applied sciences

MDPI

Article

HARMONI: A Two-Stage Hybrid Learning Framework with Dynamic Metric Learning for Interpretable NIDS

Rongxin Hu [1], Zhiqiang Zhang [2,3], Minhao Li [1], Youwen Wen [1] and Le Wang [1,*]

[1] Cyberspace Institute of Advanced Technology, Huangpu Research School, Guangzhou University, Guangzhou 510006, China; hurongxin@e.gzhu.edu.cn (R.H.); 2112333066@e.gzhu.edu.cn (M.L.); wenyouwen@e.gzhu.edu.cn (Y.W.)
[2] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China; zhangzhiqiang@stu.hit.edu.cn
[3] Department of New Networks, Pengcheng Laboratory, Shenzhen 518000, China
* Correspondence: wangle@gzhu.edu.cn

Abstract

The increasing sophistication of cyber attacks has created a growing demand for effective Network Intrusion Detection Systems (NIDSs). Although deep learning has improved NIDS performance, existing models often lack adaptive mechanisms for spatiotemporal feature fusion and struggle with complex traffic distributions characterized by severe intra-class heterogeneity and inter-class overlap. Meanwhile, current interpretability methods mainly rely on feature importance analysis and provide limited insight into the model's decision process. To address these challenges, we propose HARMONI, a two-stage hybrid learning framework that enhances both detection accuracy and model interpretability. In the first stage, a dual-branch Convolutional Neural Network and Gated Recurrent Unit (CNN-GRU) architecture extracts spatiotemporal features, which are dynamically fused through a lightweight adaptive gating network. The representation learning process is jointly optimized using a Dynamic Class-Center Loss to enforce intra-class compactness and inter-class separability in the latent space. In the second stage, the learned deep representations are concatenated with raw traffic features and fed into an ensemble classifier. This residual-style design mitigates information loss during deep encoding while leveraging the non-linear modeling capability of ensemble learning. We further develop a multi-level interpretability framework based on SHapley Additive exPlanations (SHAP) that analyzes global feature importance, individual feature contributions, and feature interactions to provide quantitative insights into the model's decision mechanisms. Experiments on four benchmark datasets show that HARMONI consistently outperforms state-of-the-art baselines, achieving 80.19% and 78.24% accuracy on NSL-KDD and UNSW-NB15 respectively, surpassing representative deep learning and ensemble methods.

Keywords: Network Intrusion Detection; spatiotemporal feature fusion; dynamic metric learning; explainable AI (XAI)

Check for updates

Academic Editor: Rui Araújo

Received: 3 April 2026
Revised: 28 May 2026
Accepted: 24 June 2026
Published: 30 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid expansion of digital infrastructure has significantly increased the scale and complexity of cyber threats, posing serious risks to enterprise ecosystems and Critical National Infrastructure (CNI) [1,2]. To mitigate these risks, modern cybersecurity architectures have gradually evolved toward a "protection–detection–response" paradigm, as illustrated in Figure 1, which serves as a fundamental framework for systematic cyber

defense [3]. Within this architecture, the Network Intrusion Detection System (NIDS) plays a central role. By continuously monitoring and analyzing high-dimensional network traffic, the NIDS enables the detection of malicious activities that bypass traditional perimeter defenses, thereby providing essential visibility for timely mitigation and response [4,5].

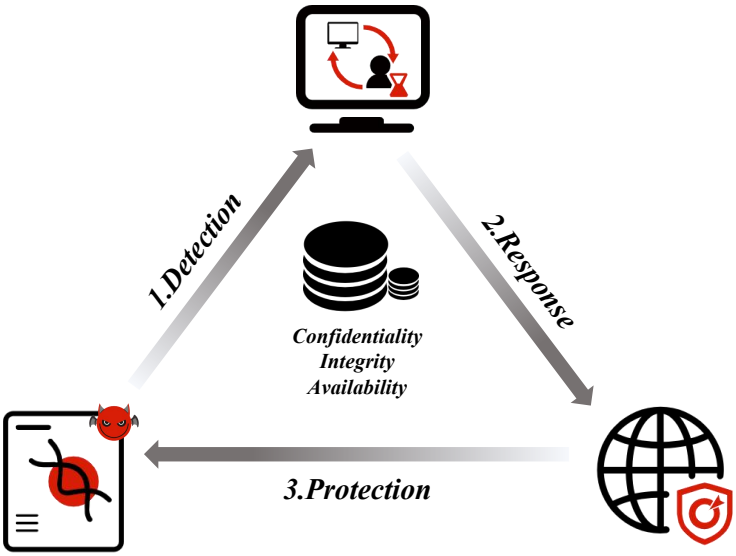


Figure 1. Protection–Detection–Responseparadigm.

Driven by the increasing demand for effective traffic analysis, deep learning (DL) has become a dominant paradigm for modern NIDSs. Existing approaches commonly employ Convolutional Neural Networks (CNNs) to extract spatial features from traffic data, while Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) and the Gated Recurrent Unit (GRU), are widely adopted to capture temporal dependencies [6–9]. More recent studies have explored hybrid architectures, attention mechanisms, and Transformer-based models to learn richer spatiotemporal representations and improve detection performance [10–13]. Despite these advances, existing methods still face significant challenges when applied to real-world network environments characterized by complex and non-stationary traffic distributions. First, commonly used spatiotemporal fusion strategies, such as heuristic concatenation or simple averaging, lack adaptive mechanisms for dynamically weighting heterogeneous feature representations, which limits the expressive capability of the learned models. Second, real-world cyber attacks often produce feature distributions with severe intra-class heterogeneity and pronounced inter-class overlap. However, most existing DL-based NIDSs lack explicit discriminative optimization mechanisms designed to address this distributional structure, which can weaken decision boundaries and ultimately limit classification performance.

Beyond detection accuracy, model interpretability is essential for connecting detection results with effective response actions. Existing efforts to improve NIDS interpretability mainly rely on intrinsic attention visualizations [14] or model-agnostic attribution methods such as SHapley Additive exPlanations (SHAP) and Integrated Gradients [15–17]. While these techniques provide feature importance estimates, they primarily reveal simple correlations between features and predictions. As a result, they offer limited insight into more complex decision behaviors, including threshold effects, non-linear influences, and interactions among multiple features. This limitation restricts the depth of explanations available to security analysts and reduces the practical usefulness of interpretability for supporting proactive defense.

To address these challenges, we propose **HARMONI**, a two-stage hybrid learning framework designed to improve both detection performance and interpretability in network intrusion detection systems. In the first stage, rather than adopting a static fusion strategy, we introduce a dual-branch CNN-GRU encoder whose outputs are aggregated through a lightweight adaptive gating network. The gating module learns, for each traffic sample, how to dynamically weight the spatial and temporal contributions, allowing the fused representation to adapt to varying traffic characteristics. To further counter the intra-class heterogeneity and inter-class overlap commonly observed in attack distributions, we formulate a Dynamic Class-center loss that is jointly optimized with the classification objective. This loss encourages features of the same class to concentrate around their centroid while simultaneously enforcing a margin between different classes; the class centroids are updated via exponential moving average to track the evolving feature distribution during training. The second stage is motivated by the observation that deep encoding can attenuate fine-grained statistical details that are often critical for recognizing specific attacks. We therefore adopt a residual-style design in which the learned latent representations are explicitly concatenated with the preprocessed raw traffic features, producing an information-augmented vector. This combined representation is then fed into an XGBoost ensemble classifier, which provides strong non-linear modeling capacity for the resulting heterogeneous feature space. To enhance model transparency, we develop a multi-level SHAP-based interpretability framework that examines the detection logic along three complementary dimensions, namely, global feature importance, univariate contribution trajectories that reveal decision thresholds, and pairwise feature interactions. Extensive experiments on four benchmark datasets show that HARMONI consistently outperforms mainstream deep learning baselines and recent NIDS models across key metrics, including accuracy and F1-score, while maintaining stable generalization across heterogeneous datasets.

The main contributions of this paper are summarized as follows:

- We propose HARMONI, a two-stage hybrid learning framework for network intrusion detection. In the first stage, a dual-branch CNN-GRU encoder incorporates a lightweight adaptive gating network to dynamically calibrate spatiotemporal feature weights across shifting traffic regimes. In the second stage, a residual-inspired feature augmentation mechanism explicitly concatenates the learned latent representations with raw traffic features, constructing an augmented and high-dimensional input space for the subsequent ensemble classifier.
- Inspired by metric learning principles, we formulate a Dynamic Class-Center Loss that is jointly optimized with the primary classification objective. This metric learning formulation imposes structural constraints during the training phase by iteratively calculating and updating class centroids, thereby actively enforcing intra-class compactness and inter-class separability within the latent feature topology.
- We design a SHAP-based multi-level interpretability framework to systematically decompose the model's non-linear decision processes. The analytical framework is constructed along three specific dimensions: quantifying global feature saliency distributions, mapping non-linear univariate contribution trajectories to isolate numerical decision thresholds, and mathematically decoding synergistic and inhibitory pairwise feature interactions.
- Extensive empirical evaluations across four heterogeneous benchmark datasets demonstrate that HARMONI consistently outperforms state-of-the-art baselines in key metrics such as accuracy and F1-score. Furthermore, the interpretability analysis enhances the overall transparency and trustworthiness of the detection model by deciphering the terminal ensemble's decision logic in response to fine-grained traffic features.

The remainder of this paper is structured as follows. Section 2 surveys the relevant literature. Section 3 details the proposed HARMONI framework. Section 4 presents the experimental evaluation, followed by a discussion of the results in Section 5. Finally, Section 6 concludes the work.

2. Related Work

The evolution of cybersecurity defense is driven by advances in intrusion detection technologies [18,19]. Evolving cyber threats drive the shift from rule-based systems to data-driven approaches [20]. However, deploying these models in practice raises two key challenges: learning robust and discriminative representations from highly non-linear network traffic [21], and ensuring transparency and interpretability of automated decisions [22]. This section reviews prior work in traditional machine learning, deep learning, and explainable AI for intrusion detection, and identifies the limitations that motivate our proposed framework. Table 1 provides a concise comparison of representative intrusion detection methods.

Table 1. Comparison of network intrusion detection methods.

Method	Datasets	Metrics	Interpretability	Limitations
CNN-LSTM [10]	NSL-KDD	Accuracy, Precision, Recall, F1-Score, ROC-AUC	None	Static concatenation cannot adapt spatiotemporal weights; intra-class heterogeneity and inter-class overlap unaddressed.
AE+LSTM [23]	NSL-KDD	Accuracy, Precision, Recall, F-Score, DR, FAR	None	Feature reduction may lose fine-grained statistics critical for minority attacks; no dynamic fusion weight calibration.
XGBoost-FS [24]	UNSW-NB15	Accuracy, Precision, Recall, F1-Score	Feature importance ranking only	Relies on manual features; cannot learn deep spatiotemporal representations; insufficient explanation of non-linear decisions.
GAN-based balancing [25]	CICIDS2017, UNSW-NB15	Precision, Recall, F1-Score	None	Data-level enhancement only; may introduce noise; does not improve intra-class compactness or inter-class separation structurally.
SHAP-based IDS [16]	VeReMi Extension	Accuracy, Precision, Recall, F1-Score, MCC, Cohen's Kappa	Global and local feature importance via SHAP	SHAP limited to single-feature importance; cannot reveal interaction effects or decision thresholds.
HARMONI (Ours)	NSL-KDD, UNSW-NB15, DDoS SDN, CIC-DDoS2019	Accuracy, Precision, Recall, F1-Score	Multi-level SHAP covers global feature importance, univariate decision threshold responses, and bivariate feature interaction effects	Limited interpretability depth into encoder dynamics; no support for online learning/adaptation; non-real-time explanatory overhead.

2.1. Traditional Machine Learning and Feature Engineering

Classical machine learning algorithms, such as Decision Trees, Support Vector Machines, and Random Forests, were widely adopted in early intrusion detection systems due to their effectiveness on structured, hand-crafted features [26]. However, their reliance on manual feature engineering limits their ability to process raw, unstructured traffic and adapt to evolving polymorphic threats.

Contemporary research has sought to extend the utility of these shallow architectures through advanced feature selection and ensemble strategies [24]. Focusing on dimensionality reduction, Al-Yaseen et al. [27] developed a wrapper-based framework that couples Differential Evolution with Extreme Learning Machines. This approach effectively identifies optimal feature subspaces to enhance real-time processing, although the iterative evaluation mechanism of wrapper methods imposes a substantial computational burden. To address the pervasive issue of class imbalance, Zhang et al. [28] proposed the Multiple Integration Model (MIM). By combining hybrid resampling techniques with a weighted ensemble of

LightGBM, XGBoost, and CatBoost, this model mitigates the bias toward majority classes. While effective, the intricate weak-coupling strategy necessitates extensive hyperparameter tuning and expert knowledge. Furthermore, Hammad et al. [29] introduced the MMM-RF architecture, which integrates Multinomial Mixture Models with Random Forests. This pipeline employs t-SNE for manifold learning and SMOTE for data augmentation to manage high-dimensional distributions. Despite its robustness, the multi-stage preprocessing workflow significantly increases the complexity of the training pipeline.

Despite these refinements, traditional machine learning methods remain constrained by limited representation capacity. Their shallow architectures struggle to capture non-linear semantics from raw traffic, making their performance heavily dependent on domain-specific feature engineering. This limitation reduces their adaptability to evolving threats and motivates the adoption of deep learning for hierarchical representation learning.

2.2. Deep Representation Learning

Compared with traditional machine learning models, deep learning enables end-to-end training to automatically learn hierarchical feature representations from raw data. This capability reduces reliance on manual feature engineering and facilitates the modeling of high-dimensional and non-linear characteristics inherent in network traffic [30].

Intrusion patterns typically exhibit both local spatial characteristics and temporal dependencies [31]. Accordingly, recent research increasingly favors hybrid spatiotemporal architectures over single-modality models, such as CNN-based spatial extractors or RNN-based sequential learners [32]. For instance, models combining Convolutional and Long Short-Term Memory networks (CNN-LSTM) are designed to capture both spatial patterns and temporal dynamics concurrently, demonstrating superior performance on benchmark datasets [10]. To further improve feature discriminability, many approaches incorporate representation learning and dimensionality reduction within unified pipelines. Representative methods integrate autoencoders for latent feature extraction with sequential models such as LSTMs for classification [23]. While effective in suppressing noise, the separation between reconstruction and classification objectives may introduce additional latency. Furthermore, the dynamic and imbalanced nature of network traffic poses a significant challenge to these models. To address the pervasive issue of class imbalance, data augmentation methods based on Generative Adversarial Networks (GANs) have been proposed, which synthesize minority class samples to balance the dataset and enhance the classifier's ability to detect rare attacks [25]. Fusion-based strategies further combine non-linear and linear transformations, enhancing robustness at the cost of increased architectural complexity. In parallel, emerging paradigms, including Graph Neural Networks (GNNs) [33] and Transformers [34], extend modeling capacity to capture structural dependencies and long-range interactions [35]. However, these data-intensive models are often constrained by limited labeled data availability and substantial computational overhead.

Despite their strong detection performance, deep learning models face two fundamental challenges in practical deployment. First, the dynamic and stochastic nature of network traffic often leads to representation ambiguity, resulting in high intra-class variance and reduced inter-class separability, which can hinder model generalization across diverse scenarios [36]. Second, the limited interpretability of deep models restricts transparency and complicates root cause analysis [22]. These challenges motivate the integration of explainable AI techniques to enhance both model reliability and decision transparency.

2.3. Interpretability and Trustworthy Analysis

In the adversarial landscape of cybersecurity, model interpretability is as critical as detection accuracy [37]. Transparent decision making is essential for fostering analyst trust,

verifying automated responses, and performing root-cause diagnosis—prerequisites for real-world operational deployment [38].

Because intrinsically interpretable models (e.g., Decision Trees) often sacrifice predictive power for structural transparency [39], modern intrusion detection predominantly relies on post hoc explainability methods to dissect complex deep architectures [40]. These techniques typically fall into perturbation-based (LIME [41]), gradient-based (IG [15]), and game-theoretic (SHAP [16]) paradigms. For instance, Ma et al. [15] utilized Integrated Gradients to quantify feature contributions via path integrals. While providing stable attribution, the IG struggles with highly non-linear models and yields coarse-grained explanations for complex feature interactions. Similarly, Durojaye et al. [41] applied LIME to generate instance-level weights. Despite offering intuitive local transparency, LIME's perturbation-based nature makes it notoriously unstable and incapable of capturing global feature patterns. Furthermore, Andresini et al. [14] incorporated attention mechanisms to generate visual heatmaps. Although visually intuitive, the underlying attention weights are highly susceptible to training stochasticity and do not strictly equate to feature importance.

Despite these advancements, current XAI frameworks in intrusion detection remain fundamentally limited. By predominantly focusing on marginal feature attribution or local linear approximations, existing methods fail to systematically quantify complex, non-linear feature interactions. This deficiency yields superficial explanations that cannot faithfully elucidate the model's latent decision boundaries. Ultimately, this lack of rigorous transparency restricts security analysts from deriving granular and actionable diagnostic evidence for mitigating sophisticated attack patterns, motivating the need for more comprehensive and interaction-aware interpretability analysis.

3. Proposed Methodology

3.1. Architecture of HARMONI

As illustrated in Figure 2, HARMONI adopts a decoupled, "divide-and-conquer" paradigm that systematically separates representation learning from classification through four synergistic processing modules. Initially, the topological analysis and preprocessing module utilizes UMAP to characterize the latent geometric manifold of diverse traffic datasets, empirically confirming the pervasive intra-class heterogeneity and inter-class overlap. This exploratory analysis provides the necessary analytical grounding for the subsequent design of Stage-I, which operationalizes a discriminative spatiotemporal representation learning module featuring a dual-branch CNN-GRU encoder. To integrate local spatial structural semantics with long-range temporal dependencies, a lightweight adaptive gating network dynamically recalibrates feature fusion at the sample level. This encoder is jointly optimized using a standard cross-entropy objective and a Dynamic Class-Center Loss, which acts as a topological constraint to explicitly enforce intra-class compactness and inter-class separability. To compensate for the fine-grained raw attributes inevitably attenuated during high-level semantic abstraction, Stage-II introduces an augmented hybrid classification mechanism that concatenates preprocessed raw features with the distilled latent representations. This skip-connection strategy feeds an information-rich augmented vector into an XGBoost ensemble classifier to delineate robust, non-linear decision boundaries. Finally, to transcend the inherent opacity of deep detection models, a multi-level interpretability engine integrates the SHAP framework. By systematically decomposing the XGBoost predictions into global feature saliency, univariate contribution trajectories, and pairwise interaction dynamics, this final module quantitatively decodes the classifier's decision manifold, providing transparent and traceable attributions for security analysts.

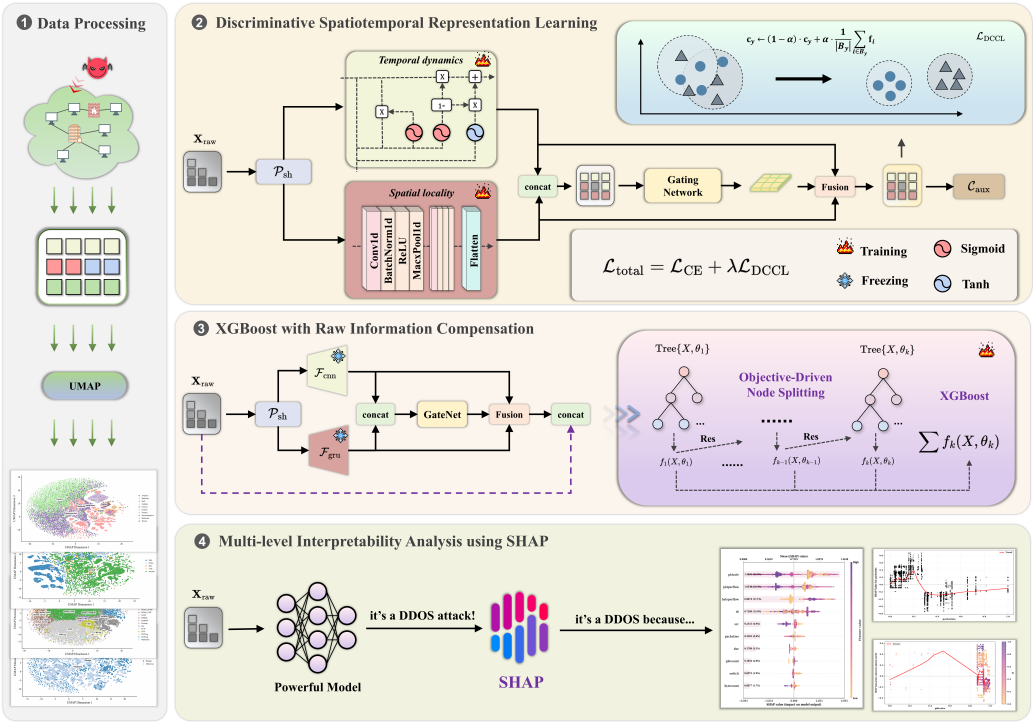


Figure 2. The workflow of HARMONI consists of four sequential modules. The pipeline begins with data preprocessing and UMAP-based topological analysis, followed by Stage-I spatiotemporal representation learning via a CNN-GRU encoder. Subsequently, Stage-II performs classification using XGBoost on concatenated raw and latent features, and the process concludes with SHAP-based interpretability analysis.

3.2. Discriminative Spatiotemporal Representation Learning

The efficacy of network traffic analysis hinges on capturing the intricate interplay between spatial semantics and temporal dynamics. Spatially, packet-level attributes (e.g., protocol fields and flags) form local semantic patterns; temporally, flow-level variations (e.g., inter-arrival times and state transitions) dictate the behavioral logic of network entities. Crucially, the manifestation of cyber attacks is often modality-dependent: while DDoS attacks typically exhibit spatial bursts and high-concurrency flows, port scanning manifests as systematic, periodic probing sequences in the temporal dimension. Consequently, a robust representation model must not only extract these dual-domain features but also adaptively modulate their contributions based on the observed attack modality. Furthermore, as empirically evidenced in Section 4.2.1, the pervasive "high intra-class heterogeneity and inter-class overlap" suggest that conventional cross-entropy supervision alone is insufficient to supervise the formation of highly discriminative feature manifolds.

To address these challenges, we propose a framework designed to learn compact and separable spatiotemporal representations, the components of which are detailed in the following subsections. We begin by delineating the Shared Projection Network in Section 3.2.1, followed by a dual-branch encoder architecture that leverages CNNs for spatial feature extraction (Section 3.2.2) and GRUs for temporal sequence modeling (Section 3.2.3). To navigate the varying modal dominance across different attack categories, Section 3.2.4 introduces an Adaptive Gated Fusion (AGF) module that is designed to dynamically recalibrate the fusion weights of dual-domain features. Finally, we formalize the Dynamic Class-Center Loss (DCCL) in Section 3.2.5, which functions to shrink intra-class variance and expand inter-class margins, thereby ensuring the formation of a highly discriminative latent space that transcends the limitations of conventional supervision.

3.2.1. Shared Projection Network

Preprocessed traffic attributes, including packet length, protocol types, and statistical metrics, naturally exhibit significant heterogeneity in both physical semantics and numerical scales. To harmonize these inputs, we designed a Shared Projection Network $\mathcal{P}_{\text{sh}}$ to map the raw features into a unified high-dimensional latent space. Acting as a foundational encoder, $\mathcal{P}_{\text{sh}}$ provides a synchronized representational basis for the subsequent spatial and temporal modules, thereby facilitating preliminary feature calibration and enhancement.

Let $\mathbf{X}_{\text{raw}} \in \mathbb{R}^{b \times d_{\text{raw}}}$ denote the input batch, where b represents the batch size and d_{raw} is the raw feature dimension. The projection network maps $\mathbf{X}_{\text{raw}}$ to a latent representation $\mathbf{Z} \in \mathbb{R}^{b \times d_{\text{sh}}}$, formally expressed as

$$\mathbf{Z} = \mathcal{P}_{\text{sh}}(\mathbf{X}_{\text{raw}}) \tag{1}$$

Architecturally, $\mathcal{P}_{\text{sh}}$ comprises a stack of non-linear projection blocks. The transition for the l-th layer is defined as

$$\mathbf{H}^{(l)} = \text{ReLU}\Big(\text{Dropout}\Big(\text{LN}\Big(\mathbf{H}^{(l-1)}\mathbf{W}^{(l)} + \mathbf{b}^{(l)}\Big)\Big)\Big) \tag{2}$$

Here, $\mathbf{H}^{(0)} = \mathbf{X}_{\text{raw}}$ serves as the initialization, while $\mathbf{W}^{(l)}$ and $\mathbf{b}^{(l)}$ denote the learnable parameters. To enhance optimization stability and generalization, each block integrates Layer Normalization (LN) and Dropout.

Through training, this network learns feature weights automatically, thereby enhancing features that are important for subsequent spatiotemporal pattern recognition while suppressing redundant or noisy information. This provides downstream modules with a more informative and better-structured starting point for feature processing. Furthermore, the network standardizes the feature dimension to a fixed size of d_{sh}. Regardless of variations in the original feature dimensions, the input structure for downstream modules remains consistent, which significantly enhances the model's versatility and deployment flexibility.

3.2.2. Spatial Encoder

Network traffic attributes in standard datasets are typically organized into functional groups, such as flow identifiers, packet length statistics, and time-related metrics. This intrinsic logical organization creates local semantic structures where proximate dimensions within the feature vector exhibit statistical dependencies. To capture these patterns, we design a spatial encoder $\mathcal{F}_{\text{cnn}}$ utilizing a 1D-CNN. By applying localized receptive fields to these structured attributes, the encoder effectively extracts multi-scale spatial representations from the latent projection $\mathbf{Z}$.

Specifically, the latent input $\mathbf{Z}$ is reshaped into a tensor $\mathbf{Z}_{\text{cnn}} \in \mathbb{R}^{b \times 1 \times d_{\text{sh}}}$ to align with the convolutional inductive bias, preserving the sequentiality of feature dimensions. The encoder comprises a cascade of one-dimensional convolutional blocks, where each block l executes a sequence of convolution, normalization, activation, and pooling. Formally, the transformation is defined as

$$\mathbf{H}_l = \text{Pool}(\text{ReLU}(\text{BN}(\text{Conv1D}(\mathbf{H}_{l-1})))) \tag{3}$$

Here, $\mathbf{H}_0 = \mathbf{Z}_{\text{cnn}}$ initializes the stack. Through this hierarchical composition, the network progressively expands its receptive field, integrating local features into global multi-scale representations.

To ensure dimension-invariant output and capture global statistics, the final convolutional feature map $\mathbf{H}_L$ undergoes adaptive average pooling. The resulting tensor is

flattened and projected via a fully connected layer to a target dimension d_c, yielding the spatial representation $\mathbf{F}_{\text{cnn}}$:

$$\mathbf{F}_{\text{cnn}} = \text{Linear}(\text{Flatten}(\text{AdaptiveAvgPool1D}(\mathbf{H}_L))) \in \mathbb{R}^{b \times d_c} \tag{4}$$

This architecture effectively distills spatial dependencies while mitigating computational complexity through pooling-induced dimensionality reduction. Crucially, by setting the output dimension d_c to match that of the temporal module, $\mathbf{F}_{\text{cnn}}$ provides a structurally aligned embedding, facilitating seamless integration in the subsequent fusion stage.

3.2.3. Temporal Modeler

To capture complex dependencies across the entire feature space, we construct a temporal modeler $\mathcal{F}_{\text{gru}}$ leveraging GRUs. We reshape the latent projection $\mathbf{Z} \in \mathbb{R}^{b \times d_{\text{sh}}}$ into a pseudo-temporal sequence $\mathbf{Z}_{\text{gru}} \in \mathbb{R}^{b \times d_{\text{sh}} \times 1}$. By treating the feature vector as a sequence, we leverage the GRU's recurrent mechanism to model the cumulative information flow across different attribute groups. This approach does not imply a literal temporal progression; instead, it serves as a mechanism to capture long-range dependencies and integrate global context that local receptive fields might overlook. Processing $\mathbf{Z}_{\text{gru}}$ enables the GRU to trace dependencies along the feature dimension, refining the robustness of the representation.

The GRU regulates this flow through its gating mechanism, mathematically formulated as

$$\mathbf{r}_t = \sigma(\mathbf{W}_r[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_r) \tag{5}$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_z[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_z) \tag{6}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h[\mathbf{r}_t \odot \mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_h) \tag{7}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \tag{8}$$

where $\mathbf{r}_t$ and $\mathbf{z}_t$ denote the reset and update gates, σ represents the sigmoid function, and $\odot$ is the Hadamard product. Crucially, the reset gate modulates the retention of prior states, while the update gate balances the integration of historical context with current input, empowering the model to adaptively capture long-range dependencies.

Processing $\mathbf{Z}_{\text{gru}}$ yields a sequence of hidden states $\mathbf{H} \in \mathbb{R}^{b \times d_{\text{sh}} \times h}$. We extract the final state $\mathbf{h}_{\text{last}} = \mathbf{H}[:, -1, :]$ as the global temporal embedding. This representation is then refined through a projection block comprising a linear layer, Layer Normalization (LN), ReLU activation, and Dropout:

$$\mathbf{F}_{\text{gru}} = \text{Dropout}(\text{ReLU}(\text{LN}(\text{Linear}(\mathbf{h}_{\text{last}})))) \in \mathbb{R}^{b \times d_g} \tag{9}$$

By enforcing $d_g = d_c$, we ensure dimensional alignment between spatial and temporal embeddings, laying the groundwork for seamless feature fusion. Ultimately, $\mathcal{F}_{\text{gru}}$ distills temporal dynamics that, when complementarily fused with spatial cues, enhance the robustness of the final decision boundary.

3.2.4. Adaptive Gated Fusion Module

To synergize the spatial features $\mathbf{F}_{\text{cnn}}$ and temporal features $\mathbf{F}_{\text{gru}}$, we propose an Adaptive Gated Fusion Module designed to accommodate the modality-specific characteristics of heterogeneous attack behaviors. For example, DDoS attacks are predominantly defined by spatial burstiness, while port scanning typically manifests as temporal persistence. Unlike static fusion strategies, our module employs a data-driven gating mechanism to dynamically recalibrate the importance of spatial versus temporal channels based on input-specific patterns.

Formally, the gating network $\mathcal{G}$ accepts the concatenation of the dual-branch outputs, $\mathbf{H}_{\text{concat}} = [\mathbf{F}_{\text{cnn}}; \mathbf{F}_{\text{gru}}] \in \mathbb{R}^{b\times(d_c+d_g)}$, as input. It computes an adaptive weight matrix $\mathbf{A} \in [0,1]^{b\times d_f}$ via a two-layer Multi-Layer Perceptron (MLP):

$$\mathbf{A} = \sigma(\text{BN}(\mathbf{H}_{\text{concat}}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2) \tag{10}$$

where $\mathbf{W}_1, \mathbf{W}_2$ and $\mathbf{b}_1, \mathbf{b}_2$ denote the learnable weights and biases, respectively. The sigmoid function σ constrains the weights to the interval [0, 1]. Consequently, the fused representation is derived via a channel-wise soft selection mechanism:

$$\mathbf{H}_{\text{fusion}} = \mathbf{A} \odot \mathbf{F}_{\text{cnn}} + (\mathbf{1} - \mathbf{A}) \odot \mathbf{F}_{\text{gru}} \tag{11}$$

The gating matrix $\mathbf{A}$ offers explicit semantic interpretability regarding feature dominance. Specifically, each element A_{ij} quantifies the contribution of the spatial modality for the i-th sample at the j-th dimension. Values approaching 1 prioritize spatial patterns, which aligns with bursty attacks like DDoS, while values near 0 emphasize temporal dynamics, capturing persistent scanning behaviors. Intermediate values indicate a balanced spatiotemporal dependency. This fine-grained, dimension-aware weighting mechanism empowers the model to adaptively modulate feature integration, significantly enhancing the representation's discriminative robustness across diverse attack scenarios.

3.2.5. Dynamic Class-Center Constraint Guided Joint Representation Learning

Upon obtaining the fused spatiotemporal features $\mathbf{H}_{\text{fusion}}$ via the Adaptive Gated Fusion Module, optimizing the model relies on establishing a robust feature space. As highlighted in Section 4.2.1, network traffic data natively exhibit significant intra-class heterogeneity and ambiguous inter-class boundaries. Relying exclusively on standard cross-entropy loss is insufficient, as it primarily focuses on linear separability in the semantic space without explicitly regularizing the underlying geometric manifold. Consequently, features often remain dispersed within classes and intermixed across classes.

To systematically address these distributional challenges, we draw inspiration from classic metric learning paradigms [42] and formulate the Dynamic Class-Center Loss ($\mathcal{L}_{\text{DCCL}}$) tailored for network intrusion detection. This objective directly shapes the feature manifold topology by simultaneously enforcing intra-class compactness and inter-class separability. Specifically, for each category $y \in \{1, 2, ..., C\}$, we maintain a dynamic centroid $\mathbf{c}_y \in \mathbb{R}^{d_f}$ serving as a prototype for feature aggregation. To mitigate intra-class variance, building upon the concept of center loss [42], we formulate an intra-class constraint that minimizes the Euclidean distance between each sample feature $\mathbf{f}_i \in \mathbf{H}_{\text{fusion}}$ and its corresponding centroid $\mathbf{c}_{y_i}$:

$$\mathcal{L}_{\text{intra}} = \frac{1}{B}\sum_{i=1}^{B}\|\mathbf{f}_i - \mathbf{c}_{y_i}\|_2^2 \tag{12}$$

where B denotes the batch size. Conversely, excessive focus on compactness risks feature collapse. To prevent distinct clusters from converging indiscriminately, we introduce a margin-based inter-class separability constraint, which is conceptually aligned with contrastive margin penalties [43]. By calculating the distance to the nearest non-target centroid, denoted as $d_i^{\text{neg}} = \min_{j\neq y_i}\|\mathbf{f}_i - \mathbf{c}_j\|_2$, we penalize distance violations below a predefined decision margin $\delta > 0$:

$$\mathcal{L}_{\text{inter}} = \frac{1}{B}\sum_{i=1}^{B}\max\left(0, \delta - d_i^{\text{neg}}\right) \tag{13}$$

The total geometric regularization is then defined as a weighted linear combination of the two constraints:

$$\mathcal{L}_{\text{DCCL}} = \omega_{\text{intra}}\mathcal{L}_{\text{intra}} + \omega_{\text{inter}}\mathcal{L}_{\text{inter}} \tag{14}$$

where ω_{intra} and ω_{inter} are hyperparameters governing the trade-off between local compactness and global separability.

Crucially, the efficacy of $\mathcal{L}_{\text{DCCL}}$ hinges on the centroids' ability to accurately track the non-stationary feature distribution during training. Rather than relying on fixed prototypes or standard gradient-based optimization, we employ an exponential moving average (EMA) mechanism to update the centroids dynamically, which is a proven strategy for maintaining feature consistency in momentum-based representation learning [44]. After each iteration, $\mathbf{c}_y$ is calibrated using the mean of the current batch samples $\mathcal{B}_y$ belonging to class y:

$$\mathbf{c}_y \leftarrow (1-\alpha)\cdot\mathbf{c}_y + \alpha\cdot\frac{1}{|\mathcal{B}_y|}\sum_{i\in\mathcal{B}_y}\mathbf{f}_i \tag{15}$$

where $\alpha \in (0,1]$ represents the momentum coefficient. By decoupling the centroids from the computational graph, this dynamic adaptation mitigates constraint degradation caused by center misalignment and ensures historical smoothing for training stability.

While $\mathcal{L}_{\text{DCCL}}$ refines the geometric structure, explicitly maximizing classification confidence requires semantic supervision. To this end, we introduce an auxiliary linear classifier $\mathcal{C}_{\text{aux}}$ that maps the high-dimensional representations $\mathbf{H}_{\text{fusion}}$ to class logits. These logits are normalized via the Softmax function to yield predicted probabilities $\mathbf{P}$. The canonical Cross-Entropy loss is then computed as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} y_{i,j}\log(p_{i,j}) \tag{16}$$

where $y_{i,j}$ and $p_{i,j}$ denote the one-hot encoded ground-truth label and the predicted probability, respectively.

By jointly optimizing the cross-entropy and Dynamic Class-Center Losses, our framework achieves a profound synergy between semantic discriminability and geometric robustness. Specifically, $\mathcal{L}_{\text{CE}}$ promotes linear classification confidence, while $\mathcal{L}_{\text{DCCL}}$ imposes explicit clustering constraints to guarantee a well-separated feature manifold. The overall multi-objective training function is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{DCCL}} \tag{17}$$

where $\lambda > 0$ balances the two objectives.

To formalize the learning paradigm, we define the overall encoder topology $\mathcal{E}_{\text{ST}}$, which encompasses the shared projection $\mathcal{P}_{\text{sh}}$, parallel spatial/temporal branches ($\mathcal{F}_{\text{cnn}}$, $\mathcal{F}_{\text{gru}}$), and the fusion module $\mathcal{G}$:

$$\mathcal{E}_{\text{ST}} = \mathcal{G}\circ(\mathcal{F}_{\text{cnn}}\oplus\mathcal{F}_{\text{gru}})\circ\mathcal{P}_{\text{sh}} \tag{18}$$

where $\oplus$ denotes concatenation and $\circ$ represents function composition. The joint training process is formulated as the empirical risk minimization over the dataset of N samples:

$$\Theta^* = \arg\min_{\Theta,\{\mathbf{c}_y\}}\frac{1}{N}\sum_{i=1}^{N}\left[\mathcal{L}_{\text{CE}}(\mathcal{C}_{\text{aux}}(\mathcal{E}_{\text{ST}}(\mathbf{x}_i)),y_i) + \lambda\cdot\mathcal{L}_{\text{DCCL}}(\mathcal{E}_{\text{ST}}(\mathbf{x}_i),\{\mathbf{c}_y\},y_i)\right] \tag{19}$$

Here, Θ encapsulates all learnable parameters of $\mathcal{E}_{\text{ST}}$, and $\{\mathbf{c}_y\}$ is the dynamically maintained centroid set. Upon convergence, both the auxiliary classifier $\mathcal{C}_{\text{aux}}$ and the centroids $\{\mathbf{c}_y\}$ are discarded, ensuring zero additional computational overhead during the inference phase. The optimized encoder $\mathcal{E}_{\text{ST}}$ is subsequently retained to extract robust representations for downstream tasks. The complete procedure for this discriminative spatiotemporal learning is summarized in Algorithm 1.

Algorithm 1: Discriminative Spatiotemporal Representation Learning

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$; Batch size B; Learning rate η; Momentum α; Margin δ; Loss weights $\lambda, \omega_{\text{intra}}, \omega_{\text{inter}}$.

Parameters: Encoder params $\Theta_{\mathcal{E}}$ (shared $\mathcal{P}_{\text{sh}}$, spatial $\mathcal{F}_{\text{cnn}}$, temporal $\mathcal{F}_{\text{gru}}$, fusion $\mathcal{G}$); Classifier params $\Theta_{\mathcal{C}}$; Class centroids $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^{C}$.

Output: Optimized Encoder $\mathcal{E}_{\text{ST}}$.

```
// Initialization
Initialize network parameters Θ_E, Θ_C randomly; Initialize centroids C;
while not converged do
    Sample a mini-batch {X, Y} from D;
    // Phase 1: Dual-Branch Spatio-Temporal Encoding
    Z ← P_sh(X) ;                                   // Project raw features to latent space
    Z_cnn ← Reshape(Z); F_cnn ← F_cnn(Z_cnn) ;       // Extract local spatial patterns
    Z_gru ← Reshape(Z); F_gru ← F_gru(Z_gru) ;       // Capture sequential temporal dependencies
    // Phase 2: Adaptive Gated Fusion
    A ← G([F_cnn; F_gru]) ;                         // Compute modality-aware attention weights
    H_fusion ← A ⊙ F_cnn + (1 − A) ⊙ F_gru ;         // Synthesize fused representations
    // Phase 3: Loss Computation
    S ← C_aux(H_fusion) ;                           // Compute class logits
    L_CE ← −(1/B) Σ log(Softmax(S)_y) ;             // Standard semantic supervision
    Compute distances d_{i,k} = ||f_i − c_k||_2;
    L_intra ← (1/B) Σ ||f_i − c_{y_i}||_2^2 ;       // Enforce intra-class compactness
    L_inter ← (1/B) Σ max(0, δ − min_{j≠y_i} ||f_i − c_j||_2) ;   // Enforce inter-class separability
    L_DCCL ← ω_intra L_intra + ω_inter L_inter;
    L_total ← L_CE + λ L_DCCL ;                     // Joint optimization objective
    // Phase 4: Optimization & Centroid Update
    Compute gradients ∇_Θ L_total;
    Update Θ_E, Θ_C using Adaw;
    for each class k in batch Y do
        f̄_k ← (1/|B_k|) Σ_{i∈B_k} f_i ;             // Calculate batch-wise mean
        c_k ← (1 − α) c_k + α f̄_k ;                 // Update dynamic centroids via EMA
return E_ST ← {P_sh, F_cnn, F_gru, G};
```

3.3. XGBoost Classification with Raw Information Compensation

Building upon the discriminative latent space constructed by the Stage-I encoder, the second stage of HARMONI uses a decoupled classification approach. While successive non-linear transformations in deep spatiotemporal encoders excel at abstracting high-level topological semantics, they inevitably induce "semantic forgetting", a phenomenon

characterized by the attenuation of fine-grained, discrete statistical details. In network intrusion detection, subtle localized anomalies, such as specific protocol flags, discrete port combinations, or minor Time-to-Live (TTL) deviations, are often the primary artifacts of sophisticated attack vectors. To mitigate this architectural limitation, we introduce a customized *Raw Information Compensation* mechanism.

Rather than relying solely on the distilled semantic abstractions, we construct a shortcut connection that directly bridges the preprocessed original input space to the downstream classifier. Formally, the robust deep representation $\mathbf{F}_{\text{fusion}}$ generated by the Stage-I encoder is explicitly augmented with the raw feature matrix $\mathbf{X}_{\text{raw}}$. This integration is realized via concatenation along the feature dimension, yielding an enhanced, information-rich feature space:

$$\mathbf{F}_{\text{enhanced}} = [\mathbf{F}_{\text{fusion}} \parallel \mathbf{X}_{\text{raw}}] \in \mathbb{R}^{b \times (d_f + d_{\text{raw}})} \tag{20}$$

Crucially, unlike standard residual connections [45] that employ element-wise addition primarily to stabilize gradient backpropagation, our architecture utilizes concatenation to explicitly maximize information fidelity. This structural design strictly preserves the determinism of low-level traffic statistics, ensuring they organically complement the high-level continuous embeddings without being mathematically diluted.

To project this augmented, highly heterogeneous feature space ($\mathbf{F}_{\text{enhanced}}$) into a precise decision manifold, we employ the XGBoost algorithm [46] as the terminal classifier. While deep neural networks exhibit formidable manifold learning capabilities, gradient tree boosting possesses a superior inductive bias for handling heterogeneous tabular data. Specifically, tree-based ensembles are inherently adept at processing concatenated vectors that comprise both dense continuous representations (derived from the deep encoder) and discrete attributes (retained from the raw input). By optimizing a regularized additive ensemble of regression trees, XGBoost effectively demarcates complex, non-linear decision boundaries while maintaining robust resilience against the curse of dimensionality.

Ultimately, this decoupled stage specifically calibrates the HARMONI framework for the fine-grained sensitivity mandated by intrusion detection. By synergizing deep topological representation with explicit feature-compensation and tree-ensemble decision making, our method achieves both robust generalization and high-fidelity signature recognition. The comprehensive inference pipeline of HARMONI is summarized in Algorithm 2.

3.4. SHAP-Based Multi-Level Interpretability Framework

To address the lack of transparency of deep representation learning and ensemble decision making in the final stage of HARMONI, we introduce a multi-level interpretability framework grounded in SHapley Additive exPlanations (SHAP). While conventional Explainable AI paradigms predominantly focus on isolated marginal feature importance, they often fail to capture the synergistic multi-vector nature of modern network intrusions. To address this limitation, our framework decodes the XGBoost decision manifold through three progressive analytical dimensions that encompass macroscopic global attribution, univariate dependence, and high-order combinatorial interaction. As the foundational interpretability engine, we leverage TreeSHAP, which exploits the topological structure of tree ensembles to reduce the intractable $O(2^d)$ complexity of exact Shapley value computation to polynomial time. Formally, given an input traffic instance $\mathbf{x} \in \mathbb{R}^d$ and the trained classifier f, the exact SHAP marginal contribution ϕ_i of the i-th feature is defined as

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \tag{21}$$

where $N = \{1, 2, \ldots, d\}$ denotes the complete feature space, and S represents any feature coalition excluding i. By satisfying the axioms of efficiency and symmetry, SHAP rigorously decomposes the model's prediction into the base expected value and a linear combination of localized feature attributions.

Algorithm 2: HARMONI: Two-Stage Classification with Raw Information Compensation

Input: Training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$; Pre-trained Stage-I Encoder $\mathcal{E}_{\text{ST}}$.
Parameters: XGBoost hyperparameters Θ (tree depth, learning rate, etc.).
Output: Optimized Classifier Φ; Inference Interface $\mathcal{H}_{\text{HARMONI}}$.

```
// Phase 1: Feature Augmentation via Raw Information Compensation
Initialize augmented training set D_enh ← ∅;
Freeze E_ST parameters;
for each sample (x_i, y_i) ∈ D do
    h_i ← E_ST(x_i) ;                          // Extract deep spatiotemporal semantics
    f_i^enh ← [h_i ‖ x_i] ;                    // Concatenate raw features with latent embeddings
    D_enh ← D_enh ∪ {(f_i^enh, y_i)} ;         // Construct information-rich dataset

// Phase 2: Tree Ensemble Optimization
Φ ← XGBoost.Train(D_enh, Θ) ;   // Optimize gradient boosting trees on augmented space

// Phase 3: Online Inference Pipeline
Function Predict(x_query) is
    h_query ← E_ST(x_query) ;                  // Deep semantic extraction
    f_query^enh ← [h_query ‖ x_query] ;        // Raw information compensation
    p ← Φ.predict(f_query^enh) ;               // Ensemble decision making
    return ŷ ← arg max(p) ;                    // Final classification
return H_HARMONI ← Predict;
```

Based on this mathematical foundation, our framework first unpacks the classification logic at a macroscopic level through global feature attribution, which isolates the primary drivers dictating the overall decision boundaries. We aggregate instance-level Shapley values into a unified global importance score I_i, quantified as the expected absolute SHAP magnitude across the evaluation distribution $\mathcal{D}_{\text{test}}$:

$$I_i = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{test}}}[|\phi_i(f, \mathbf{x})|] = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{\mathbf{x} \in \mathcal{D}_{\text{test}}} |\phi_i(f, \mathbf{x})| \tag{22}$$

Features yielding higher I_i values serve as the principal discriminators separating benign traffic from anomalous vectors. Moving beyond macroscopic aggregation, the framework then inspects granular univariate mappings through feature dependence analysis. By projecting original feature values against their localized SHAP attributions, we map the non-linear relationship between individual traffic fluctuations and model confidence. This stage explicitly reveals critical micro-patterns, such as precise activation thresholds and monotonic trends, which are essential for profiling specific attack signatures.

Because advanced intrusions inherently manifest as coordinated multi-dimensional anomalies, relying solely on isolated feature analysis provides an incomplete forensic picture. For instance, specific packet rates often need to be analyzed in conjunction with anomalous payload lengths to accurately identify an attack. Thus, the final dimension of our framework focuses on feature interaction analysis to disentangle intrinsic main effects

from high-order joint dependencies. The SHAP interaction value $\Phi_{i,k}$ for a feature pair (i,k) is formulated as

$$\Phi_{i,k} = \sum_{S \subseteq N \setminus \{i,k\}} \frac{|S|!(|N|-|S|-2)!}{2(|N|-1)!} \delta_{i,k}(S) \tag{23}$$

where the discrete second-order difference, $\delta_{i,k}(S) = f(S \cup \{i,k\}) - f(S \cup \{i\}) - f(S \cup \{k\}) + f(S)$, captures the purely interactive effect of the joint coalition relative to their independent presence. A positive $\Phi_{i,k}$ signifies synergistic amplification, whereas a negative value denotes antagonistic suppression. By systematically mapping from overarching feature salience to intricate pairwise dynamics, this analytical framework not only provides transparent and forensic-grade attributions for security analysts but also validates whether HARMONI establishes robust and logically sound criteria for intrusion detection.

Note that while the Stage-II classifier utilizes both latent and raw features, our interpretability analysis primarily focuses on the raw compensated variables. This choice is made because raw features (e.g., packet rate, flow duration) possess direct physical meanings for network administrators, making the SHAP attributions more actionable for real-world forensic analysis compared to high-dimensional latent representations.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets and Evaluation Metrics

To validate the efficacy and generalization capability of our proposed framework, we employ four representative benchmark datasets. UNSW-NB15 [47] and NSL-KDD [48] serve as standard baselines for general intrusion detection, utilizing their official predefined training and testing partitions to ensure comparability with prior studies. Additionally, to investigate adaptability within Software-Defined Networking (SDN) architectures and against modern high-volume threats, we incorporate the DDoS attack SDN [49] and CIC-DDoS2019 [50] datasets. For these specialized datasets, we apply a stratified 7:3 split for training and evaluation. Collectively, these benchmarks cover a diverse spectrum of network behaviors, ranging from traditional attack families to complex DDoS floods. As summarized in Table 2, which reports total samples, class distributions, and feature dimensions, UNSW-NB15 (10 classes), NSL-KDD (5 classes), and CIC-DDoS2019 (16 classes) are multiclass classification tasks, while DDoS Attack SDN is a binary task.

Table 2. Summary statistics of the four datasets used in this study.

Dataset	Total Samples	Features	Benign	Anomaly
UNSW-NB15	257,673	42	Normal (36.10%)	Generic (22.85%), Exploits (17.28%), Fuzzers (9.41%), DoS (6.35%), Reconnaissance (5.43%), Analysis (1.04%), Backdoor (0.90%), Shellcode (0.59%), Worms (0.07%)
NSL-KDD	148,517	41	Normal (51.88%)	DoS (35.95%), Probe (9.48%), R2L (2.52%), U2R (0.17%)
DDoS Attack SDN	103,839	22	BENIGN (60.99%)	ATTACK (39.01%)
CIC-DDoS2019	431,371	77	Benign (22.68%)	DrDoS_NTP (28.14%), TFTP (22.93%), Syn (11.44%), UDP (4.19%), UDP-lag (2.06%), MSSQL (1.98%), DrDoS_MSSQL (1.44%), DrDoS_DNS (0.85%), DrDoS_SNMP (0.63%), LDAP (0.44%), DrDoS_LDAP (0.33%), Portmap (0.16%), NetBIOS (0.15%), DrDoS_NetBIOS (0.14%), UDPLag (0.01%), WebDoS (0.01%)

For performance assessment, we adopt Accuracy as the primary metric for comparison with the existing literature. Furthermore, we report precision, recall, and the F1-score in our baseline and ablation analyses to provide a granular view of the model's robustness.

4.1.2. Implementation Details and Hyperparameter Configuration

The proposed HARMONI framework is implemented using PyTorch 2.6.0, with all computational experiments executed on a workstation equipped with a single NVIDIA GeForce RTX 4090 GPU (NVIDIA Corporation, Santa Clara, CA, USA). The data preprocessing pipeline is standardized across all benchmarks: categorical attributes (e.g., 'proto', 'service') are mapped via ordinal encoding, while numerical features are normalized to the range $[0, 1]$ using Min-Max scaling to mitigate gradient instability caused by varying feature magnitudes. Crucially, these preprocessing steps are performed column-wise, ensuring that the original relative ordering of attributes within the feature vector remains invariant.

A critical design principle of this study is to validate the intrinsic universality of the model. Unlike existing approaches that often necessitate bespoke architectural modifications for distinct tasks, we maintain an invariant core architecture across all four diverse datasets described in Section 4.1.1. Specifically, the output dimension of the spatiotemporal encoder $\mathcal{E}_{\text{ST}}$ is fixed at 128, and the GRU hidden dimension is set to 64. Adaptation to specific datasets is limited to hyperparameter tuning via grid search, as detailed in Table 3. This rigorous constraint ensures that the observed generalization capability is attributable to the robustness of the proposed representation learning mechanism rather than task-specific structural heuristics.

Table 3. Training hyperparameter settings of HARMONI on NSL-KDD, UNSW-NB15, DDoS Attack SDN, and CIC-DDoS2019 datasets.

Hyperparameter	NSL-KDD	UNSW-NB15	DDoS Attack SDN	CIC-DDoS2019
Training epochs of $\mathcal{E}_{\text{ST}}$	15	15	15	15
Learning rate of $\mathcal{E}_{\text{ST}}$	0.0005	0.0005	0.0005	0.0005
Batch size of $\mathcal{E}_{\text{ST}}$	128	128	128	128
Dynamic Class-Center Loss weight λ	0.4	0.4	0.4	0.4
Intra-class loss weight ω_{intra}	1.0	1.0	1.0	1.0
Inter-class loss weight ω_{inter}	0.7	0.5	0.7	0.7
Inter-class margin δ	1.2	1.0	1.2	1.2
momentum coefficient α	0.3	0.4	0.3	0.3
XGBoost learning rate	0.25	0.01	0.25	0.25
Number of base learners (trees)	40	50	70	70
Maximum depth of trees	3	3	3	4

4.2. Results Analysis

4.2.1. Feature Distribution Analysis via UMAP

To investigate the intrinsic complexity of high-dimensional network traffic and identify underlying classification bottlenecks, we conduct a visual analysis of the feature space using Uniform Manifold Approximation and Projection (UMAP) [51]. UMAP is employed to project the high-dimensional data into a 2D embedding because of its superior ability to preserve both local geometric relationships and global topology, effectively revealing non-linear data distributions. In this experiment, we apply UMAP across four heterogeneous public datasets: DDoS Attack SDN, CIC-DDoS2019, NSL-KDD, and UNSW-NB15.

As illustrated in Figure 3, the UMAP visualizations reveal two pervasive topological characteristics across the evaluated datasets: pronounced **intra-class heterogeneity** and **inter-class overlap**. Regarding intra-class heterogeneity, samples sharing the same semantic label exhibit significant spatial dispersion rather than forming compact clusters. For instance, benign and malicious traffic are widely fragmented in the DDoS Attack SDN dataset (Figure 3a), while the Normal class in NSL-KDD (Figure 3c) demonstrates a multimodal

distribution extensively woven into attack regions. This indicates that single categories inherently encompass high structural diversity due to varying attack implementations and session dynamics. Conversely, inter-class overlap manifests as ambiguous boundaries between distinct categories. In NSL-KDD (Figure 3c) and UNSW-NB15 (Figure 3d), distinct attacks (e.g., Probe/DoS and Reconnaissance/Shellcode) occupy highly entangled manifold regions, likely due to shared underlying protocol behaviors. Furthermore, the extensive entanglement of benign and malicious traffic in CIC-DDoS2019 (Figure 3b) highlights the severe structural ambiguity present in modern datasets.

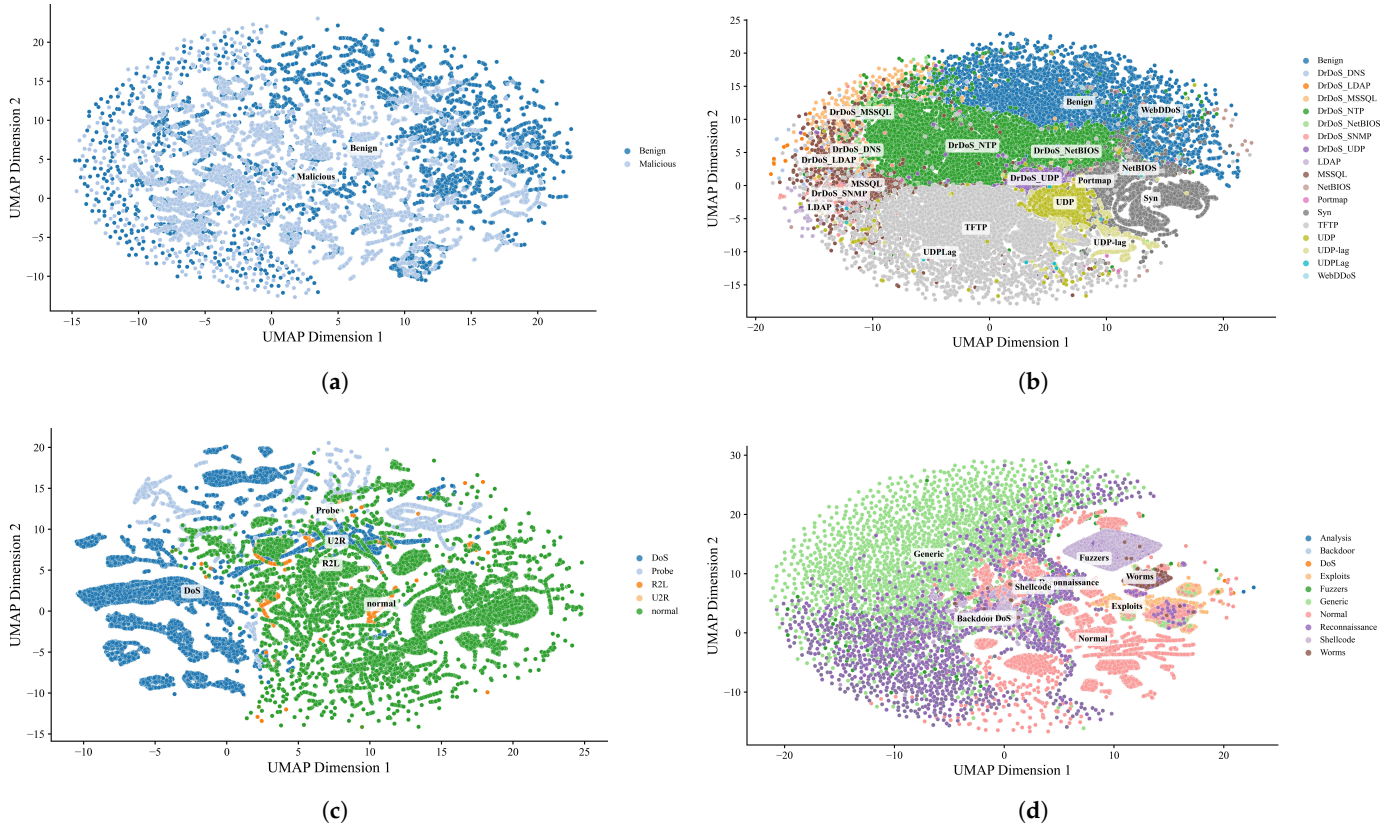


Figure 3. UMAP visualizations of feature distributions across four heterogeneous datasets. The four subplots correspond to (**a**) DDoS Attack SDN; (**b**) CIC-DDoS2019; (**c**) NSL-KDD; (**d**) UNSW-NB15.

In summary, the UMAP-based visual analysis confirms that modern network traffic data inherently possess a highly complex and ambiguous topological structure. The widespread presence of intra-class heterogeneity and inter-class overlap demonstrates severe feature entanglement across different datasets. Consequently, the core conclusion of this experiment is that the raw feature space of network traffic is intrinsically non-linear and heavily intertwined. This inherent data distribution characteristic naturally introduces structural difficulties for classification tasks, as delineating clear boundaries between widely dispersed and deeply overlapping traffic categories remains a significant data-level challenge.

4.2.2. Comparative Analysis with Representative Architectures

We evaluate HARMONI against eight baseline models spanning traditional machine learning (RF), deep learning (CNN, RNN variants), and attention mechanisms (Transformer). All experiments utilize unified dataset partitioning and evaluation protocols to ensure equitable comparisons. As detailed in Table 4, HARMONI demonstrates consistent performance margins across the majority of evaluation scenarios. On the NSL-KDD and

UNSW-NB15 datasets, which exhibit high-dimensional heterogeneity and complex attack distributions, HARMONI secures leading precision rates of 83.36% and 80.51%, respectively. Furthermore, the framework achieves an F1-score of 99.97% on the DDoS Attack SDN dataset. While we attribute this near-upper-bound metric primarily to the highly structured and synthetic nature of SDN traffic, HARMONI converges to this result more stably than the evaluated baselines.

Table 4. Quantitative performance comparison with representative baselines across four benchmark datasets. Bold values indicate the best performance.

Method	NSL-KDD			UNSW-NB15			DDoS Attack SDN			CIC-DDoS2019		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
RF	66.67%	71.21%	66.29%	77.37%	67.37%	67.88%	97.14%	96.92%	96.94%	84.92%	89.36%	86.66%
CNN	82.05%	76.67%	73.16%	79.07%	69.72%	71.03%	96.09%	95.94%	95.96%	90.91%	92.48%	90.98%
RNN	78.66%	72.83%	68.99%	78.45%	70.89%	71.93%	97.16%	97.07%	97.08%	73.58%	84.38%	78.12%
GRU	78.17%	73.70%	70.63%	77.51%	71.71%	72.01%	98.95%	98.95%	98.95%	92.86%	93.22%	92.39%
LSTM	78.63%	75.10%	72.13%	76.13%	71.07%	71.64%	97.91%	97.88%	97.89%	92.62%	93.22%	91.93%
BiLSTM	77.83%	73.36%	69.36%	80.48%	74.74%	74.85%	97.58%	97.48%	97.49%	91.98%	93.03%	91.54%
Transformer	79.56%	73.50%	68.70%	79.44%	73.52%	73.70%	99.93%	99.91%	99.92%	92.12%	92.80%	91.15%
Autoencoder	75.21%	75.22%	70.44%	**80.84%**	74.12%	74.24%	96.74%	96.62%	96.63%	89.87%	92.26%	90.47%
HARMONI (Ours)	**83.36%**	**80.19%**	**76.59%**	80.51%	**78.24%**	**77.82%**	**99.97%**	**99.97%**	**99.97%**	**94.17%**	**94.08%**	**93.31%**

The performance advantage of HARMONI stems directly from its architectural response to the high intra-class heterogeneity and inter-class overlap identified in Section 3.1. Unlike unimodal baselines that isolate spatial or temporal dependencies, the dual-branch encoder and Adaptive Gated Fusion mechanism within HARMONI synergistically extract multi-dimensional traffic representations. This capability drives the precision and recall improvements on complex benchmarks like NSL-KDD and CIC-DDoS2019. To further refine classification boundaries amidst these complex distributions, the Dynamic Class-Center Loss explicitly optimizes intra-class compactness and inter-class separability.

Moreover, the second-stage XGBoost classifier utilizes a raw-information compensation mechanism to fuse deep abstract features with original statistical attributes. This hybrid strategy mitigates the risk of semantic forgetting and establishes a more robust decision boundary. This is validated on the UNSW-NB15 dataset, where HARMONI surpasses the Autoencoder in F1-score (77.82% vs. 74.24%) despite a marginal precision deficit, indicating a better equilibrium in detecting minority classes. Conversely, while the Transformer baseline performs commendably on the simpler SDN dataset, it suffers substantial degradation on the heterogeneous NSL-KDD dataset (F1-score of 68.70%). This discrepancy suggests that relying solely on sequential modeling lacks the structural constraints required to handle highly overlapping feature spaces. HARMONI overcomes this limitation through its systematic integration of multi-view feature extraction and discriminative metric learning.

In summary, the cross-dataset evaluations substantiate the efficacy of HARMONI. By integrating spatiotemporal feature extraction with dynamic metric learning and raw-information compensation, the architecture effectively addresses the challenges of high-dimensional traffic classification. These results validate the structural coherence of the design, providing a robust empirical foundation for building more reliable and generalizable network intrusion detection systems.

4.2.3. Comparative Evaluation Against Existing Methods

To objectively assess the competitive advantages of HARMONI, this study conducts a systematic evaluation against established baselines on the NSL-KDD and UNSW-NB15 benchmarks. The selection criteria prioritize technical diversity, encompassing deep neural

networks, generative adversarial models, and hybrid architectures. The comparison relies on results reported in the original publications, as the source code for most competitors is not publicly available to guarantee exact reproduction. However, the comparison remains fair because all methods, including HARMONI, are evaluated on the official benchmark versions of the datasets with their standard train/test splits, a practice commonly accepted in the NID literature. [14]. Accuracy serves as the primary metric for this evaluation because it fundamentally quantifies overall classification performance. Furthermore, given the critical necessity of trustworthy AI in cybersecurity, the analysis extends to evaluate the interpretability capabilities of each method, as detailed in Table 5 and elaborated in Section 4.3.

Table 5. Quantitative comparison with existing state-of-the-art methods on NSL-KDD and UNSW-NB15 benchmarks. Bold values indicate the best performance. ✗ = Not available; ✓ = Available; ✗̸ = Partially available.

Method	Architectural Paradigm	Accuracy	Interpretability
NSL-KDD			
DL-IDS [30]	Standard Deep Neural Network	77.80%	✗
TMG-GAN [25]	Multi-generator GAN for data imbalance	79.32%	✗
AE+LSTM [23]	Sequential modeling with Autoencoder	79.40%	✗
AEDNN+UOSW [35]	Adversarial training with AE and DNN	78.26%	✗
HARMONI (Ours)	**Spatiotemporal Encoder + Hybrid Classification**	**80.19%**	✓
UNSW-NB15			
AE+PCA+LSTM [52]	Feature fusion with dimensionality reduction	76.28%	✗
XGB-FS-IDS [24]	XGBoost-based feature selection	75.60%	✗
MIM [28]	Multi-stage ensemble with feature selection	76.30%	✗
ROULETTE [14]	Attention-based CNN with multi-output learning	76.40%	✗̸
HARMONI (Ours)	**Spatiotemporal Encoder + Hybrid Classification**	**78.24%**	✓

On the NSL-KDD dataset, HARMONI achieves a superior accuracy of **80.19%**, which significantly outperforms competing methods by effectively addressing the distribution challenges inherent in network traffic. As visualized via UMAP in Section 4.2.1, this dataset exhibits substantial intra-class heterogeneity. While traditional architectures like DL-IDS struggle to resolve such complexity due to their reliance on single-modal feature extraction, HARMONI explicitly constrains the feature space through a Dynamic Class-Center Loss. This mechanism enforces intra-class compactness to directly optimize decision boundaries. The performance differential with TMG-GAN further elucidates the advantage of direct manifold optimization over indirect data augmentation, demonstrating that structural representation learning yields greater efficiency. Moreover, unlike the AE+LSTM baseline, the proposed adaptive gating mechanism enables sample-level dynamic feature weighting. When coupled with the raw-information compensation mechanism, this design preserves subtle statistical attributes often filtered by deep networks, thereby modeling complex attack patterns with higher fidelity.

Consistent performance gains are observed on the UNSW-NB15 dataset, where HARMONI leads with an accuracy of **78.24%** despite the increased diversity of modern attack vectors. This result validates the robustness of the framework against complex manifold structures that challenge other methods. For instance, the dimensionality reduction process in AE+PCA+LSTM risks losing non-linear features sensitive to novel attacks, a limitation circumvented by the end-to-end learning approach of HARMONI. Similarly, compared to feature engineering techniques like XGB-FS-IDS and MIM, the dual-branch spatiotemporal

encoder facilitates automated discriminative feature extraction, which drastically reduces the dependence on manual heuristics. Furthermore, regarding model transparency, while ROULETTE reveals simple feature correlations via attention maps, it struggles to capture complex non-linear influences and threshold effects. In contrast, HARMONI employs a multi-level SHAP framework to explicitly decode global importance, decision thresholds, and pairwise interactions, offering deeper and more objective diagnostic insights.

In summary, the empirical evaluations across the NSL-KDD and UNSW-NB15 datasets demonstrate that HARMONI consistently outperforms established baselines, achieving accuracies of 80.19% and 78.24%, respectively. These gains stem from the framework's spatiotemporal encoder, which captures intricate traffic patterns through dual-branch processing, and the Dynamic Class-Center Loss, which enhances intra-class compactness and inter-class separation to address manifold complexities. The raw-information compensation further preserves critical statistical features, enabling superior generalization to diverse attack vectors compared to single-modal, augmentation-based, or feature-selection reliant approaches. Overall, HARMONI's architectural synergies yield robust advancements in intrusion detection performance under standardized conditions.

4.2.4. Ablation Study

To evaluate the contribution of individual components in HARMONI, we conduct a leave-one-out ablation study on the NSL-KDD and UNSW-NB15 datasets. We isolate eight core components spanning feature processing (SPN), spatiotemporal modeling (SE, TM, ST), loss optimization (DCCL), information fusion (AGF), and architectural design (HA, RIC). As shown in Table 6, the full HARMONI framework consistently outperforms its ablated variants across almost all metrics, validating the efficacy of its modular synergy.

Table 6. Ablation study of key components in HARMONI (SPN: Shared Projection Network, SE: Spatial Encoder, TM: Temporal Modeler, ST: Spatial and Temporal components, AGF: Adaptive Gated Fusion, DCCL: Dynamic Class-Center Loss, HA: Hybrid Architecture, RIC: Raw Information Compensation). Bold values indicate the best performance.

Method	NSL-KDD			UNSW-NB15		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
HARMONI (Ours)	**83.36%**	**80.19%**	**76.59%**	80.51%	**78.24%**	**77.82%**
w/o SPN	81.00%	76.44%	71.59%	80.03%	75.36%	75.77%
w/o SE	81.37%	77.50%	73.06%	79.45%	75.67%	75.67%
w/o TM	82.40%	78.52%	73.93%	79.94%	76.30%	76.17%
w/o ST	80.09%	76.37%	71.49%	79.52%	75.93%	75.96%
w/o AGF	82.62%	78.23%	74.08%	79.51%	76.68%	76.59%
w/o DCCL	82.43%	77.83%	73.59%	80.50%	75.75%	76.48%
w/o HA	80.49%	77.97%	74.44%	**81.05%**	76.33%	76.58%
w/o RIC	81.66%	77.99%	74.42%	79.81%	77.62%	77.20%

The most severe performance degradation occurs when core feature processing modules are excluded. For instance, the omission of the Shared Projection Network (SPN) and the combined spatiotemporal components (ST) causes the F1-score on NSL-KDD to drop significantly to 71.59% and 71.49%, respectively. This indicates that unified feature projection and collaborative spatiotemporal modeling are indispensable for extracting high-fidelity representations from raw traffic. Beyond foundational feature extraction, the optimization and fusion mechanisms play a critical role in handling complex sample distributions. Specifically, removing the Dynamic Class-Center Loss (DCCL) results in a 2.49% drop in Recall on the UNSW-NB15 dataset. This empirical result strongly sup-

ports our topological analysis in Section 3.1, demonstrating that explicitly constraining intra-class compactness and inter-class separability effectively mitigates the issue of high class overlap. Similarly, omitting the Adaptive Gated Fusion (AGF) uniformly degrades performance across both datasets, which underscores its necessity for effective sample-level feature integration.

Regarding architectural refinement, modifying the Hybrid Architecture (HA) by reverting to a pure deep encoder without XGBoost presents a notable precision–recall trade-off. While the variant without HA achieves a slightly higher precision of 81.05% on UNSW-NB15, it suffers from inferior recall and F1-scores. This implies that the tree-based rectification in Stage-II mitigates the overfitting risks inherent in pure deep learning models and yields a more robust decision boundary. Furthermore, the Raw Information Compensation (RIC) serves as an effective complementary mechanism by recovering fine-grained attributes that are inevitably attenuated during deep semantic abstraction.

Overall, the ablation study confirms that the decoupled design of HARMONI is highly effective. By integrating robust spatiotemporal representation learning in Stage-I with an augmented hybrid classifier in Stage-II, the framework accurately models the complex and heterogeneous distributions of network intrusion traffic.

4.3. SHAP-Based Multi-Level Interpretability Analysis

4.3.1. Global Feature Importance Analysis

To quantify the global importance and directional influence of traffic features on the intrusion detection model, we apply SHAP TreeExplainer to the trained XGBoost classifier using the DDoS Attack SDN test set. Figure 4 presents the SHAP summary plot, where features are ranked according to their mean absolute SHAP values ($\mathbb{E}[|\phi|]$). The horizontal axis denotes the SHAP value (ϕ), where positive values increase the predicted probability of a DDoS attack and negative values correspond to benign traffic. Each point represents an individual sample, and the color encodes the normalized feature magnitude.

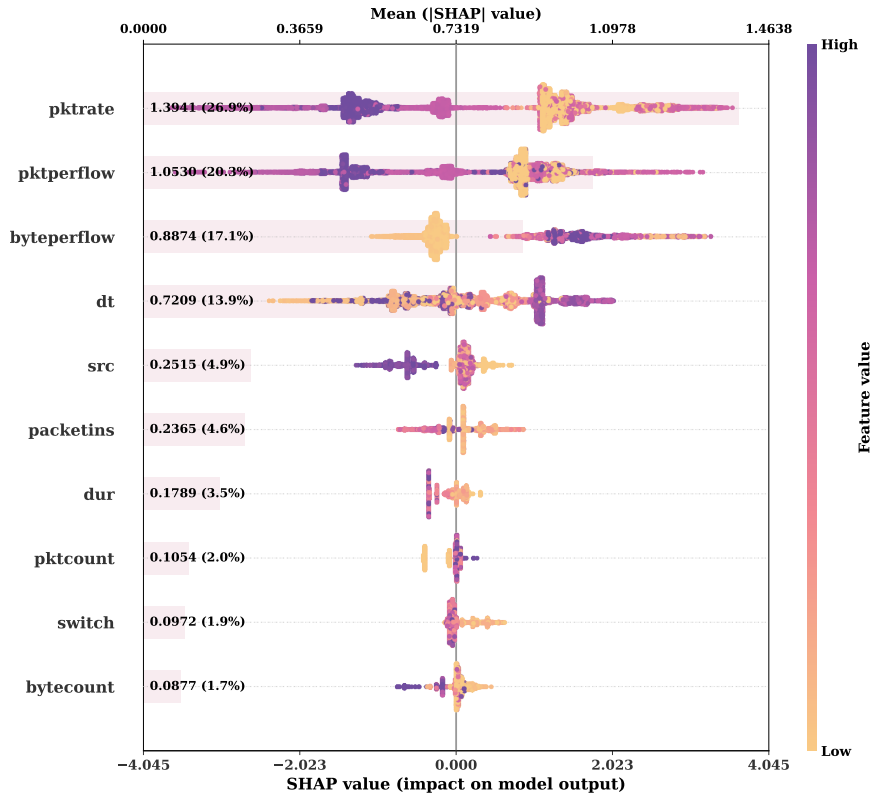


Figure 4. SHAP feature summary and importance plot.

According to $\mathbb{E}[|\phi|]$, the most influential features are *pktrate* (26.9%), *pktperflow* (20.3%), and *byteperflow* (17.1%), which together contribute 64.3% of the overall feature importance. The summary plot shows that lower values of *pktrate* and *pktperflow* tend to produce positive SHAP values, indicating a higher likelihood of attack classification. This pattern is consistent with the behavior of flow-exhaustion DDoS attacks such as SYN or UDP floods, where attackers generate a large number of short-lived flows. As a result, the packet rate and packet count per flow are significantly reduced compared to benign long-lived connections.

In contrast, *byteperflow* exhibits an opposite trend. Higher feature values are associated with positive SHAP values, suggesting that flows with unusually large payload sizes are more likely to be classified as attacks. This behavior reflects volumetric attack patterns designed to exhaust bandwidth resources. Together, these features capture both flow-exhaustion and volumetric characteristics commonly observed in DDoS traffic.

Interestingly, environmental attributes such as *dt* (timestamp) and *src* (encoded source IP) also exhibit relatively high importance. The contribution of *dt* suggests a temporal bias in the dataset, where attack traffic is concentrated in specific simulation periods. Similarly, the importance of *src* indicates that the model partially relies on specific source address patterns. Such dependencies are not directly observable from aggregate evaluation metrics such as F1-score, but can be revealed through feature attribution analysis.

Lower-ranked features, such as *packetins* (4.6% contribution), have comparatively minor influence on the model predictions and are therefore not analyzed further. These features likely correspond to simulation-specific SDN controller interactions rather than fundamental traffic behavior.

Overall, the SHAP analysis indicates that the model primarily relies on a small subset of flow-level statistics for decision making. In particular, the combination of low packet rates, low packet counts, and elevated payload volumes forms the dominant behavioral signature associated with DDoS traffic in the SDN environment. In addition, the analysis highlights the presence of dataset-specific temporal and source-related dependencies, demonstrating the usefulness of global interpretability techniques for auditing intrusion detection models. It should be noted that while the Stage-II classifier operates on a concatenated vector of 128-dimensional latent features and raw compensated features, our interpretability analysis focuses on the latter. This prioritization is based on the fact that raw statistical variables (e.g., packet rate, byte count) provide direct physical intuition for network security analysts, whereas individual dimensions of high-dimensional latent embeddings lack explicit semantic meaning. Consequently, the SHAP results presented here primarily decode the terminal decision logic in response to actionable network metrics. We acknowledge that this provides a localized view of the final ensemble's behavior rather than a complete transparency of the internal gating dynamics in Stage-I, which remains a subject for further granular investigation.

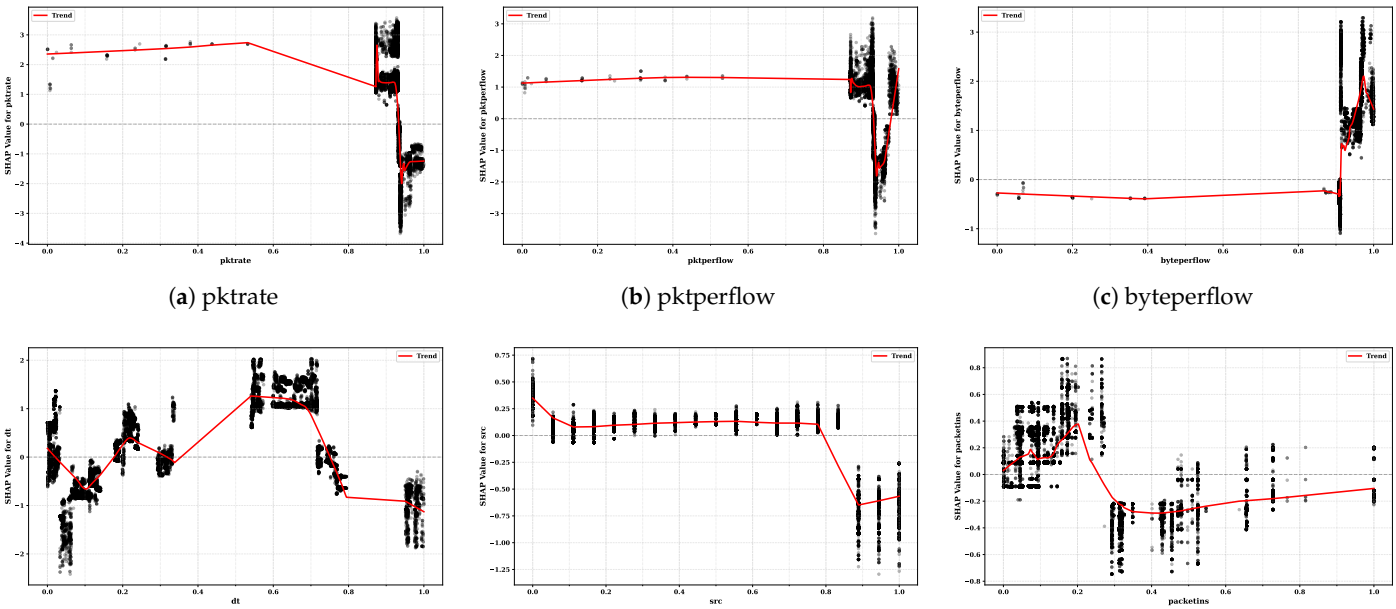
4.3.2. Feature Response Analysis

To further examine how feature magnitudes influence the model output, we analyze SHAP dependence plots for the six most influential features identified in the global importance analysis. These features collectively account for 90.9% of the total feature importance. The dependence plots map normalized feature values to their corresponding SHAP attributions (ϕ), revealing how variations in feature magnitude affect the model prediction. All features are min-max normalized to the $[0, 1]$ interval.

Figure 5a shows that *pktrate* exhibits a clear threshold-like behavior. In the normalized feature space, values below approximately 0.93 are associated with positive SHAP attributions ($\phi > 0$), indicating a higher likelihood of attack prediction. This behavior is consistent

with flow-exhaustion DDoS attacks such as TCP SYN floods, where a large number of short-lived flows leads to reduced packet rates per flow.

A related pattern is observed for *pktperflow*. As shown in Figure 5b, the dependence plot reveals a non-linear response within the high-value region. Lower normalized values ($\leq$0.93) tend to produce positive SHAP values, which align with connection-oriented flooding behavior. In contrast, the intermediate region ($0.93 < x \leq 0.98$) is dominated by negative SHAP values, corresponding to stable, long-lived benign connections. At the extreme high end (>0.98), SHAP values become positive again, suggesting that unusually large packet counts per flow may also indicate high-intensity attack traffic. Together, the response patterns of *pktrate* and *pktperflow* help explain how the model distinguishes different traffic behaviors.



(**a**) pktrate (**b**) pktperflow (**c**) byteperflow

(**d**) dt (**e**) src (**f**) packetins

Figure 5. SHAP dependence plots for the most influential features. Each dot represents an individual sample; darker colors indicate higher sample density, while lighter grey dots correspond to lower density regions.

The dependence plot for *byteperflow* (Figure 5c) reveals another distinct threshold. Normalized values exceeding approximately 0.92 are associated with positive SHAP values, indicating that flows with unusually large payload volumes contribute strongly to attack predictions. This pattern is consistent with volumetric attacks designed to exhaust bandwidth resources, such as oversized ICMP or heavy-payload UDP floods. In contrast, benign interactive protocols typically exhibit lower byte counts per flow.

In addition to behavioral features, environmental attributes such as *dt* (timestamp) and *src* (source IP) also exhibit noticeable attribution patterns. As shown in Figure 5d, the SHAP values of *dt* fluctuate across the normalized time range, suggesting that the model partially captures temporal patterns present in the dataset. Similarly, Figure 5e indicates a sharp transition around a normalized value of approximately 0.83 for *src*, implying that certain source address ranges are more frequently associated with attack predictions. These patterns highlight the presence of dataset-specific spatiotemporal dependencies, which may affect model generalization.

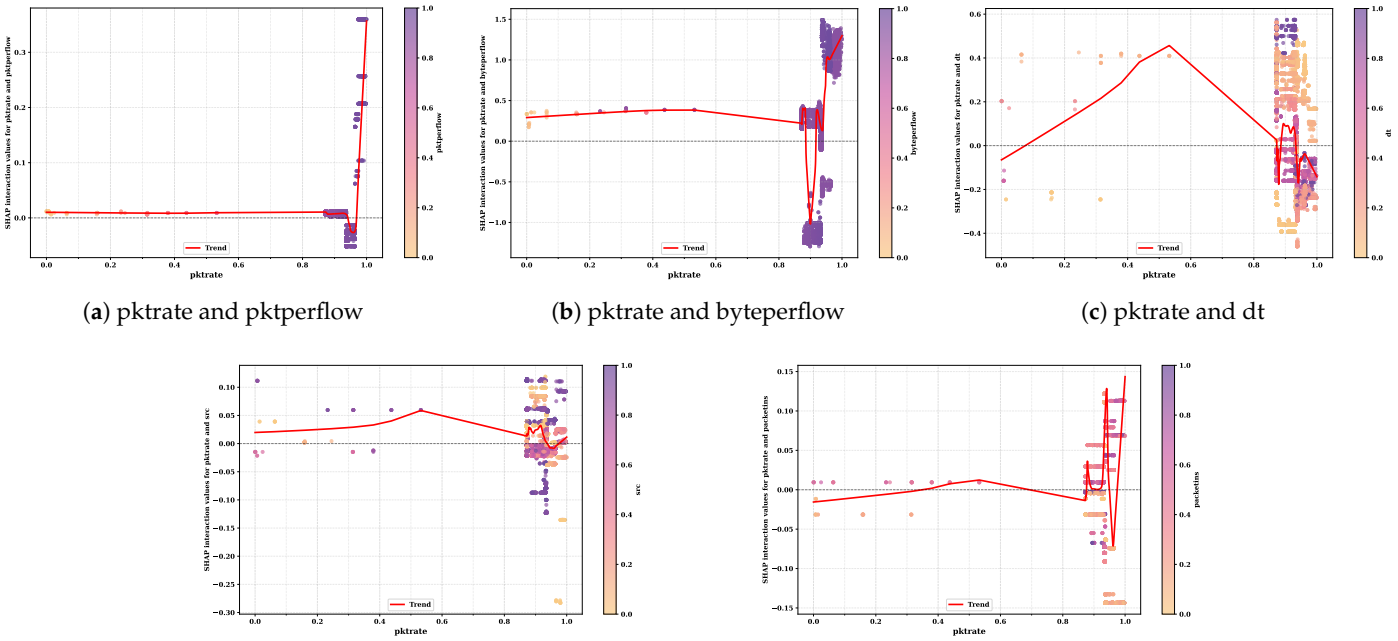
Finally, the dependence plot of *packetins* (Figure 5f) shows a moderate attribution shift around the normalized value of 0.3. This suggests that variations in controller interaction

frequency can influence the prediction, although the overall impact of this feature is smaller compared with the primary traffic statistics.

Overall, the feature response analysis reveals several non-linear relationships between traffic characteristics and model predictions. In particular, low packet rates, low packet counts, and unusually large payload volumes form common behavioral signatures associated with DDoS traffic. At the same time, the analysis exposes temporal and source-related dependencies learned from the dataset, demonstrating the usefulness of feature attribution techniques for auditing the decision patterns of intrusion detection models.

4.3.3. Feature Interaction Analysis

Building upon the univariate response patterns identified in the previous section, we further examine pairwise feature dependencies using SHAP interaction analysis centered on *pktrate*, the most influential global feature. Specifically, we compute SHAP interaction values $\Phi_{i,j}$ between *pktrate* and the top five auxiliary features. Figure 6 visualizes these interactions, where the x-axis represents the normalized *pktrate*, the y-axis denotes the interaction value, and the color gradient encodes the magnitude of the paired feature (yellow: low; purple: high). These plots reveal how flow-level statistics jointly influence the model's decision boundary.



(**a**) pktrate and pktperflow (**b**) pktrate and byteperflow (**c**) pktrate and dt

(**d**) pktrate and src (**e**) pktrate and packetins

Figure 6. SHAP interaction effects between *pktrate* and auxiliary features.

As illustrated in Figure 6a, the interaction between *pktrate* and *pktperflow* exhibits a multi-phase dependency across the *pktrate* spectrum. At lower *pktrate* values, the interaction is predominantly positive, indicating that elevated packet counts per flow reinforce the attack signal captured by reduced packet rate. As *pktrate* increases toward the mid-range, the interaction becomes negative, suggesting attenuation when both features approach values typical of benign traffic. At very high *pktrate* levels, the interaction shifts back to positive and intensifies when *pktperflow* simultaneously exhibits extreme values, consistent with the extreme-value behavior observed in the dependence analysis. This pattern indicates that the model emphasizes anomalous combinations of high packet rate and elevated per-flow packet volume, a characteristic often associated with persistent high-intensity DDoS traffic.

Figure 6b presents the interaction between *pktrate* and *byteperflow*. Within the attack-correlated region of *pktrate*, the interaction varies depending on the magnitude of *byteperflow*. When both features exhibit attack-indicative values, the interaction is moderately positive, reflecting complementary evidence. In contrast, when *byteperflow* indicates typical traffic behavior, the interaction becomes strongly negative, reflecting conflicting signals between the two indicators. When *pktrate* transitions to the benign regime, the interaction shifts to a consistently positive range. In this case, elevated *byteperflow* amplifies the predictive contribution despite the benign packet rate baseline, highlighting anomalous flow payload patterns that may correspond to volumetric attack variants.

In contrast, the interaction plots involving *dt*, *src*, and *packetins* (Figure 6c–e) exhibit diffuse distributions without consistent structural patterns. Consistent with the earlier global analysis, *dt* and *src* primarily capture dataset-specific temporal or spatial artifacts, whereas *packetins* contributes limited predictive influence. Consequently, these interactions likely reflect localized dataset characteristics rather than stable protocol-level behaviors, and therefore are not examined in further detail.

Overall, the interaction analysis highlights the importance of multivariate feature relationships in the detection process. Instead of relying on isolated thresholds, the model integrates packet rate indicators with flow-level volume statistics to capture diverse DDoS manifestations across varying traffic intensities. These observations complement the univariate analysis and provide insight into the model's decision logic, thereby strengthening the interpretability of its detection behavior in SDN-based environments.

5. Discussion

The experimental results demonstrate that HARMONI consistently outperforms representative baselines across multiple datasets. Rather than merely reporting these gains, we examine the underlying mechanisms that explain this performance. The UMAP visualizations in Section 4.2.1 reveal that network traffic data inherently exhibit severe intra-class heterogeneity and inter-class overlap, a structural challenge that conventional unimodal or fixed-fusion architectures struggle to address. Within HARMONI, the Dynamic Class-Center Loss explicitly counteracts this distributional complexity by imposing simultaneous intra-class compactness and inter-class separation constraints directly on the latent manifold. The ablation study corroborates this interpretation: removing the DCCL component results in a pronounced recall decline on the UNSW-NB15 dataset, where class ambiguity is particularly severe. Beyond the optimization objective, the adaptive gating module provides a complementary mechanism by dynamically re-weighting spatial and temporal representations on a per-sample basis. This design accommodates the heterogeneous modality signatures of diverse attack types, which range from spatially bursty floods to temporally persistent probing behaviors. The observed degradation when the gating mechanism is ablated confirms its contribution to representation flexibility. Furthermore, the Raw Information Compensation strategy preserves fine-grained protocol-level attributes that are inevitably attenuated during deep semantic encoding, and the tree-based ensemble classifier effectively leverages this augmented feature space to delineate robust, non-linear decision boundaries. Collectively, these architectural components form an integrated pipeline that directly targets the specific distributional and representational challenges identified in our data analysis.

It is necessary to acknowledge the limitations of the benchmark datasets employed in this study. NSL-KDD and UNSW-NB15, although widely adopted as standard evaluation benchmarks in the NIDS literature, were collected in environments that no longer fully reflect contemporary network landscapes. Their attack diversity, traffic volume, and protocol composition differ from those observed in modern enterprise or cloud infrastructures.

Nevertheless, their continued use in this work serves two purposes. First, they provide a common comparative ground against the extensive body of prior research that relies on these same datasets, ensuring that the reported performance margins are both interpretable and reproducible. Second, the topological complexity revealed by our UMAP analysis confirms that even these relatively classical datasets pose nontrivial distributional challenges for existing models; the ability of HARMONI to address these challenges therefore carries methodological significance beyond the specific datasets. Future evaluation on more diverse and recently collected traffic corpora remains an essential step to further validate the generalization capability of the proposed framework.

Regarding deployment practicality, certain considerations for operational network monitoring should be noted. The inference pipeline of HARMONI, once trained, is relatively lightweight: the deep encoder performs a single forward pass, and the subsequent XGBoost prediction is computationally inexpensive, suggesting that per-flow detection latency could be compatible with near-real-time requirements. However, the current framework is designed for offline training and static inference; it does not incorporate an online learning mechanism to adapt to the concept drift inherent in live traffic. While the exponential moving average update of class centroids offers a conceptual starting point, substantial challenges remain in extending this to a fully incremental learning setting. Additionally, the SHAP-based interpretability analysis, though valuable for retrospective auditing and alert triage, is computationally intensive and is not intended for per-packet or high-throughput real-time explanation. These practical constraints delineate the current application scope of HARMONI and motivate further investigation into lightweight, adaptive deployment strategies.

In addition, this work adopts the standard fixed feature order of public network intrusion detection datasets, and further exhaustive feature permutation and sensitivity ablation experiments will be supplemented in future research to quantitatively verify the model's robustness to feature order variation.

6. Conclusions

This paper addresses the challenge of achieving both high detection effectiveness and transparent decision reasoning in deep learning-based network intrusion detection. To address this problem, we present HARMONI, a two-stage hybrid learning framework that integrates adaptive spatiotemporal representation learning, dynamic metric optimization, and a raw information compensated ensemble classifier. In the first stage, convolutional and recurrent encoders capture spatial and temporal characteristics of network traffic, while an adaptive gating mechanism aggregates heterogeneous representations and a Dynamic Class-Center Loss improves feature compactness and class separability. In the second stage, the learned latent representation is combined with preprocessed raw traffic features and processed by an XGBoost classifier to strengthen non-linear decision boundaries and preserve fine grained statistical patterns that may be attenuated in deep representations. Extensive experiments and ablation analyses demonstrate that HARMONI consistently improves detection performance while maintaining stable generalization across datasets. Furthermore, the SHAP-based interpretability analysis provides systematic insights into feature importance, decision patterns, and feature interactions, enabling a transparent understanding of model behavior. Future work should investigate online adaptation for streaming environments, robustness against adversarial traffic manipulation, lightweight deployment strategies, and more rigorous evaluation protocols such as time-aware or source-disjoint train/test splits to further audit model robustness. Overall, this work contributes a principled framework for accurate and interpretable intrusion detection, offering

practical insights into the integration of deep representation learning and explainable security analytics.

Author Contributions: Conceptualization, R.H. and Z.Z.; Methodology, R.H.; Validation, R.H., Z.Z. and M.L.; Formal analysis, R.H. and Y.W.; Investigation, R.H. and M.L.; Resources, L.W.; Writing—original draft, R.H.; Writing—review and editing, R.H., Z.Z. and L.W.; Visualization, R.H.; Supervision, L.W.; Project administration, L.W.; Funding acquisition, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (Grant No. 62572136 and No. W2511073).

Data Availability Statement: The data presented in this study are available on request.

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

References

1. Yu, T.; Liu, G.; Wang, C.; Yang, Y. LLMGRAPH: Label-free Detection against APTs in Edge Networks via LLM and GCN. *IEEE Trans. Dependable Secur. Comput.* **2025**, *22*, 7256–7271. [CrossRef]
2. Sameera, N.; Tripathy, R.; Maram, B.; Venkatakotireddy, G.; Muppidi, S.; Velpuru, M.S. Wide residual quantum dilated convolutional neural network for zero-day attack detection. *Eng. Appl. Artif. Intell.* **2026**, *166*, 113404. [CrossRef]
3. Zhang, X.; Zhu, Y.; Li, X.; Zhang, Y.; Niu, W.; Xu, F.; He, J.; Yan, R.; Huang, S. Active cybersecurity: Vision, model, and key technologies. *Front. Inf. Technol. Electron. Eng.* **2025**, *26*, 1243–1278. [CrossRef]
4. Zhang, C.; Hu, P.; Tan, L. Spatiotemporal Feature Correlation with Feature Space Transformation for Intrusion Detection. *Appl. Sci.* **2025**, *15*, 11168. [CrossRef]
5. Dilini, N.; Sun, N.; Miao, Y.; Moustafa, N. A Comprehensive Review on Graph-Based Anomaly Detection: Approaches for Intrusion Detection. *Appl. Sci.* **2026**, *16*, 1906. [CrossRef]
6. Torre, D.; Chennamaneni, A.; Jo, J.; Vyas, G.; Sabrsula, B. Toward enhancing privacy preservation of a federated learning CNN intrusion detection system in iot: Method and empirical study. *ACM Trans. Softw. Eng. Methodol.* **2025**, *34*, 1–48. [CrossRef]
7. Deore, B.; Bhosale, S. Intrusion detection system based on RNN classifier for feature reduction. *SN Comput. Sci.* **2022**, *3*, 114.
8. Devendiran, R.; Turukmane, A.V. Dugat-LSTM: Deep learning based network intrusion detection system using chaotic optimization strategy. *Expert Syst. Appl.* **2024**, *245*, 123027. [CrossRef]
9. Ansari, M.S.; Bartoš, V.; Lee, B. GRU-based deep learning approach for network intrusion alert prediction. *Future Gener. Comput. Syst.* **2022**, *128*, 235–247. [CrossRef]
10. Bamber, S.S.; Katkuri, A.V.R.; Sharma, S.; Angurala, M. A hybrid CNN-LSTM approach for intelligent cyber intrusion detection system. *Comput. Secur.* **2025**, *148*, 104146. [CrossRef]
11. Manocchio, L.D.; Layeghy, S.; Lo, W.W.; Kulatilleke, G.K.; Sarhan, M.; Portmann, M. Flowtransformer: A transformer framework for flow-based network intrusion detection systems. *Expert Syst. Appl.* **2024**, *241*, 122564. [CrossRef]
12. Akuthota, U.C.; Bhargava, L. Transformer based intrusion detection for iot networks. *IEEE Internet Things J.* **2025**, *12*, 6062–6067. [CrossRef]
13. Liu, Y.; Wu, L. Intrusion Detection Model Based on Improved Transformer. *Appl. Sci.* **2023**, *13*, 6251. [CrossRef]
14. Andresini, G.; Appice, A.; Caforio, F.P.; Malerba, D.; Vessio, G. ROULETTE: A neural attention multi-output model for explainable network intrusion detection. *Expert Syst. Appl.* **2022**, *201*, 117144. [CrossRef]
15. Ma, Z.; Liu, Y.; Chen, Y.; Liu, Z.; Li, Y. XMF-GNN: A cross-modality dynamic fusion heterogeneous graph neural network for network intrusion detection. *Neurocomputing* **2025**, *655*, 131285.
16. Khan, W.; Ahmad, J.; Alasbali, N.; Al Mazroa, A.; Alshehri, M.S.; Khan, M.S. A novel transformer-based explainable AI approach using SHAP for intrusion detection in vehicular ad hoc networks. *Comput. Netw.* **2025**, *270*, 111575. [CrossRef]
17. Dong, J.; Chen, H.; Shen, S.; Xu, H.; Liu, Z. Interpretable Intrusion Detection for IoT Security: A SHAP-Enhanced NGBoost Model. *Comput. Netw.* **2025**, *275*, 111937.
18. Zhang, Z.; Zeng, L.; Zhu, D.; Tan, H.; Wang, L.; Li, Z.; Gu, Z. Gbfkan: An adaptive multi-layer interpretable architecture for intrusion detection in various internet of things scenarios. *IEEE Internet Things J.* **2025**, *12*, 30379–30397.
19. Zhang, Z.; Wang, L.; Chen, G.; Gu, Z.; Tian, Z.; Du, X.; Guizani, M. STG2P: A two-stage pipeline model for intrusion detection based on improved LightGBM and K-means. *Simul. Model. Pract. Theory* **2022**, *120*, 102614.
20. Shyaa, M.A.; Ibrahim, N.F.; Zainol, Z.; Abdullah, R.; Anbar, M.; Alzubaidi, L. Evolving cybersecurity frontiers: A comprehensive survey on concept drift and feature dynamics aware machine and deep learning in intrusion detection systems. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109143. [CrossRef]

21. Nandanwar, H.; Katarya, R. Deep learning enabled intrusion detection system for Industrial IOT environment. *Expert Syst. Appl.* **2024**, *249*, 123808. [CrossRef]
22. Mutalib, N.H.A.; Sabri, A.Q.M.; Wahab, A.W.A.; Abdullah, E.R.M.F.; AlDahoul, N. Explainable deep learning approach for advanced persistent threats (APTs) detection in cybersecurity: A review. *Artif. Intell. Rev.* **2024**, *57*, 297. [CrossRef]
23. Mushtaq, E.; Zameer, A.; Umer, M.; Abbasi, A.A. A two-stage intrusion detection system with auto-encoder and LSTMs. *Appl. Soft Comput.* **2022**, *121*, 108768.
24. Kasongo, S.M.; Sun, Y. Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. *J. Big Data* **2020**, *7*, 105. [CrossRef]
25. Ding, H.; Sun, Y.; Huang, N.; Shen, Z.; Cui, X. TMG-GAN: Generative adversarial networks-based imbalanced learning for network intrusion detection. *IEEE Trans. Inf. Forens. Secur.* **2023**, *19*, 1156–1167.
26. Costa, V.G.; Pedreira, C.E. Recent advances in decision trees: An updated survey. *Artif. Intell. Rev.* **2023**, *56*, 4765–4800.
27. Al-Yaseen, W.L.; Idrees, A.K.; Almasoudy, F.H. Wrapper feature selection method based differential evolution and extreme learning machine for intrusion detection system. *Pattern Recognit.* **2022**, *132*, 108912. [CrossRef]
28. Zhang, Z.; Wang, L.; Zhu, J.; Zhu, D.; Gu, Z.; Zhang, Y. MIM: A multiple integration model for intrusion detection on imbalanced samples. *World Wide Web* **2024**, *27*, 47. [CrossRef]
29. Hammad, M.; Hewahi, N.; Elmedany, W. MMM-RF: A novel high accuracy multinomial mixture model for network intrusion detection systems. *Comput. Secur.* **2022**, *120*, 102777. [CrossRef]
30. Vinayakumar, R.; Alazab, M.; Soman, K.P.; Poornachandran, P.; Al-Nemrat, A.; Venkatraman, S. Deep learning approach for intelligent intrusion detection system. *IEEE Access* **2019**, *7*, 41525–41550. [CrossRef]
31. Cao, Y.; Jiang, H.; Deng, Y.; Wu, J.; Zhou, P.; Luo, W. Detecting and mitigating DDoS attacks in SDN using spatial-temporal graph convolutional network. *IEEE Trans. Dependable Secur. Comput.* **2021**, *19*, 3855–3872.
32. Neto, E.C.P.; Iqbal, S.; Buffett, S.; Sultana, M.; Taylor, A. Deep learning for intrusion detection in emerging technologies: A comprehensive survey and new perspectives. *Artif. Intell. Rev.* **2025**, *58*, 340. [CrossRef]
33. Fu, M.; Wang, P.; Liu, S.; Chen, X.; Zhou, X. FIR-GNN: A Graph Neural Network using Flow Interaction Relationships for Intrusion Detection of Consumer Electronics in Smart Home Network. *IEEE Trans. Consum. Electron.* **2025**, *71*, 4892–4902.
34. Kheddar, H. Transformers and large language models for efficient intrusion detection systems: A comprehensive survey. *Inf. Fusion* **2025**, *124*, 103347. [CrossRef]
35. Yang, H.; Zeng, R.; Xu, G.; Zhang, L. A network security situation assessment method based on adversarial deep learning. *Appl. Soft Comput.* **2021**, *102*, 107096. [CrossRef]
36. Aceto, G.; Ciuonzo, D.; Montieri, A.; Persico, V.; Pescapé, A. AI-powered internet traffic classification: Past, present, and future. *IEEE Commun. Mag.* **2023**, *62*, 168–175. [CrossRef]
37. Charmet, F.; Tanuwidjaja, H.C.; Ayoubi, S.; Gimenez, P.F.; Han, Y.; Jmila, H.; Blanc, G.; Takahashi, T.; Zhang, Z. Explainable artificial intelligence for cybersecurity: A literature survey. *Ann. Telecommun.* **2022**, *77*, 789–812. [CrossRef]
38. Calzarossa, M.C.; Giudici, P.; Zieni, R. An assessment framework for explainable AI with applications to cybersecurity. *Artif. Intell. Rev.* **2025**, *58*, 150. [CrossRef]
39. Moustafa, N.; Koroniotis, N.; Keshk, M.; Zomaya, A.Y.; Tari, Z. Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions. *IEEE Commun. Surv. Tut.* **2023**, *25*, 1775–1807. [CrossRef]
40. Vilone, G.; Longo, L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion* **2021**, *76*, 89–106. [CrossRef]
41. Durojaye, H.; Naiseh, M. Explainable AI for intrusion detection systems: A model development and experts' evaluation. In *Proceedings of the Intelligent Systems Conference*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 301–318.
42. Wen, Y.; Zhang, K.; Li, Z.; Qiao, Y. A discriminative feature learning approach for deep face recognition. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 499–515.
43. Hadsell, R.; Chopra, S.; LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006*; IEEE: New York, NY, USA, 2006; Volume 2, pp. 1735–1742.
44. He, K.; Fan, H.; Wu, Y.; Xie, S.; Girshick, R. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9729–9738.
45. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
46. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794.
47. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 10–12 November 2015*; IEEE: New York, NY, USA, 2015; pp. 1–6.

48. Tavallaee, M.; Bagheri, E.; Lu, W.; Ghorbani, A.A. A detailed analysis of the KDD CUP 99 data set. In *Proceedings of the 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, Canada, 8–10 July 2009*; IEEE: New York, NY, USA, 2009; pp. 1–6.
49. Ahuja, N.; Singal, G.; Mukhopadhyay, D. DDOS attack SDN dataset. *Mendeley Data* **2020**, *1*, 17632.
50. Sharafaldin, I.; Lashkari, A.H.; Hakak, S.; Ghorbani, A.A. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In *Proceedings of the 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, 1–3 October 2019*; IEEE: New York, NY, USA, 2019; pp. 1–8.
51. Castellano-Escuder, P.; Zachman, D.K.; Han, K.; Hirschey, M.D. GAUDI: Interpretable multi-omics integration with UMAP embeddings and density-based clustering. *Nat. Commun.* **2025**, *16*, 5771. [CrossRef] [PubMed]
52. Thakkar, A.; Kikani, N.; Geddam, R. Fusion of linear and non-linear dimensionality reduction techniques for feature reduction in LSTM-based Intrusion Detection System. *Appl. Soft Comput.* **2024**, *154*, 111378.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.