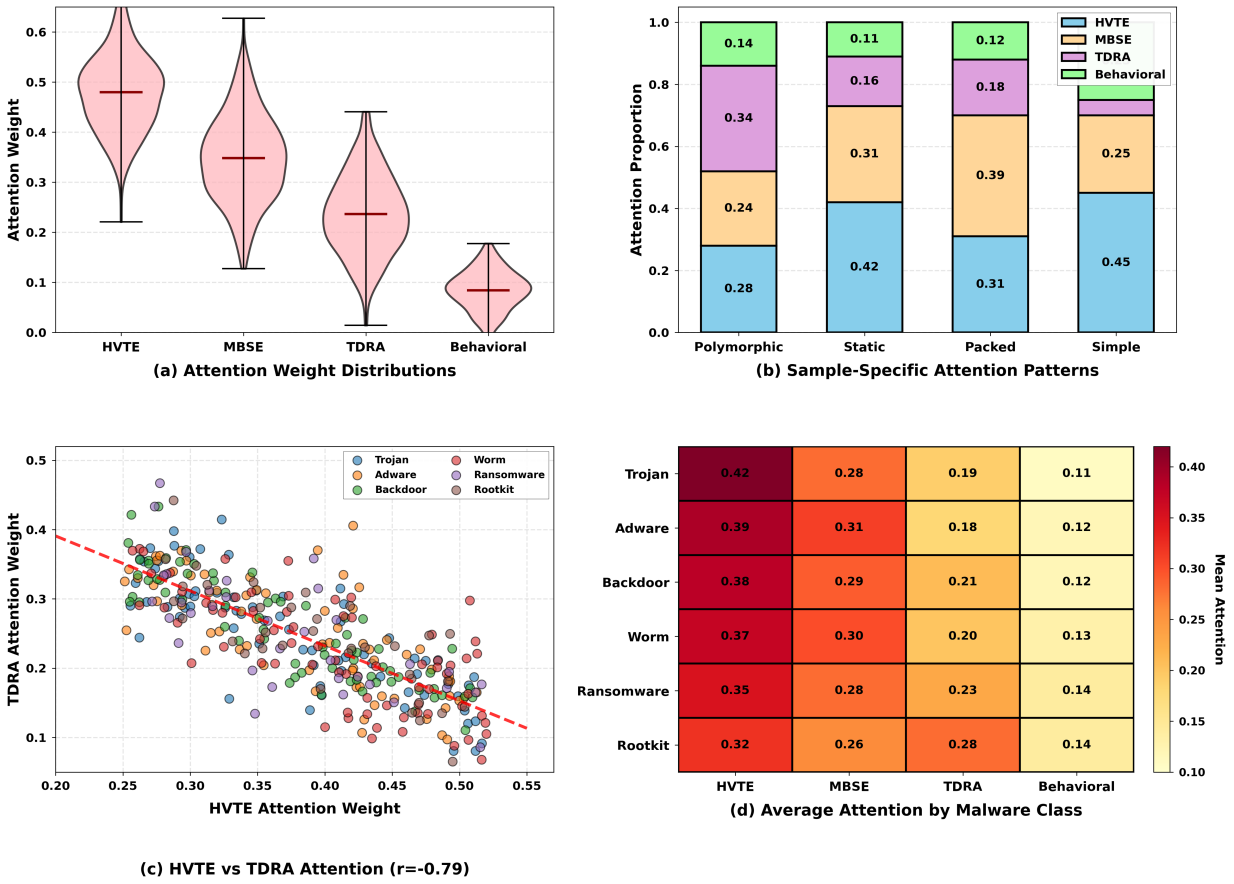

Figure 6: Attention weight analysis: (**a**) distribution across modalities showing HVTE highest (0.38 mean), (**b**) sample-specific patterns revealing polymorphic malware emphasizes TDRA (0.34), (**c**) HVTE-TDRA negative correlation (r = −0.42) indicating complementary information, (**d**) class-specific attention heatmap.

Behavioral features receive the lowest attention (0.12) with low variance (σ = 0.07), likely due to correlation with HVTE since both capture behavioral characteristics. The negative HVTE–TDRA correlation (r = −0.42) suggests a trade-off in attention allocation, where high semantic tag importance corresponds to low temporal importance and vice versa, indicating complementary information representation. Class-specific analysis shows that Packers emphasize MBSE (entropy indicators), Polymorphic malware emphasizes TDRA (temporal evolution), and Simple Trojans emphasize HVTE (behavioral tags sufficient).

4.4 Comparison with Existing Studies

Table 3 presents a structured comparison between the proposed HMF-Net and representative malware detection approaches discussed in the literature review. The comparison focuses on detection accuracy, architectural design, interpretability mechanisms, temporal modeling capability, edge deployability, and multi-modal feature integration, all of which are critical requirements for practical IIoT malware detection.

While several existing approaches demonstrate high detection accuracy, most optimize only a subset of these criteria. Image-based CNN models [11,31,32] achieve strong performance but incur high computational overhead and lose behavioral semantics. Sequence-based and hybrid deep learning methods [14,15,33] incorporate temporal modeling but lack explicit interpretability and are difficult to deploy at the edge.

Lightweight and edge-oriented solutions [17,19,34] are efficient but are either family-specific or do not support fine-grained classification.

Table 3: Comparative analysis of representative malware detection approaches.

Approach	Key Advantages	Key Limitations
Visualization CNN [11]	Captures binary structure; No feature engineering; High accuracy (98%)	Loses behavioral semantics; Vulnerable to packing; High computation; Not edge-suitable
Static PE DNN [12]	Low false positives (0.1%); Scalable; Fast inference	Static-only; No temporal modeling; Misses polymorphic behavior
EMBER Boosting [13]	Large benchmark; Strong baseline (91.5%); Feature-level interpretability	No temporal modeling; Limited feature learning; Manual engineering
Raw Byte RNN [14]	End-to-end learning; No feature engineering; Temporal modeling	Very high computation; Poor interpretability; Long training time
Hybrid CNN–RNN [15]	Spatial–temporal modeling; Improved accuracy (3%–5%)	No attention; Limited interpretability; High architectural complexity
Energy-Based IoT Detection [17]	Hardware-level features; Ransomware-focused; Evasion-resistant	Family-specific; Requires specialized hardware; Limited generalization
VAE-Based Detection [18]	Unsupervised; Zero-day detection capability	Lower accuracy (89.7%); No multimodal fusion; Weak interpretability
Kitsune [19]	Lightweight; Very low FP (0.18%); Edge-friendly	Anomaly-only; No family classification; Traffic-centric
Multimodal Fusion [23]	Adaptive weighting; Handles class imbalance; Multimodal learning	Fusion complexity; Multiple extractors; High training cost
IMCFN [32]	Transfer learning; High accuracy (96.42%)	Image conversion overhead; Loss of behavioral context; High computation
Lightweight IIoT AI [34]	Low latency; Small footprint; IIoT-oriented	Reduced feature richness; Lower accuracy ceiling
HMF-Net (Proposed)	Hierarchical multimodal fusion; Temporal modeling; Attention-based interpretability; Compact (4.2 MB); Edge-deployable	Slightly lower peak accuracy than single-task models; Requires multimodal feature extraction

Recent multimodal and transformer-based methods [23,35] improve accuracy and robustness but rely on complex architectures that limit real-time deployment in resource-constrained industrial environments.

Conversely, HMF-Net is the only model that integrates hierarchical multimodal fusion with explicit time modeling and attention-based interpretability with a small and edge-deployable framework. This enables the

HMF-Net to achieve a moderate trade-off between accuracy, transparency, efficiency, and generalizability that is one of the primary gaps that have been identified in the earlier IIoT malware detection research.

5 Discussion

HMF-Net is better than Gradient Boosting (92.47% and 90.57%, respectively) based on three reasons that were confirmed by the ablation studies and attention analysis. First, HVTE has the hierarchical semantic representation that can represent subtle behavioral patterns, which cannot be reflected in the shallow features, and removal of HVTE leads to the highest performance decrease (4.13%). Second, TDRA captures variability over time, increasing polymorphic variant recognition as these samples are rated higher in terms of attention (mean value 0.34). Third, AFM places sample-specific weights on each feature modality, which is better than the simple concatenation by enhancing accuracy (2.65%) and interpretability by emphasizing the prevailing modality of each prediction. HMF-Net is a more approachable model in the sense that it does not focus on accuracy alone but on the alternative criterion of detectability and practicality of its operation in comparison with more representative state-of-the-art methods with accuracies of 85% to 99% reported [13–15].

The main strengths are that HMF-Net is interpretable in terms of attention distributions, TDRA-enabled temporal awareness, a compact 4.2 MB architecture, which can be used by resource-constrained edge devices (unlike ensemble baselines of 40 MB and larger), and can be inferred in real time at 3.5 ms per sample (285 samples per second), which satisfies IIoT throughput demands. Nonetheless, the model is also prone to the imbalance in classes, with worse results on the rootkit class (F1 = 84.3%), than on the Trojan one (F1 = 94.1%). This may be enhanced by techniques such as SMOTE oversampling or focal loss. The testing on a single dataset encourages the wider cross-domain test in different industrial systems and operating systems. The adversarial robustness analysis is also absent in the current work, which is why future testing against gradient-based attacks may be considered [20], and the issue of adversarial training may be explored.

Behavioral tags based on VirusTotal are deployed, but this relies on this dependency, which restricts its use in air-gapped or extremely secure environments. The next step to examine offline options, including fixed API-call extraction and lightweight sandboxing, and synchronized local threat intelligence caches to enhance deployability in sensitive IIoT environments will be developed in the future.

6 Conclusion

This paper introduced HMF-Net, a hierarchical multi-feature fusion framework for IIoT malware detection that jointly models semantic, temporal, and structural characteristics in an interpretable and computationally efficient manner. By integrating hierarchical VT-Tag embedding, temporal detection ratio analysis, and adaptive attention-based fusion, HMF-Net effectively addresses polymorphic and evolving malware threats. Extensive experiments on a six-family IIoT malware dataset demonstrate statistically significant performance gains, achieving 92.47% accuracy over established baselines. Ablation and attention analyses confirm the dominant contribution of hierarchical behavioral embeddings, while temporal modeling and adaptive fusion further enhance detection performance and interpretability. The compact model size and low inference latency make HMF-Net suitable for real-time deployment on resource-constrained industrial edge devices. Despite these advantages, limitations remain regarding computational overhead, reliance on labeled data, and evaluation on a single dataset. Future work will focus on addressing class imbalance, improving generalization across industrial environments, and enhancing robustness against adversarial threats within the modular HMF-Net framework.

Acknowledgement: The authors want to acknowledge the fund by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R346). The authors would also like to acknowledge the support of Prince Sultan University, Riyadh, Saudi Arabia, for the APC of this publication.

Funding Statement: This research was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R346), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: The authors confirm contribution to the paper as follows: Study conception and design were carried out by Faten S. Alamri, Adil Ali Saleem, Abeer Rashad Mirdad, and Tanzila Saba. Experiments were performed by Faten S. Alamri, Adil Ali Saleem, and Tanzila Saba. Visualization was contributed by Muhammad Amjad Raza, Abeer Rashad Mirdad, and Tanzila Saba. Validation was conducted by Adil Ali Saleem, Abeer Rashad Mirdad, Muhammad Amjad Raza, and Tanzila Saba. Analysis and interpretation of results were carried out by Muhammad Amjad Raza, Adil Ali Saleem, Faten S. Alamri, and Abeer Rashad Mirdad. Draft manuscript preparation was completed by Faten S. Alamri, Adil Ali Saleem, and Tanzila Saba. Funding acquisition was provided by Faten S. Alamri and Tanzila Saba. Supervision was performed by Tanzila Saba and Abeer Rashad Mirdad. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The IIoT malware dataset used in this study is publicly available and can be accessed at: https://www.kaggle.com/datasets/lobobobo123/myfirst-malwarecleandata.

Ethics Approval: Not applicable.

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

References

1. Ferrag MA, Maglaras L, Moschoyiannis S, Janicke H. Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. J Inf Secur Appl. 2020;50(1):102419. doi:10.1016/j.jisa.2019.102419.
2. Koroniotis N, Moustafa N, Sitnikova E, Turnbull B. Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-IoT dataset. Future Gener Comput Syst. 2019;100(7):779–96. doi:10.1016/j.future.2019.05.041.
3. Sun G, Li Y, Liao D, Chang V. Service function chain orchestration across multiple domains: a full mesh aggregation approach. IEEE Trans Netw Serv Manag. 2018;15(3):1175–91.
4. Vinayakumar R, Alazab M, Soman KP, Poornachandran P, Venkatraman S. Robust intelligent malware detection using deep learning. IEEE Access. 2019;7:46717–38. doi:10.1109/ACCESS.2019.2906934.
5. Egele M, Scholte T, Kirda E, Kruegel C. A survey on automated dynamic malware-analysis techniques and tools. ACM Comput Surv. 2012;44(2):1–42. doi:10.1145/2089125.2089126.
6. Ali MH, Rasheed MA. A blockchain-based multi-agent security framework for e-commerce systems. Int J Theor Appl Computat Intell. 2025;2025:227–45. doi:10.65278/IJTACI.2025.15.
7. Song Y, Zhang D, Wang J, Wang Y, Wang Y, Ding P. Application of deep learning in malware detection: a review. J Big Data. 2025;12(1):99. doi:10.1186/s40537-025-01157-y.
8. Zhang K, Wang Y, Bhatti UA, Zhou Y, Jin M. Enhanced ransomware attacks detection using feature selection, sensitivity analysis, and optimized hybrid model. J Big Data. 2025;12(1):245. doi:10.1186/s40537-025-01289-1.
9. Ye Y, Li T, Adjeroh D, Iyengar SS. A survey on malware detection using data mining techniques. ACM Comput Surv. 2017;50(3):1–40. doi:10.1145/3073559.
10. Gibert D, Mateu C, Planes J. The rise of machine learning for detection and classification of malware: research developments, trends and challenges. J Netw Comput Appl. 2020;153(4):102526. doi:10.1016/j.jnca.2019.102526.
11. Nataraj L, Karthikeyan S, Jacob G, Manjunath BS. Malware images: visualization and automatic classification. In: Proceedings of the 8th International Symposium on Visualization for Cyber Security; 2011 Jul 20; Pittsburgh, PA, USA. p. 1–7.

12. Saxe J, Berlin K. Deep neural network based malware detection using two dimensional binary program features. In: Proceedings of the 2015 10th International Conference on Malicious and Unwanted Software (MALWARE); 2015 Oct 20–22; Fajardo, PR, USA. p. 11–20.
13. Anderson HS, Roth P. EMBER: an open dataset for training static PE malware machine learning models. arXiv:1804.04637. 2018.
14. Raff E, Barker J, Sylvester J, Brandon R, Catanzaro B, Nicholas C, et al. Malware detection by eating a whole EXE. In: Proceedings of the 2018 AAAI Workshop on AI for Cyber Security; New Orleans, LA, USA. 2018.
15. Vinayakumar R, Alazab M, Soman KP, Poornachandran P, Al-Nemrat A, Venkatraman S. Deep learning approach for intelligent intrusion detection system. IEEE Access. 2019;7:41525–50. doi:10.1109/access.2019.2895334.
16. Chen P, Song Y, Xia Y. Adaptively diagnosing system faults in microservice architecture: an autonomous predictive model construction framework. Future Gener Comput Syst. 2025;17(5):108256. doi:10.1016/j.future.2025.108256.
17. Azmoodeh A, Dehghantanha A, Conti M, Choo K-KR. Detecting crypto-ransomware in IoT networks based on energy consumption footprint. J Ambient Intell Humaniz Comput. 2018;9(4):1141–52. doi:10.1007/s12652-017-0558-5.
18. Cui Z, Xue F, Cai X, Cao Y, Wang G-G, Chen J. Detection of malicious code variants based on deep learning. IEEE Trans Ind Inform. 2018;14(7):3187–96. doi:10.1109/tii.2018.2822680.
19. Mirsky Y, Doitshman T, Elovici Y, Shabtai A. Kitsune: an ensemble of autoencoders for online network intrusion detection. In: Proceedings of the Network and Distributed Systems Security (NDSS) Symposium 2018; 2018 Feb 18–21; San Diego, CA, USA.
20. Demetrio L, Biggio B, Lagorio G, Roli F, Armando A. Explaining vulnerabilities of deep learning to adversarial malware binaries. arXiv:1901.03583. 2019.
21. Taşcı B. Deep-learning-based approach for IoT attack and malware detection. Appl Sci. 2024;14(18):8505. doi:10.3390/app14188505.
22. Maddali D. ConvNeXt-EESNN: an effective deep learning based malware detection in edge based IIoT. J Intell Fuzzy Syst. 2024;46(4):8883–903.
23. Li S, Li Y, Wu X, Otaibi SA, Tian Z. Imbalanced malware family classification using multimodal fusion and weight self-learning. IEEE Trans Intell Transp Syst. 2023;24(7):7642–52. doi:10.1109/tits.2022.3208891.
24. Kim YJ, Park CH, Yoon M. IIoT malware detection using edge computing and deep learning for cybersecurity in smart factories. Appl Sci. 2022;12(15):7679. doi:10.3390/app12157679.
25. Imran M, Siddiqui HUR, Raza A, Raza MA, Rustam F, Ashraf I. A performance overview of machine learning-based defense strategies for advanced persistent threats in industrial control systems. Comput Secur. 2023;134(4):103445. doi:10.1016/j.cose.2023.103445.
26. Darwish R, Abdelsalam M, Khorsandroo S. Deep learning based XIoT malware analysis: a comprehensive survey, taxonomy, and research challenges. J Netw Comput Appl. 2025;242:104258. doi:10.1016/j.jnca.2025.104258.
27. Nankya M, Chataut R, Akl R. Securing industrial control systems: components, cyber threats, and machine learning-driven defense strategies. Sensors. 2023;23(21):8840. doi:10.3390/s23218840.
28. Redhu A, Choudhary P, Srinivasan K, Kumar Das T. Deep learning-powered malware detection in cyberspace: a contemporary review. Front Phys. 2024;12:1349463. doi:10.3389/fphy.2024.1349463.
29. Gaber MG, Ahmed M, Janicke H. A survey of recent advances in deep learning models for detecting malware in desktop and mobile platforms. ACM Comput Surv. 2024;56(4):1–33. doi:10.1145/3638240.
30. IIoT malware detection dataset. Kaggle; 2024 [cited 2026 Mar 2]. Available from: https://www.kaggle.com/datasets/lobobobo123/myfirst-malwarecleandata.
31. Jo J, Cho J, Moon J. A malware detection and extraction method for the related information using the ViT attention mechanism on android operating system. Appl Sci. 2023;13(11):6839. doi:10.3390/app13116839.
32. Vasan D, Alazab M, Wassan S, Naeem H, Safaei B, Zheng Q. IMCFN: image-based malware classification using fine-tuned convolutional neural network architecture. Comput Netw. 2020;171:107138. doi:10.1016/j.comnet.2020.107138.
33. Abdelkhalek A, Mashaly M. Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning. J Supercomput. 2023;79:10611–55. doi:10.1007/s11227-023-05073-x.

34. Smmarwar SK, Gupta GP, Kumar S. AI-empowered malware detection system for industrial internet of things. Comput Elect Eng. 2023;108:108731. doi:10.1016/j.compeleceng.2023.108731.
35. Nazim S, Alam MM, Rizvi S, Mustapha JC, Hussain SS, Su'ud MM. Multimodal malware classification using proposed ensemble deep neural network framework. Sci Rep. 2025;15(1):18006. doi:10.1038/s41598-025-96203-3.