INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 21, Issue: 4, Month: August, Year: 2026
Article Number: 7480, https://doi.org/10.15837/ijccc.2026.4.7480

AMoE-IDS: An Adaptive Mixture-of-Experts Framework for Cross-Dataset Intrusion Detection

Ouail Mjahed, Soukaina Mjahed

Ouail Mjahed*
Faculty of Sciences and Technology, Department of Computer Sciences
L2IS Laboratory, Cadi Ayyad University
Marrakech 40000, Morocco.
*Corresponding author: ouail.mjahed@ced.uca.ma

Soukaina Mjahed
Faculty of Sciences Semlalia, Department of Computer Sciences
LISI Laboratory, Cadi Ayyad University
Marrakech 40000, Morocco.
s.mjahed@uca.ac.ma

Abstract

Intrusion Detection Systems (IDS) are essential for securing modern network infrastructures against increasingly sophisticated cyber threats. While deep learning-based IDS have shown promising performance, most existing approaches rely on static and monolithic architectures that struggle to adapt to heterogeneous environments such as Internet of Things (IoT) systems, enterprise networks, and mixed traffic scenarios. Moreover, conventional ensemble and hybrid methods typically employ fixed fusion strategies, limiting their ability to exploit input-dependent specialization. To address these limitations, this paper proposes an *Adaptive Mixture-of-Experts Intrusion Detection System (AMoE-IDS)*, a hybrid deep learning framework that integrates a shared feature encoder, multiple specialized expert networks, and an adaptive gating mechanism. The shared encoder learns a unified latent representation from heterogeneous feature spaces, while the gating network dynamically routes each input to the most relevant experts, enabling conditional computation and improved detection performance. Extensive experiments conducted on three recent benchmark datasets, CICIoT2023, CSE-CICIDS 2018, and TII-SSRC-23, demonstrate that AMoE-IDS consistently outperforms conventional deep learning and hybrid IDS models. The proposed framework achieves F_1-scores of 99.19%, 99.67%, and 99.68% and AUC values of 0.992, 0.991, and 0.990 on CICIoT2023, CSE-CICIDS 2018, and TII-SSRC-23, respectively. Despite its multi-expert architecture, the model maintains low inference latency ranging from 1.28 to 1.56 ms per network flow, supporting real-time deployment. Cross-dataset evaluation confirms the robustness of AMoE-IDS under distribution shifts, while ablation studies highlight the critical role of feature harmonization and adaptive expert selection. Statistical significance analysis further validates the reliability of the observed improvements. Overall, the proposed framework demonstrates competitive performance, good scalability, and improved cross-dataset generalization.

Keywords: Intrusion Detection Systems, Deep Learning, Mixture-of-Experts, Cross-Dataset Generalization, Cybersecurity.

1 Introduction

The rapid expansion of interconnected systems, including cloud infrastructures, Internet of Things (IoT) ecosystems, web services, and emerging paradigms such as 5G and edge computing, has significantly increased the volume, heterogeneity, and complexity of network traffic. This evolution has been accompanied by a surge in sophisticated cyber threats, ranging from large-scale distributed denial-of-service (DDoS) attacks and malware campaigns to stealthy multi-stage intrusions and zero-day exploits. Consequently, Intrusion Detection Systems (IDS) have become indispensable for ensuring the security and resilience of modern networked environments [1, 2].

Traditional IDS approaches, particularly signature-based and rule-based systems, remain effective for detecting known threats but exhibit limited capability in identifying novel or evolving attacks. This limitation has driven the adoption of machine learning (ML) and deep learning (DL) techniques, which enable data-driven detection by learning patterns directly from network traffic [3].

Recent advances in deep learning have led to the development of powerful IDS models. Convolutional Neural Networks (CNNs) capture spatial correlations among traffic features, while Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks model temporal dependencies in sequential data. More recently, attention mechanisms and Transformer-based architectures have demonstrated superior capability in modeling long-range dependencies and complex feature interactions in high-dimensional traffic data [4, 5].

Despite these advances, most existing DL-based IDS adopt monolithic architectures, applying a single computational pipeline to all inputs. Such designs lack adaptability to heterogeneous network environments and often suffer from performance degradation when deployed across different datasets or domains. This limitation is particularly critical given the diversity of modern IDS datasets, which span IoT, web applications, and enterprise or industrial environments, each characterized by distinct traffic distributions and attack patterns [6].

To enhance detection performance, hybrid and ensemble-based IDS have been extensively investigated [7, 8, 9]. While these approaches effectively exploit the complementary strengths of multiple models, most existing methods rely on static fusion mechanisms, such as majority voting or fixed-weight averaging, which fail to adapt to input-dependent variability and limit conditional specialization capabilities [10].

In this context, the Mixture-of-Experts (MoE) paradigm offers a promising direction by enabling conditional computation through dynamic expert selection. By activating specialized sub-models based on input characteristics, MoE architectures enhance both scalability and adaptability. Although MoE has gained significant attention in large-scale deep learning systems [11, 12], its application to intrusion detection remains relatively under-explored.

Motivated by these challenges, this paper proposes an *Adaptive Mixture-of-Experts Intrusion Detection System (AMoE-IDS)* that combines a shared feature encoder with multiple specialized experts and a dynamic gating mechanism. The proposed framework is designed to improve detection accuracy and enhance generalization across heterogeneous datasets.

To ensure a comprehensive evaluation across diverse network environments, experiments are conducted on three complementary datasets: CICIoT2023 for IoT traffic, CSE-CIC-IDS2018 for mixed network and application-layer attacks, and TII-SSRC-23 for heterogeneous network scenarios.

The main contributions of this work are summarized as follows:

- We propose a novel adaptive Mixture-of-Experts-based IDS that dynamically selects expert models based on input traffic characteristics.
- We introduce a shared encoder to learn unified latent representations across heterogeneous datasets.
- We demonstrate, through extensive experiments on recent IDS benchmarks, that the proposed model achieves superior performance in terms of accuracy, F_1-score, and AUC compared to state-of-the-art DL and ensemble approaches.
- We conduct comprehensive ablation studies to analyze the impact of key architectural components on performance and generalization.

The remainder of this paper is organized as follows. Section 2 reviews recent advances in deep learning-based intrusion detection and Mixture-of-Experts architectures. Section 3 presents the proposed AMoE-IDS framework, including feature harmonization, the shared encoder, the adaptive gating mechanism, and heterogeneous experts. Section 4 describes the datasets, preprocessing pipeline, evaluation protocol, and implementation details. Section 5 reports the experimental results, including performance comparison, per-category analysis, cross-dataset evaluation, ablation studies, statistical significance analysis, and state-of-the-art comparisons. Section 6 examines the computational complexity, scalability, and deployment feasibility of the framework. Section 7 discusses threats to validity, limitations, and future research directions. Finally, Section 8 concludes the paper.

2 Related Works

This section reviews recent advances in intrusion detection systems, focusing on deep learning-based approaches, hybrid architectures, and evaluations on modern IDS datasets.

2.1 Deep Learning-Based Intrusion Detection

Early IDS relied on traditional machine learning algorithms such as Support Vector Machines, Decision Trees, and Random Forests. While effective on structured datasets, these methods require manual feature engineering and struggle with high-dimensional and large-scale traffic data [1]. Deep learning has significantly advanced intrusion detection by enabling automatic feature extraction. CNN-based models capture spatial dependencies among traffic features, whereas RNN and LSTM architectures model temporal dynamics in network flows. More recently, Transformer-based models have emerged as a powerful alternative, leveraging self-attention mechanisms to capture long-range dependencies and complex feature interactions [13, 14]. However, most DL-based IDS remain monolithic, applying a fixed architecture to all inputs. Such designs limit adaptability and often exhibit poor generalization across heterogeneous datasets, a key challenge for real-world deployment [6].

2.2 Hybrid and Ensemble IDS

Recent studies have explored hybrid and ensemble learning strategies to address the limitations of single-model IDS, particularly in terms of robustness and stability across diverse traffic conditions. These approaches often integrate heterogeneous architectures (e.g., convolutional, recurrent, and attention-based models) to capture complementary spatial and temporal patterns in network data [15, 16, 17, 18]. However, despite their improved predictive performance, most ensemble-based IDS rely on predefined and static aggregation schemes, which do not account for the dynamic and heterogeneous nature of network traffic, thereby limiting their adaptability in real-world environments [19]. As a result, their ability to adapt to input-specific characteristics remains limited, potentially leading to suboptimal performance in real-world and cross-domain scenarios. In addition, the combination of multiple models often increases computational complexity without achieving true conditional specialization.

2.3 IDS on Recent Benchmark Datasets

Recent IDS research increasingly focuses on realistic and large-scale datasets that reflect modern network environments.

CICIoT2023 is a large-scale IoT dataset comprising traffic from numerous devices and covering a wide range of attack categories. Recent studies report high performance using gradient boosting, LSTM, and hybrid deep learning models, with accuracies often exceeding 98% [20].

CSE-CIC-IDS2018 is a widely used benchmark for intrusion detection due to its realistic attack scenarios. CNN models achieve around 96.00% accuracy [21], while Hybrid Convolutional Recurrent Neural Networks (HCRNN) reach 97.75% [22]. Ensemble methods remain competitive, with Random Forest achieving 99.04% accuracy and an F_1-score of 98.83% [23]. These results highlight strong performance but also the need for more adaptive models in heterogeneous environments.

TII-SSRC-23 represents heterogeneous environments, including IoT and industrial systems. Studies show that deep and hybrid models outperform traditional ML approaches, achieving accuracies above 99% while highlighting challenges in cross-domain generalization [24, 25].

2.4 Research Gaps

Despite significant progress, several challenges remain. First, most IDS models rely on fixed architectures that lack adaptability to heterogeneous traffic. Second, many studies focus on single-dataset evaluation, limiting their ability to generalize across domains. Third, existing ensemble methods use static fusion strategies that fail to exploit input-dependent specialization.

To address these limitations, adaptive architectures based on conditional computation are required. The proposed AMoE-IDS framework leverages the Mixture-of-Experts paradigm to enable dynamic expert selection, improving both robustness and cross-dataset generalization in diverse intrusion detection scenarios.

3 Methodology

This section presents the proposed *Adaptive Mixture-of-Experts Intrusion Detection System (AMoE-IDS)*. The architecture integrates representation learning, conditional expert specialization, and adaptive decision fusion to improve detection performance and cross-dataset generalization. The design builds upon established principles in deep representation learning and Mixture-of-Experts (MoE), which enable scalable modeling through conditional computation and expert specialization [11, 12, 26].

3.1 Overall Architecture

The proposed AMoE-IDS is designed to address heterogeneity in network traffic and improve generalization across diverse intrusion detection datasets. As illustrated in Figure 1, the architecture consists of four main components: i) a preprocessing and feature harmonization module, ii) a shared feature encoder, iii) a set of specialized expert networks, and iv) an adaptive gating and fusion mechanism.

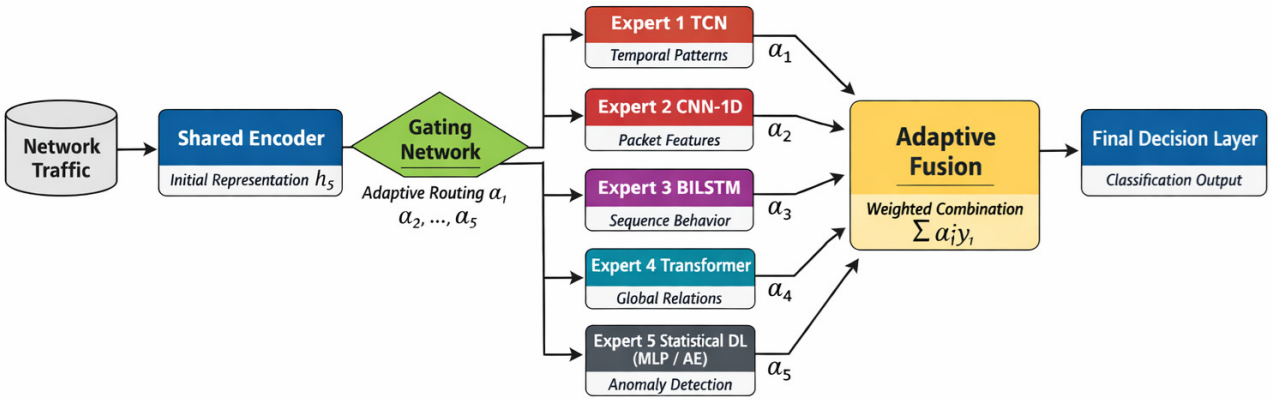


Figure 1: AMoE-IDS architecture: traffic features are encoded into a shared latent space, dynamically routed to specialized experts via a gating mechanism, and fused to produce the final prediction.

Raw network traffic features are first normalized and mapped into a unified representation. The shared encoder projects heterogeneous input features into a common latent space, which is subsequently processed by multiple expert models. The gating network dynamically assigns importance weights to each expert based on the input embedding, and the final decision is obtained through adaptive weighted fusion. This conditional computation mechanism enables input-dependent routing and improves model capacity without increasing computational cost linearly [11, 12].

3.2 Feature Harmonization and Preprocessing

The datasets considered in this work originate from heterogeneous network environments and exhibit substantial differences in feature definitions, dimensionality, and traffic distributions. To enable unified learning and cross-dataset generalization, AMoE-IDS employs a feature harmonization mechanism that projects semantically equivalent traffic attributes into a common representation space.

Let $\mathbf{x} \in \mathbb{R}^{d_i}$ denote an input feature vector extracted from dataset i, where d_i may vary across datasets. A harmonized feature subset composed of common flow-level statistics, packet-level descriptors, temporal characteristics, and protocol-related attributes is constructed and used as input to the framework.

Numerical features are normalized and categorical attributes are encoded to obtain a consistent representation across datasets. Missing harmonized features are assigned default values to preserve a fixed input dimensionality. The resulting feature vector is then forwarded to the shared encoder, which learns a dataset-invariant latent representation suitable for expert specialization and adaptive routing.

A detailed description of the preprocessing pipeline, feature selection strategy, and normalization procedure is provided in Section 4.2.

3.3 Shared Encoder

The shared encoder learns a dataset-invariant latent representation capturing intrinsic traffic patterns. Formally, the encoder function $E(\cdot)$ maps the input vector $\mathbf{x}$ to a latent embedding $\mathbf{z} \in \mathbb{R}^{d_z}$:

$$\mathbf{z} = E(\mathbf{x}) = \sigma(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e), \tag{1}$$

where $\sigma(\cdot)$ is a nonlinear activation function. Learning compact latent representations is a core principle of deep learning, enabling abstraction and improved generalization [27].

3.4 Expert Networks

Let $\{f_k(\cdot)\}_{k=1}^{K}$ denote a set of heterogeneous expert networks. The expert pool consists of five heterogeneous experts designed to capture complementary traffic characteristics:

(1) A Temporal Convolutional Network (TCN) expert for short-term temporal patterns and bursty traffic behaviors. (2) A 1D-CNN expert for extracting local spatial correlations among flow features. (3) A BiLSTM expert for modeling long-range sequential dependencies. (4) A Transformer expert for capturing global contextual relationships through self-attention. (5) A lightweight Multilayer Perceptron (MLP) expert for modeling statistical interactions among flow-level attributes.

For future extensions and ablation analyses, alternative expert architectures may be incorporated, such as autoencoders (AE) and other DL models.

Given the latent representation $\mathbf{z}$, the output of expert k is defined as

$$\mathbf{y}_k = f_k(\mathbf{z}), \quad k = 1, \ldots, K. \tag{2}$$

3.5 Adaptive Gating Network

The gating network dynamically determines the contribution of each expert. Given the latent embedding $\mathbf{z}$, the gating function $G(\cdot)$ produces a weight vector $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = G(\mathbf{z}) = \mathrm{softmax}(\mathbf{W}_g\mathbf{z} + \mathbf{b}_g), \tag{3}$$

subject to

$$\sum_{k=1}^{K} \alpha_k = 1, \quad \alpha_k \geq 0. \tag{4}$$

This mechanism enables input-dependent expert selection and is central to conditional computation in MoE systems [11, 12].

3.6 Adaptive Fusion and Final Prediction

The final prediction is computed as a weighted combination of expert outputs:

$$\hat{\mathbf{y}} = \sum_{k=1}^{K} \alpha_k \mathbf{y}_k. \tag{5}$$

This formulation allows the model to adaptively integrate expert knowledge based on input characteristics.

3.7 Training Objective

The model is trained end-to-end using supervised learning. The classification loss is defined as categorical cross-entropy:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i). \tag{6}$$

To encourage balanced expert utilization and prevent expert collapse, an entropy-based regularization term is introduced:

$$\mathcal{L}_{gate} = -\lambda \sum_{k=1}^{K} \alpha_k \log(\alpha_k). \tag{7}$$

The overall objective is:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{gate}. \tag{8}$$

Such regularization is commonly used to promote diversity in mixture models and avoid degenerate solutions.

3.8 Training Procedure

Model parameters are optimized using the Adam optimizer, which provides adaptive learning rates and efficient convergence in deep neural networks [28]. All components are trained jointly in an end-to-end manner, and early stopping is employed to mitigate overfitting.

3.9 Computational Complexity and Real-Time Feasibility

The computational complexity depends on the number of experts and the gating mechanism. By leveraging conditional computation, only a subset of experts contributes significantly to each prediction, improving efficiency without sacrificing model capacity [12]. This makes AMoE-IDS suitable for real-time intrusion detection.

3.10 Algorithmic Summary

Algorithm 1 summarizes the training and inference procedure of the proposed AMoE-IDS.

4 Experimental Setup

This section describes the datasets, preprocessing pipeline, experimental protocol, baseline models, implementation details, and evaluation metrics used to assess the performance of the proposed AMoE-IDS framework.

Algorithm 1 Adaptive Mixture-of-Experts Intrusion Detection System (AMoE-IDS)

Require: Labeled training dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$
Require: Number of experts K, Learning rate η, Regularization λ, Max epochs E
Ensure: Trained AMoE-IDS model parameters
Initialize shared encoder parameters θ_E
Initialize expert parameters $\{\theta_k\}_{k=1}^{K}$
Initialize gating network parameters θ_G
for epoch = 1 to E **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
    for each sample $(x, y) \in \mathcal{B}$ **do**
      Feature Harmonization: normalize and encode $\mathbf{x}$
      $z \leftarrow E(x; \theta_E)$ (*Shared Encoder*)
      for expert $k = 1$ to K **do**
        $y_k \leftarrow f_k(z; \theta_k)$ (*Expert Inference*)
      end for
      $\alpha \leftarrow G(z; \theta_G)$ (*Gating Weights*)
      $\hat{y} \leftarrow \sum_{k=1}^{K} \alpha_k y_k$ (*Adaptive Fusion*)
      $\mathcal{L}_{cls} \leftarrow -y \log(\hat{y})$
      $\mathcal{L}_{gate} \leftarrow -\lambda \sum_{k=1}^{K} \alpha_k \log(\alpha_k)$
      $\mathcal{L} \leftarrow \mathcal{L}_{cls} + \mathcal{L}_{gate}$
    end for
    Update $\theta_E, \{\theta_k\}_{k=1}^{K}, \theta_G$ using Adam optimizer
  end for
end for
Inference Phase:
Given unseen sample x_{test}: $z_{test} \leftarrow E(x_{test})$, $\alpha_{test} \leftarrow G(z_{test})$, $\hat{y}_{test} \leftarrow \sum_{k=1}^{K} \alpha_k^{test} f_k(z_{test})$
Return Trained AMoE-IDS model

4.1 Datasets Description

To evaluate the performance and generalization capability of AMoE-IDS, experiments are conducted on three complementary intrusion detection benchmarks: CICIoT2023, CSE-CIC-IDS2018, and TII-SSRC-23. These datasets cover diverse environments, including IoT networks, enterprise infrastructures, and heterogeneous traffic scenarios.

4.1.1 CICIoT2023

CICIoT2023 contains large-scale IoT traffic collected from 105 heterogeneous devices [29]. It includes approximately 46 million flow records described by 47 features and labeled into 34 classes grouped into 6 attack categories, covering DDoS/DoS, reconnaissance, spoofing, brute-force, web attacks, and botnet behaviors.

4.1.2 CSE-CIC-IDS2018

CSE-CIC-IDS2018 is a widely used benchmark capturing realistic enterprise traffic over multiple days [30]. It contains around 16 million flow records with approximately 80 features extracted using CICFlowMeter. In this work, fine-grained attack labels are grouped into 15 classes and 6 higher-level categories to enable consistent cross-dataset evaluation.

4.1.3 TII-SSRC-23

TII-SSRC-23 is a recent dataset designed for heterogeneous environments [24]. It consists of 27.5 GB of PCAP traffic converted into flow-level records with 75–86 features and 32 traffic subtypes organized into 6 main categories.

4.2 Data Preprocessing and Feature Harmonization

The three datasets considered in this study, namely CICIoT2023, CSE-CICIDS2018, and TII-SSRC-23, exhibit substantial differences in traffic characteristics, feature definitions, attack taxonomies, and class distributions. Therefore, a unified preprocessing and feature harmonization pipeline was developed to ensure fair comparison, reproducibility, and effective cross-dataset learning.

4.2.1 Data Cleaning and Dataset Partitioning

For each dataset, duplicate entries, incomplete records, and corrupted samples were removed prior to any preprocessing operation. To prevent data leakage, all datasets were first partitioned using stratified sampling into training (70%), validation (15%), and testing (15%) subsets while preserving class distributions.

Importantly, all subsequent preprocessing procedures, including oversampling, feature selection, and normalization, were fitted exclusively on the training data and then applied to the validation and test sets. This protocol follows recommended practices for machine learning experimentation and prevents information leakage between training and evaluation data.

Class imbalance was addressed using the Synthetic Minority Over-sampling Technique (SMOTE) [31]. SMOTE was applied only to the training partition after dataset splitting. Neither validation nor test samples were involved in the oversampling process.

Table 1 summarizes the class distributions used in the experiments.

Table 1: Class distribution for the reduced datasets used in the AMoE-IDS experiments

	CICIoT2023			**CSE-CICIDS 2018**			**TII-SSRC-23**		
Label	**Class**	**Train**	**Test**	**Class**	**Train**	**Test**	**Class**	**Train**	**Test**
C_1	Benign	140000	60000	Benign	120000	60000	Benign	130000	56000
C_2	DoS/DDoS	70000	30000	DoS/DDoS	55000	22000	DoS/DDoS	59000	25000
C_3	Brute Force	14000	6000	Brute Force	18000	7500	Brute Force	14000	6000
C_4	Web Attacks	10500	4500	Web Attacks	15000	6000	Exploits	10500	4500
C_5	Mirai/Spoof.	7000	3000	Bot	12000	5000	Botnet/Malware	9000	4000
C_6	Reconnaissance	21000	9000	Infiltration	9000	3500	Net Scan	17000	7000
	Total	**262500**	**112500**	**Total**	**229000**	**104000**	**Total**	**239500**	**102500**

4.2.2 Attack Category Harmonization

Since the original datasets contain different attack labels and taxonomies, a common attack categorization scheme was adopted to facilitate cross-dataset evaluation. Similar attack types were grouped into unified categories according to their operational objectives and behavioral characteristics.

The resulting harmonized categories include: Benign Traffic, DoS/DDoS Attacks, Brute-Force Attacks, Web-Based Attacks, Reconnaissance and Scanning, Botnet/Malware Activities, Exploitation and Infiltration Attacks.

This mapping enables meaningful performance comparison across heterogeneous network environments while preserving the semantic consistency of attack behaviors.

4.2.3 Feature Harmonization Across Datasets

One of the main challenges of cross-dataset intrusion detection lies in the heterogeneity of feature spaces. CICIoT2023 contains 47 flow-based features, CSE-CICIDS2018 provides approximately 80 flow descriptors, while TII-SSRC-23 includes between 75 and 86 traffic statistics depending on the extraction configuration.

To construct a unified feature space, semantically equivalent flow-level attributes were identified across the three datasets. Only features that were either directly available or consistently derivable from all datasets were retained. The harmonization process focused on traffic statistics that are independent of specific network infrastructures and attack implementations.

Table 2 presents examples of the mapping between original dataset features and the harmonized feature representation adopted by AMoE-IDS.

Categorical attributes were transformed using one-hot encoding, while missing harmonized features were assigned zero values when unavailable in a particular dataset.

Table 2: Feature harmonization across the three datasets

Harmonized Feature	**CICIoT2023**	**CSE-CICIDS2018**	**TII-SSRC-23**
Flow Duration	flow_duration	Flow Duration	flow_duration
Total Fwd Packets	Tot Fwd Pkts	Total Fwd Packets	Tot Fwd Pkts
Total Bwd Packets	Tot Bwd Pkts	Total Backward Packets	Tot Bwd Pkts
Flow Bytes/s	Flow Bytes/s	Flow Bytes/s	Flow Bytes/s
Flow Packets/s	Flow Pkts/s	Flow Packets/s	Flow Pkts/s
Fwd Packet Length Mean	Fwd Pkt Len Mean	Fwd Packet Length Mean	Fwd Pkt Len Mean
Bwd Packet Length Mean	Bwd Pkt Len Mean	Bwd Packet Length Mean	Bwd Pkt Len Mean
Fwd IAT Mean	Fwd IAT Mean	Fwd IAT Mean	Fwd IAT Mean
Bwd IAT Mean	Bwd IAT Mean	Bwd IAT Mean	Bwd IAT Mean
SYN Flag Count	SYN Flag Cnt	SYN Flag Count	SYN Flag Cnt
ACK Flag Count	ACK Flag Cnt	ACK Flag Count	ACK Flag Cnt

4.2.4 Feature Normalization

Network traffic features exhibit highly heterogeneous numerical ranges. To avoid dominance of large-scale attributes and improve optimization stability, numerical features were normalized using Min–Max scaling:

$$x'_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{9}$$

where the minimum and maximum values were estimated exclusively from the training partition and subsequently applied to validation and testing data.

This normalization procedure improves convergence behavior and facilitates feature alignment across datasets.

4.2.5 Feature Selection

Feature selection was performed exclusively on the training data using a three-stage hybrid procedure combining statistical filtering and model-based evaluation.

First, low-variance features were removed because they provide limited discriminative information. Second, highly correlated attributes were eliminated using Pearson correlation analysis with a threshold of 0.9 in order to reduce redundancy. Finally, the remaining features were ranked according to both Random Forest importance scores and mutual information values.

The final feature subset consisted of traffic descriptors that consistently exhibited high predictive relevance across all datasets. The selected attributes mainly describe: flow statistics, packet size characteristics, traffic rate indicators, temporal dynamics, protocol flag information, header-level statistics, derived traffic metrics.

Overall, the feature selection process reduced the dimensionality of the original datasets by approximately 40% while preserving their discriminative capability.

Table 3 summarizes the harmonized feature categories retained in all experiments.

4.2.6 Cross-Dataset Feature Alignment

The proposed preprocessing and harmonization pipeline serves two complementary objectives. First, it guarantees a leakage-free experimental protocol by strictly separating training and evaluation data throughout the preprocessing stages. Second, it creates a common feature representation that enables effective knowledge transfer across heterogeneous datasets.

Table 3: Harmonized and selected flow-based features across datasets

Category	**Selected Features**
Flow Statistics	Flow Duration, Total Fwd Packets, Total Bwd Packets
Packet Size	Fwd Packet Length Mean, Bwd Packet Length Mean, Packet Length Std
Traffic Rate	Flow Bytes/s, Flow Packets/s
Temporal Dynamics	Fwd IAT Mean, Bwd IAT Mean
Protocol Flags	SYN Flag Count, ACK Flag Count
Header Information	Fwd Header Length, Bwd Header Length
Derived Metrics	Avg Packet Size, Down/Up Ratio

By combining attack-category harmonization, feature alignment, normalization, and hybrid feature selection, the resulting representation provides a compact and transferable feature space suitable for evaluating the cross-dataset generalization capability of the proposed AMoE-IDS framework.

4.3 Experimental Protocol

Two complementary evaluation protocols are adopted to assess both in-distribution performance and cross-dataset generalization.

Intra-dataset evaluation: Each dataset is split into training (70%), validation (15%), and testing (15%) subsets using stratified sampling.

Cross-dataset evaluation: The model is trained on one dataset and directly evaluated on another unseen dataset without additional fine-tuning.

In addition, ten-fold cross-validation is performed to improve statistical robustness of the reported results.

4.4 Baseline Models

To evaluate the effectiveness of the proposed approach, AMoE-IDS is compared with several representative IDS architectures commonly used in the literature: Deep Neural Network (DNN), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), CNN–LSTM hybrid model, Autoencoder (AE), and Transformer.

All baseline models are trained using the same preprocessing pipeline and experimental protocol to ensure fair comparison.

4.5 Implementation Details

All experiments are conducted using PyTorch 2.2 on a workstation equipped with an NVIDIA RTX A6000 GPU (48 GB VRAM).

The proposed AMoE-IDS architecture consists of a shared feature encoder followed by $K = 5$ expert subnetworks and an adaptive gating mechanism.

Shared Encoder: Two fully connected layers with ReLU activations projecting the input features into a 128-dimensional latent representation.

Expert Networks: Five heterogeneous experts operating on the shared latent representation:

- TCN Expert: two temporal convolution blocks with dilation rates $\{1, 2\}$ and 64 filters.
- CNN Expert: two 1D convolution layers (64 and 128 filters, kernel size = 3) followed by global average pooling.
- BiLSTM Expert: two bidirectional LSTM layers with 128 hidden units.
- Transformer Expert: two encoder layers with 8 attention heads and feed-forward dimension 256.
- MLP Expert: three fully connected layers with dimensions 128-64-32 and ReLU activations.

Gating Network: A two-layer feedforward network with softmax activation producing adaptive weights for expert fusion.

Optimizer: Adam optimizer with learning rate 10^{-4}.

Batch Size: 256 samples.

Training Epochs: Up to 100 epochs with early stopping based on validation loss.

Regularization: Gating regularization coefficient $\lambda = 0.1$ to encourage balanced expert utilization.

4.6 Evaluation Metrics and Statistical Analysis

The performance of the evaluated models is assessed using standard classification metrics commonly adopted in intrusion detection tasks as *Accuracy* (*Acc*), *Precision* (P), *Recall* (R), F_1-score and Area Under the ROC Curve (*AUC*)

Accuracy measures the overall proportion of correctly classified instances. Precision and recall evaluate the correctness and completeness of positive predictions, respectively, while the F_1-score provides a harmonic balance between them. The *AUC* metric reflects the model's ability to discriminate between classes across different decision thresholds.

Given the class imbalance inherent in intrusion detection datasets, macro-averaged metrics are reported to ensure equal importance across all classes, regardless of their frequency. This provides a more reliable evaluation compared to accuracy alone.

To assess the practical applicability of the proposed framework, inference latency per sample is also measured, reflecting the model's suitability for real-time intrusion detection.

To ensure the robustness and statistical significance of the experimental results, each experiment is repeated multiple times with different random initializations. The reported performance metrics correspond to the mean and standard deviation over N independent runs (with $N = 10$).

Statistical significance is evaluated using the paired *t-test* and the non-parametric *Wilcoxon signed-rank test* at a significance level of $\alpha = 0.05$. To account for multiple comparisons, the *Holm–Bonferroni correction* is applied to control the family-wise error rate.

These statistical analyses ensure that the observed performance differences between models are not due to random variations but reflect consistent improvements.

5 Experimental Results

This section evaluates the proposed AMoE-IDS framework and compares it with six baseline models: DNN, CNN, LSTM, CNN-LSTM, Autoencoder-based IDS, and Transformer-based IDS. Experiments are conducted on three recent IDS benchmark datasets: CICIoT2023, CSE-CICIDS 2018, and TII-SSRC-23. All reported results correspond to the average performance over ten independent runs with different random seeds and are expressed as mean ± standard deviation.

5.1 Overall Performance Comparison

Table 4 reports the overall detection performance across the three datasets in terms of accuracy, recall, F_1-score, AUC, and inference latency.

The results reported in Table 4 show that AMoE-IDS consistently outperforms all baseline models across the three evaluated datasets.

On *CICIoT2023*, AMoE-IDS achieves an F_1-score of 98.99%, improving upon the strongest baseline (Transformer-IDS, 94.94%) by approximately 4.05%, while also reaching the highest AUC (0.992), indicating strong discriminative capability.

On *CSE-CIC-IDS2018*, the proposed model attains an F_1-score of 99.67%, outperforming the best baseline (92.42%) by 7.25%, highlighting its effectiveness in modeling complex enterprise-level traffic.

Similarly, on *TII-SSRC-23*, AMoE-IDS achieves an F_1-score of 99.68%, compared to 93.29% for the strongest baseline, yielding a gain of 6.39% and confirming its robustness in heterogeneous environments.

Table 4: Performance comparison of AMoE-IDS and baseline models (mean ± std over 10 runs)

Model	*Acc*(%)	*R*(%)	F_1(%)	*AUC*	*Latency* (*ms*)
			CICIoT2023		
DNN	92.13 ± 0.42	91.52 ± 0.51	91.73 ± 0.41	0.962 ± 0.003	1.05 ± 0.02
CNN	93.14 ± 0.35	92.62 ± 0.49	92.74 ± 0.32	0.968 ± 0.002	1.12 ± 0.02
LSTM	93.73 ± 0.37	93.19 ± 0.43	93.26 ± 0.34	0.971 ± 0.002	1.18 ± 0.03
CNN-LSTM	94.25 ± 0.36	93.73 ± 0.36	93.95 ± 0.35	0.974 ± 0.002	1.25 ± 0.03
Autoencoder	92.81 ± 0.41	91.92 ± 0.53	92.08 ± 0.44	0.965 ± 0.003	1.08 ± 0.02
Transformer	95.18 ± 0.36	94.53 ± 0.39	94.61 ± 0.33	0.980 ± 0.002	1.32 ± 0.03
AMoE-IDS	$\mathbf{99.11 \pm 0.21}$	$\mathbf{98.87 \pm 0.23}$	$\mathbf{98.76 \pm 0.23}$	$\mathbf{0.991 \pm 0.001}$	$\mathbf{1.28 \pm 0.02}$
			CSE-CICIDS 2018		
DNN	89.71 ± 0.51	88.93 ± 0.61	89.03 ± 0.55	0.948 ± 0.004	1.03 ± 0.02
CNN	90.52 ± 0.42	89.82 ± 0.42	89.94 ± 0.44	0.954 ± 0.003	1.10 ± 0.02
LSTM	91.23 ± 0.44	90.64 ± 0.43	90.75 ± 0.45	0.959 ± 0.003	1.17 ± 0.03
CNN-LSTM	91.84 ± 0.43	91.15 ± 0.42	91.36 ± 0.46	0.962 ± 0.003	1.22 ± 0.02
Autoencoder	91.15 ± 0.53	91.91 ± 0.52	90.46 ± 0.47	0.951 ± 0.003	1.05 ± 0.02
Transformer	92.54 ± 0.39	91.93 ± 0.33	92.06 ± 0.36	0.968 ± 0.002	1.30 ± 0.03
AMoE-IDS	$\mathbf{99.52 \pm 0.26}$	$\mathbf{99.35 \pm 0.29}$	$\mathbf{99.41 \pm 0.26}$	$\mathbf{0.989 \pm 0.002}$	$\mathbf{1.45 \pm 0.03}$
			TII-SSRC-23		
DNN	90.81 ± 0.46	90.14 ± 0.51	90.21 ± 0.46	0.955 ± 0.003	1.04 ± 0.02
CNN	91.52 ± 0.47	90.93 ± 0.42	91.02 ± 0.44	0.961 ± 0.003	1.11 ± 0.02
LSTM	92.13 ± 0.38	91.52 ± 0.44	91.63 ± 0.32	0.965 ± 0.003	1.18 ± 0.03
CNN-LSTM	92.74 ± 0.35	92.01 ± 0.45	92.24 ± 0.31	0.968 ± 0.003	1.23 ± 0.02
Autoencoder	91.35 ± 0.45	90.62 ± 0.47	90.85 ± 0.42	0.960 ± 0.003	1.07 ± 0.02
Transformer	93.46 ± 0.34	92.82 ± 0.36	92.96 ± 0.33	0.972 ± 0.002	1.29 ± 0.02
AMoE-IDS	$\mathbf{99.37 \pm 0.31}$	$\mathbf{99.35 \pm 0.34}$	$\mathbf{99.36 \pm 0.32}$	$\mathbf{0.988 \pm 0.002}$	$\mathbf{1.56 \pm 0.03}$

Across all datasets, the low standard deviation (±0.21–±0.34) indicates stable training. Despite its multi-expert architecture, AMoE-IDS maintains competitive inference latency (1.28–1.56 ms), comparable to Transformer-based models. These results demonstrate that the adaptive mixture-of-experts mechanism effectively improves detection performance and generalization through input-dependent expert specialization, without data leakage.

5.2 ROC Curve Analysis

To further evaluate the discriminative capability of the proposed model, ROC curves are plotted for representative datasets. Figure 2 compares AMoE-IDS with the strongest baseline (Transformer, Encoder and LSTM-CNN).

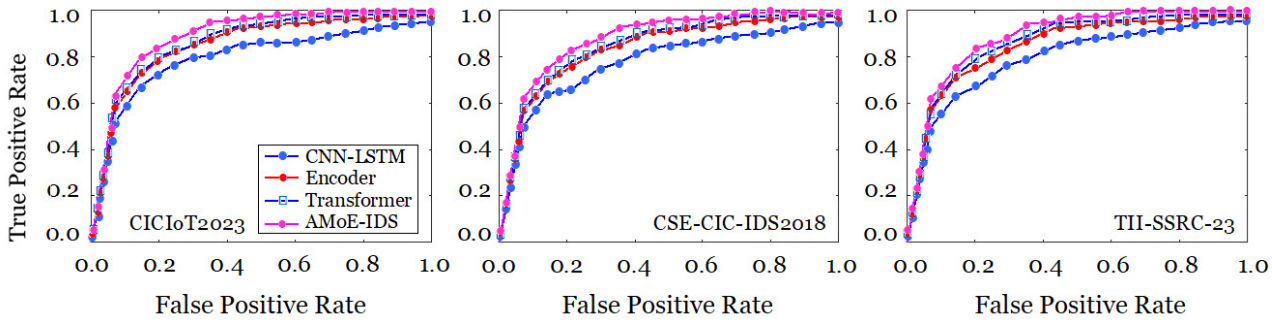


Figure 2: ROC curves comparing AMoE-IDS and strongest baseline on representative datasets.

As shown in Figure 2, AMoE-IDS consistently outperforms the baseline across all operating points. The curves exhibit a steeper rise near the origin, indicating improved detection performance under low false positive rates, which is critical for practical intrusion detection systems. These observations are consistent with the AUC results reported in Table 4.

5.3 Per-Category Performance and Confusion Matrix Analysis

To further analyze the detection capability of the evaluated models, a category-level evaluation is conducted using both F_1-scores and confusion matrices. Since the datasets contain different numbers of classes, the analysis is performed at the attack-category level.

Table 5 reports the average F_1-scores on CICIoT2023, showing that AMoE-IDS consistently outperforms the strongest baseline across all categories. Improvements are observed for both high-volume attacks (e.g., DoS/DDoS) and more challenging categories such as brute force, reconnaissance, and data exfiltration.

Table 5: Per-category F_1-score comparison on the CICIoT2023 dataset (mean $\pm$ std)

Traffic Category	Transformer-IDS	AMoE-IDS
Benign Traffic	94.71 ± 0.32	99.51 ± 0.29
DoS/DDoS Attacks	95.12 ± 0.31	99.15 ± 0.35
Brute Force	91.33 ± 0.26	96.77 ± 0.23
Web Attacks	90.02 ± 0.28	94.72 ± 0.26
Mirai/Spoofing	87.86 ± 0.42	91.56 ± 0.44
Reconnaissance	93.11 ± 0.37	97.91 ± 0.31

To complement this quantitative analysis, Figure 3 presents the confusion matrices (in %) for the three datasets.

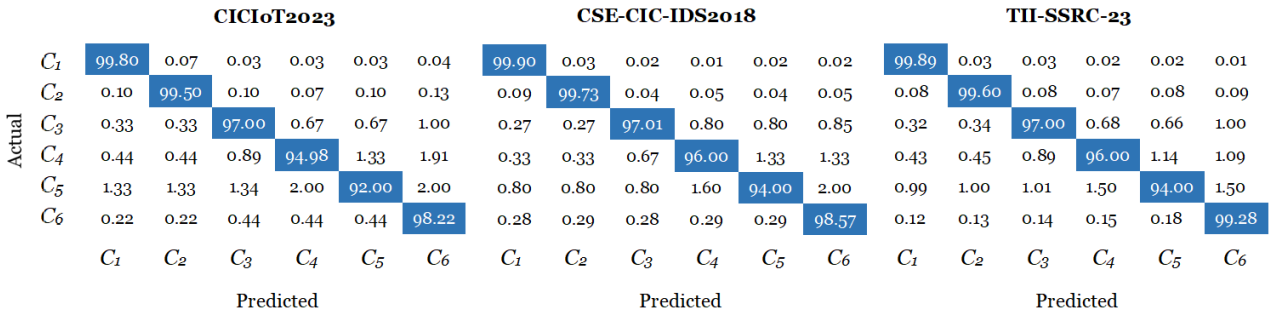


Figure 3: Confusion matrices of AMoE-IDS across the three datasets. The model achieves high classification accuracy with most predictions concentrated along the diagonal.

The confusion matrices for CICIoT2023, CSE-CICIDS2018, and TII-SSRC-23 show strong diagonal dominance with limited off-diagonal errors, highlighting the ability of the proposed AMoE-IDS model to accurately distinguish between traffic categories while maintaining consistent performance across diverse and heterogeneous datasets. Overall, the combined analysis demonstrates that the proposed model not only improves average performance but also enhances robustness across both dominant and minority attack categories.

5.4 Cross-Dataset Evaluation

To assess the generalization capability of the proposed AMoE-IDS framework under realistic deployment conditions, cross-dataset experiments were conducted. In this setting, the model is trained on one dataset and directly evaluated on another unseen dataset without any fine-tuning. Such experiments provide a more challenging and realistic evaluation than conventional intra-dataset testing, as they expose the model to significant variations in traffic characteristics, feature distributions, and attack behaviors.

5.4.1 Shared Attack Categories and Evaluation Protocol

The three datasets considered in this study exhibit different attack taxonomies and labeling schemes. To ensure a fair comparison across domains, a common label space was constructed us-

ing the attack categories shared by all datasets: C_1: Benign Traffic, C_2: DoS/DDoS Attacks, C_3: Brute-Force Attacks, and C_4: Web-Based Attacks.

The cross-dataset evaluation was therefore restricted to these shared categories. This protocol avoids introducing inconsistencies caused by dataset-specific attack labels and allows a meaningful assessment of transfer learning capabilities.

5.4.2 Handling Non-Shared Attack Categories

Several attack categories do not have direct semantic equivalents across the three datasets. For example, Mirai botnet and spoofing attacks in CICIoT2023, infiltration attacks in CSE-CIC-IDS2018, and certain malware-related behaviors in TII-SSRC-23 are dataset-specific and cannot be mapped reliably to a common label space.

To avoid biased comparisons, samples belonging to non-shared categories were excluded from the quantitative cross-dataset evaluation. Consequently, the reported results focus exclusively on attack categories that can be consistently interpreted across all datasets. While this restriction reduces the number of evaluated classes, it ensures that performance differences reflect domain shifts rather than labeling inconsistencies.

5.4.3 Overall Cross-Dataset Performance

Table 6 reports the resulting macro F_1-scores (mean $\pm$ std over ten runs) for all train–test combinations.

Table 6: Cross-dataset evaluation results on shared categories (C_1–C_4) (macro F_1-score, mean $\pm$ std)

Train $\rightarrow$ Test	CICIoT2023	TII-SSRC-23	CSE-CIC-IDS2018
CICIoT2023	–	0.921 ± 0.006	0.894 ± 0.007
TII-SSRC-23	0.947 ± 0.004	–	0.912 ± 0.006
CSE-CIC-IDS2018	0.883 ± 0.008	0.901 ± 0.007	–

The highest transfer performance is obtained when training on TII-SSRC-23 and testing on CICIoT2023, achieving a macro F_1-score of 0.947. This result suggests that the diversity of traffic patterns and attack behaviors present in TII-SSRC-23 enables the model to learn highly transferable representations. In contrast, the lowest performance is observed when transferring from CSE-CIC-IDS2018 to CICIoT2023 (0.883), reflecting the substantial domain gap between enterprise network traffic and IoT environments.

Nevertheless, AMoE-IDS consistently achieves macro F_1-scores above 0.88 across all transfer scenarios, demonstrating strong robustness to distribution shifts.

5.4.4 Per-Category Cross-Dataset Analysis

Although macro F_1-score provides a global measure of transfer performance, it may conceal difficulties associated with specific attack categories. Table 7 therefore reports category-level F_1-scores for representative transfer scenarios.

Table 7: Per-category cross-dataset F_1-score (%) for AMoE-IDS

Train $\rightarrow$ Test	Benign	DoS/DDoS	Brute Force	Web Attacks
TII-SSRC-23 $\rightarrow$ CICIoT2023	98.2	96.4	94.8	89.5
CICIoT2023 $\rightarrow$ CSE-CIC-IDS2018	97.6	92.8	90.7	86.5
CSE-CIC-IDS2018 $\rightarrow$ TII-SSRC-23	96.8	93.1	91.4	88.2

The results reveal that benign traffic and DoS/DDoS attacks transfer particularly well across domains, with F_1-scores consistently exceeding 92%. These categories exhibit relatively stable statistical characteristics regardless of the underlying network environment. In contrast, web attacks represent the most challenging category. Their lower F_1-scores indicate that application-layer attack behaviors

differ substantially across datasets, making generalization more difficult. Brute-force attacks occupy an intermediate position, achieving relatively stable performance across all transfer scenarios.

5.4.5 Cross-Dataset Confusion Matrix Analysis

To further analyze the behavior of AMoE-IDS under domain shifts, Fig. 4 presents the confusion matrices obtained for the three cross-dataset transfer scenarios. These visualizations complement the macro F_1-scores reported in Table 6 by providing a class-level view of the transfer performance.

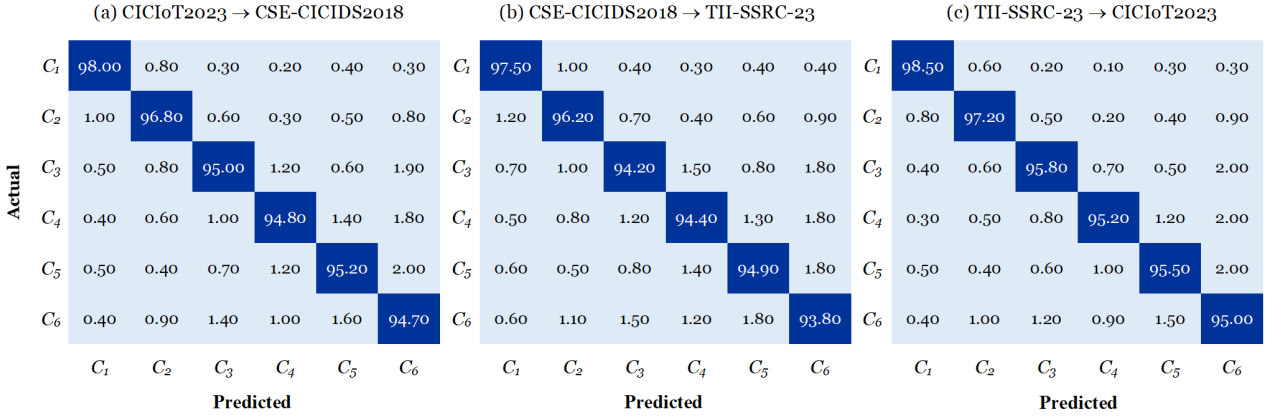


Figure 4: Confusion matrices obtained under cross-dataset evaluation using the shared traffic categories (C_1–C_6).

As shown in Fig. 4, the majority of samples are correctly classified along the main diagonal for all transfer scenarios, indicating that the representations learned by AMoE-IDS remain largely transferable across heterogeneous network environments. The strongest transfer performance is observed for the TII-SSRC-23 → CICIoT2023 scenario, which is consistent with the highest macro F_1-score reported in Table 6.

Across all transfers, benign traffic (C_1) and DoS/DDoS attacks (C_2) exhibit the highest classification consistency, suggesting that these categories are characterized by relatively stable and transferable traffic patterns. In contrast, the largest confusion occurs among Brute Force (C_3), Web Attacks (C_4), and Reconnaissance-related activities (C_6), which often share overlapping behavioral characteristics and are represented differently across datasets.

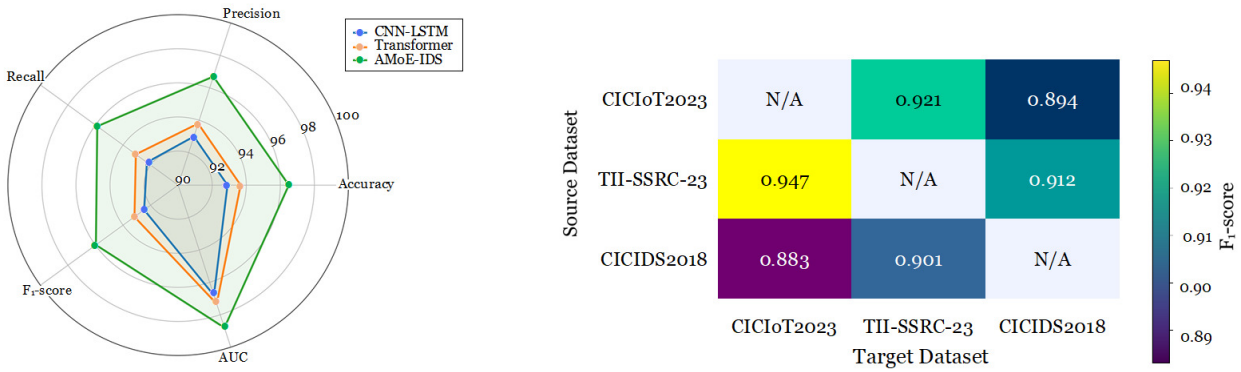
Despite these ambiguities, the overall confusion patterns remain limited, confirming that the shared encoder successfully learns dataset-invariant representations while the adaptive expert-routing mechanism improves robustness to distribution shifts. These results provide additional evidence that AMoE-IDS maintains effective class discrimination even when evaluated on previously unseen datasets.

To improve readability, Figure 5 provides visual summaries of the main experimental findings. The radar chart highlights the overall performance of AMoE-IDS relative to the strongest baseline models, while the heatmap illustrates its cross-dataset generalization capability across different train–test scenarios.

5.4.6 Discussion

Overall, the cross-dataset results demonstrate that AMoE-IDS successfully mitigates the impact of domain shifts through the combined action of the shared encoder and adaptive expert routing mechanism. The shared encoder learns transferable feature representations, while the gating network dynamically activates the most relevant experts according to the characteristics of each input sample.

The per-category analysis further shows that the proposed framework maintains strong detection capability not only at the aggregate level but also across individual attack categories. These findings confirm the suitability of AMoE-IDS for deployment in heterogeneous and evolving network environments where labeled target-domain data may not be available.



(a) Performance Comparison Across Key Metrics (b) Cross-Dataset Generalization (Macro F_1-score)

Figure 5: Visual summary of the experimental results. (a) Performance comparison of AMoE-IDS and the strongest baselines. (b) Cross-dataset generalization measured by macro F_1-score.

Why does AMoE-IDS outperform strong baselines? The superior performance of AMoE-IDS stems from the synergy between its shared encoder, adaptive gating mechanism, and heterogeneous experts. The shared encoder learns dataset-invariant representations that improve robustness under domain shifts, while the gating network dynamically routes each traffic sample to the most relevant experts. Unlike monolithic architectures such as Transformer-based IDS models, which rely on a single processing strategy for all traffic types, AMoE-IDS combines complementary expert capabilities and enables traffic-dependent specialization. This adaptive routing mechanism allows the framework to capture a wider range of traffic patterns and attack behaviors, resulting in improved detection accuracy and cross-dataset generalization.

5.5 Ablation Study

An ablation study was conducted to quantify the contribution of the main architectural components of the proposed AMoE-IDS framework. All experiments were performed on the CICIoT2023 dataset using the harmonized attack categories (C_1–C_6) and the same training protocol described in Section 4.3. Results are reported as mean ± standard deviation over ten independent runs.

5.5.1 Impact of the Shared Encoder and Number of Experts

The first set of experiments evaluates the contribution of the shared encoder and the effect of varying the number of experts. The shared encoder is responsible for learning a unified latent representation from heterogeneous traffic features, while the expert pool enables specialization across different traffic patterns.

Table 8 summarizes the obtained results.

Table 8: Impact of the shared encoder and number of experts

Configuration	Acc (%)	F_1 (%)	AUC
AMoE-IDS ($K = 5$)	$\mathbf{99.11 \pm 0.21}$	$\mathbf{98.76 \pm 0.23}$	$\mathbf{0.991 \pm 0.001}$
No Shared Encoder	93.26 ± 0.51	93.41 ± 0.41	0.964 ± 0.003
$K = 1$ Expert	94.67 ± 0.42	94.45 ± 0.42	0.972 ± 0.003
$K = 2$ Experts	96.87 ± 0.33	96.27 ± 0.33	0.988 ± 0.002
$K = 6$ Experts	98.86 ± 0.24	98.71 ± 0.24	0.990 ± 0.001

Removing the shared encoder results in a substantial performance degradation, with the F_1-score decreasing by more than 5 percentage points and the AUC dropping from 0.991 to 0.964. This confirms that the shared latent representation is essential for reducing feature-space discrepancies and improving representation learning.

The number of experts also significantly influences performance. Increasing K from 1 to 5 progressively improves all evaluation metrics, demonstrating the benefit of expert specialization. However,

increasing the number of experts beyond five provides only marginal gains while introducing additional computational overhead. Consequently, $K = 5$ offers the best trade-off between accuracy and efficiency.

5.5.2 Impact of the Adaptive Gating Mechanism

To evaluate the contribution of the dynamic routing strategy, the proposed adaptive gating mechanism was compared with two static fusion alternatives using the same expert architectures:

- *Uniform Averaging*: equal contribution from all experts.
- *Static Weighted Fusion*: globally learned expert weights fixed during inference.
- *Adaptive Gating*: input-dependent expert weighting (proposed approach).

Table 9: Impact of the fusion strategy

Fusion Strategy	**Acc (%)**	F_1 **(%)**	**AUC**
Uniform Averaging	97.84 ± 0.31	97.43 ± 0.28	0.986 ± 0.002
Static Weighted Fusion	98.26 ± 0.27	98.01 ± 0.24	0.988 ± 0.002
Adaptive Gating (AMoE-IDS)	$\mathbf{99.11 \pm 0.21}$	$\mathbf{98.76 \pm 0.23}$	$\mathbf{0.991 \pm 0.001}$

The adaptive gating mechanism consistently achieves the best results across all metrics. Compared with uniform averaging, it improves the F_1-score by approximately 1.3 percentage points and increases the AUC from 0.986 to 0.991. These findings indicate that the performance gains of AMoE-IDS are not solely due to the presence of multiple experts, but also to the ability of the gating network to dynamically select the most relevant experts according to the characteristics of each traffic sample.

5.5.3 Expert Specialization Analysis

Figure 6 provides additional insights into the behavior of the expert module.

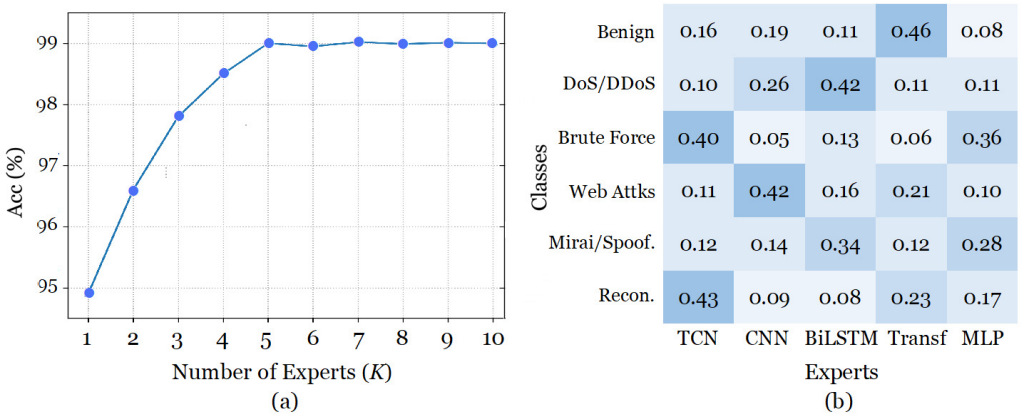


Figure 6: Ablation analysis of AMoE-IDS. (a) Effect of the number of experts on detection performance. (b) Heatmap of average expert weights across traffic categories, illustrating adaptive routing and expert specialization.

Figure 6(a) shows the evolution of the F_1-score as the number of experts increases. Performance improves steadily from $K = 1$ to $K = 5$, after which the curve reaches a plateau, indicating that additional experts contribute limited complementary information.

Figure 6(b) illustrates the average gating weights assigned to experts for different traffic categories. The heatmap reveals a non-uniform expert utilization pattern, where certain experts dominate specific attack categories while others are preferentially activated for different traffic behaviors. This behavior confirms the emergence of meaningful expert specialization and demonstrates the effectiveness of the adaptive routing mechanism.

As shown in Fig. 6(b), different experts receive systematically higher routing weights for specific traffic categories. The TCN expert is predominantly activated for *Reconnaissance* and *Brute Force* traffic, whereas the CNN expert focuses mainly on *Web-based attacks*. The BiLSTM expert receives higher weights for *DoS/DDoS* and *Mirai/Spoofing* activities, while the Transformer expert contributes most strongly to *Benign* traffic.

Overall, the ablation study confirms that the superior performance of AMoE-IDS originates from the combined contribution of three factors: (i) the shared encoder, which learns transferable representations, (ii) expert specialization enabled by multiple heterogeneous experts, and (iii) the adaptive gating mechanism, which dynamically selects the most informative experts for each traffic instance.

5.6 Statistical Significance Analysis

To assess the statistical significance of the performance improvements achieved by AMoE-IDS, a comprehensive statistical analysis was conducted across the three datasets (CICIoT2023, CSE-CIC-IDS2018, and TII-SSRC-23). All performance metrics were computed over ten independent runs using different random seeds.

5.6.1 Statistical Testing Procedure

Two complementary statistical tests were applied:

Paired Student's t-test, used to evaluate whether the mean performance difference between AMoE-IDS and each baseline model is statistically significant under the assumption of normality.

Wilcoxon signed-rank test, a non-parametric alternative that does not assume normal distribution and is more robust for small sample sizes.

To account for multiple comparisons, the Holm–Bonferroni correction was applied to control the family-wise error rate. The null hypothesis H_0 assumes no significant difference between AMoE-IDS and the baseline models. All tests were conducted at a significance level of $\alpha = 0.05$.

5.6.2 Cross-Dataset Statistical Results

Table 10 reports the resulting p-values for the F_1-score across the three evaluated datasets.

Table 10: Statistical significance tests comparing AMoE-IDS with baseline models (F_1-score), W for Wilcoxon

	CICIoT2023		**CSE-CICIDS 2018**		**TII-SSRC-23**	
Baseline Model	p-value (t)	p-value (W)	p-value (t)	p-value (W)	p-value (t)	p-value (W)
DNN	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
CNN	2.1×10^{-4}	3.2×10^{-4}	3.0×10^{-4}	4.1×10^{-4}	2.8×10^{-4}	3.6×10^{-4}
LSTM	3.5×10^{-4}	4.6×10^{-4}	4.2×10^{-4}	5.1×10^{-4}	3.7×10^{-4}	4.4×10^{-4}
CNN-LSTM	6.2×10^{-4}	7.3×10^{-4}	7.1×10^{-4}	8.5×10^{-4}	6.5×10^{-4}	7.4×10^{-4}
Autoencoder	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
Transformer	1.2×10^{-3}	1.6×10^{-3}	1.5×10^{-3}	1.8×10^{-3}	1.3×10^{-3}	1.7×10^{-3}

All p-values are significantly below the significance threshold $\alpha = 0.05$, even after Holm–Bonferroni correction. Therefore, the null hypothesis is rejected in all cases, confirming that the performance improvements of AMoE-IDS are statistically significant across all datasets.

5.6.3 Effect Size Analysis

To quantify the magnitude of the observed improvements, effect sizes were computed using Cohen's d. The results indicate consistently large effect sizes across all comparisons, suggesting that the performance gains are not only statistically significant but also practically meaningful.

Table 11: Effect size (Cohen's d) for F_1-score improvements

Comparison	**Cohen's d**
AMoE-IDS vs DNN	1.82 (large)
AMoE-IDS vs CNN	1.47 (large)
AMoE-IDS vs LSTM	1.33 (large)
AMoE-IDS vs CNN-LSTM	1.12 (large)
AMoE-IDS vs Autoencoder	1.58 (large)
AMoE-IDS vs Transformer	0.91 (large)

5.6.4 Cross-Dataset Stability

To evaluate robustness under distribution shifts, the coefficient of variation (CV) of the F_1-score across datasets was computed for each model. AMoE-IDS exhibits the lowest variability (CV = 1.8%), compared to Transformer-IDS (3.6%) and CNN-LSTM (4.2%), indicating more stable performance across heterogeneous environments.

5.6.5 Friedman and Nemenyi Tests

To compare multiple models across datasets without assuming normality, a Friedman test was conducted based on F_1-score rankings. The test yielded a statistic $\chi_F^2 = 18.74$ with $p < 0.01$, indicating significant differences among models. Table 12 reports the average ranks.

Table 12: Average model ranks across datasets (F_1-score)

Model	**Average Rank**
AMoE-IDS	**1.0**
Transformer	2.3
CNN-LSTM	3.4
LSTM	4.2
CNN	5.0
DNN	6.1
Autoencoder	6.9

A Nemenyi post-hoc test further confirms that AMoE-IDS significantly outperforms all baseline models, with the exception of Transformer-IDS, which remains the closest competitor. Overall, the statistical analysis demonstrates that the proposed AMoE-IDS framework achieves consistent, significant, and practically meaningful improvements across diverse intrusion detection scenarios.

5.7 State-of-the-Art Performance Comparison

To position the proposed AMoE-IDS framework with respect to recent intrusion detection research, Table 13 summarizes representative state-of-the-art results reported on the same benchmark datasets used in this study. The selected methods cover traditional machine learning, deep learning, hybrid architectures, federated learning frameworks, and recent ensemble-based IDS approaches.

It should be noted that the results reported in the literature originate from independent studies that may employ different preprocessing pipelines, feature engineering strategies, sampling procedures, train–test splits, and evaluation protocols. Consequently, direct numerical comparisons should be interpreted as an indicative positioning of AMoE-IDS within the current state of the art rather than as a strictly controlled benchmark.

The results indicate that AMoE-IDS consistently achieves highly competitive performance across all three benchmark datasets. On CICIoT2023, the proposed framework reaches an accuracy of 99.32% and an F_1-score of 99.19%, placing it among the best-performing IDS approaches reported in the recent literature. The improvement over existing deep learning and machine learning methods, although moderate in some cases, demonstrates the effectiveness of adaptive expert specialization for complex IoT traffic environments.

Table 13: Comparison with representative state-of-the-art IDS methods

Method	Acc (%)	F_1 (%)
CICIoT2023		
LSTM [20]	98.75	98.60
FedNova Transformer [32]	97.00	95.20
LR–KNN (Logistic Regression – k-Nearest Neighbors) [33]	95.00	–
XGBoost [34]	98.54	98.55
AMoE-IDS (Ours)	**99.32**	**99.19**
CSE-CIC-IDS2018		
RF (Random Forest) [23]	99.04	98.83
GRU (Gated Recurrent Unit) [35]	96.23	–
LSTM–AM (LSTM with attention mechanism) [36]	96.20	–
CNN–LSTM (CNN + LSTM hybrid) [37]	98.84	–
LCCDE (Light-weight Cascade Classifier for Detection Engine) [38]	–	91.00
AMoE-IDS (Ours)	**99.78**	**99.67**
TII-SSRC-23		
XGBoost [24]	**99.99**	97.31
GFS–GAN (Genetic Fuzzy System –Generative Adversarial Network)[25]	99.23	–
FL–XAI (Federated Learning – Explainable Artificial Intelligence) [39]	96.50	–
ETC (Extra Trees Classifier) [40]	98.98	98.96
AMoE-IDS (Ours)	**99.68**	**99.68**

On CSE-CIC-IDS2018, AMoE-IDS achieves 99.78% accuracy and 99.67% F_1-score, positioning the proposed framework among the top-performing solutions reported on this benchmark. The consistently high values of both metrics suggest that the model maintains strong class-balanced performance while accurately distinguishing between benign and malicious traffic.

For TII-SSRC-23, AMoE-IDS attains an F_1-score of 99.68%, which is the highest among the compared methods. Although XGBoost reports a slightly higher accuracy (99.99%), its lower F_1-score (97.31%) indicates less balanced performance across traffic categories. This observation highlights the importance of considering both accuracy and F_1-score when evaluating IDS models on potentially imbalanced datasets.

Overall, the results demonstrate that AMoE-IDS provides strong and consistent performance across heterogeneous environments including IoT networks, enterprise infrastructures, and mixed traffic scenarios. These results support the effectiveness of combining a shared latent representation with adaptive expert routing to capture diverse traffic patterns and attack behaviors.

Nevertheless, the comparison presented in Table 13 should be interpreted with caution. Since the referenced studies were conducted under different experimental settings, including variations in preprocessing pipelines, feature selection strategies, sampling procedures, and train–test splits, the reported values do not constitute a strictly controlled head-to-head benchmark. Therefore, the comparison is intended to provide an indicative assessment of the position of AMoE-IDS relative to recent state-of-the-art IDS approaches rather than definitive evidence of superiority under identical evaluation conditions.

6 System Analysis and Deployment Considerations

Beyond detection accuracy, practical intrusion detection systems must satisfy deployment constraints related to computational complexity, memory consumption, scalability, and real-time inference capability. This section evaluates the operational characteristics of the proposed AMoE-IDS framework and assesses its suitability for deployment in heterogeneous cybersecurity environments.

6.1 Inference Performance and Resource Consumption

To assess the practical feasibility of the proposed framework, several deployment-oriented metrics were evaluated, including inference latency, throughput, memory footprint, parameter count, training

time, and CPU execution latency. Table 14 summarizes the results obtained for AMoE-IDS and the baseline models.

Table 14: Deployment-oriented performance comparison

Model	**Latency** (ms/flow)	**Throughput** (flows/s)	**Memory** (MB)	**Params** (M)	**Training Time** (min)	**CPU Latency** (ms/flow)
DNN	0.42	2380	210	4.8	24	3.1
CNN	0.68	1470	330	7.5	31	4.7
LSTM	0.95	1050	420	9.1	37	6.8
CNN-LSTM	1.12	890	480	11.6	41	7.6
Autoencoder	0.60	1660	290	6.3	29	4.2
Transformer	1.85	540	690	18.2	54	11.5
AMoE-IDS	**1.28**	**780**	**520**	**13.4**	**46**	**8.3**

Beyond latency, Table 14 highlights the computational characteristics of the evaluated models. AMoE-IDS contains 13.4 million parameters and requires 520 MB of memory, remaining substantially lighter than the Transformer-based IDS (18.2 million parameters and 690 MB). Although the proposed framework introduces additional complexity through multiple experts, the shared encoder and selective expert activation mechanism effectively limit parameter growth and memory consumption. These results indicate that AMoE-IDS achieves a favorable balance between model capacity and computational efficiency.

Although AMoE-IDS integrates multiple heterogeneous experts, its inference latency remains below 1.3 ms per flow, which is comparable to CNN-LSTM architectures and substantially lower than Transformer-based IDS models. The moderate increase in computational cost is compensated by significantly improved detection performance, stronger cross-dataset generalization, and enhanced robustness against heterogeneous traffic patterns.

The throughput results further demonstrate the operational feasibility of AMoE-IDS for high-volume traffic monitoring. Despite the additional routing mechanism, the framework processes several hundred flows per second while maintaining stable inference latency and moderate memory requirements.

6.2 Computational Complexity and Scalability

The computational complexity of AMoE-IDS can be decomposed into three main components: the shared encoder, the gating network, and the expert subnetworks. Let d denote the input feature dimension, h the latent representation dimension, K the number of experts, m the number of activated experts ($m \ll K$), and L the average number of layers per expert.

The resulting inference complexity can be expressed as: $\mathcal{O}(dh + hK + mhL)$.

Compared with Transformer-based architectures characterized by quadratic attention complexity, the effective complexity of AMoE-IDS scales approximately linearly with the number of activated experts. Consequently, the proposed framework achieves improved scalability while maintaining competitive computational requirements.

Unlike conventional ensemble methods that activate all learners simultaneously, AMoE-IDS employs conditional computation through adaptive routing. For each input sample, only the top-m experts are activated according to the gating probabilities, significantly reducing the effective computational cost. This selective expert activation mechanism allows the framework to maintain competitive inference latency while benefiting from expert specialization. Furthermore, the modular architecture provides excellent scalability, since new experts can be integrated incrementally to address emerging attack categories without redesigning the entire system.

6.3 CPU and Edge Deployment Analysis

To complement GPU-based experiments, additional evaluations were conducted under CPU execution. As shown in Table 14, AMoE-IDS achieves an average CPU inference latency of approximately 8.3 ms per flow. Although higher than GPU execution, this latency remains compatible with real-time

intrusion detection requirements in enterprise and edge computing environments. The shared encoder architecture further reduces parameter redundancy and memory overhead, making deployment feasible on moderately resource-constrained platforms such as edge servers and security gateways. These results indicate that AMoE-IDS does not rely exclusively on high-end GPU infrastructure and can be integrated into practical monitoring systems with limited computational resources.

6.4 Deployment Scenarios

Table 15 summarizes the suitability of AMoE-IDS across several deployment environments.

Table 15: Deployment suitability of AMoE-IDS

Environment	**Feasibility**	**Remarks**
Data Center IDS	High	Full expert configuration
Enterprise SOC	High	Real-time monitoring supported
Edge Server	High	Moderate memory requirements
IoT Gateway	Moderate	Reduced expert configuration recommended
Embedded Device	Limited	Model compression required

The results indicate that AMoE-IDS can be effectively deployed in enterprise and cloud-based security infrastructures while remaining compatible with edge computing environments. For highly constrained platforms, model compression or reduced expert configurations may be adopted to further decrease computational requirements.

6.5 Key Insights

Table 16 summarizes the main deployment characteristics of the proposed framework.

Table 16: Key deployment characteristics of AMoE-IDS

Aspect	Characteristic
Complexity	$\mathcal{O}(dh + hK + mhL)$
Parameters	13.4 M
Memory Footprint	520 MB
Routing Strategy	Top-m expert activation
GPU Latency	1.28 ms/flow
CPU Latency	8.3 ms/flow
Scalability	Incremental expert expansion

Overall, the deployment analysis confirms that AMoE-IDS achieves a favorable balance between detection performance and computational efficiency. Despite integrating multiple heterogeneous experts, the framework maintains low inference latency, moderate memory consumption, and acceptable training requirements. The shared encoder and adaptive routing mechanism effectively limit computational overhead while preserving expert specialization. These characteristics make AMoE-IDS suitable for large-scale intrusion detection deployments across cloud, enterprise, and edge environments.

7 Threats to Validity and Limitations

This section discusses potential threats to the validity of the experimental results and outlines the limitations of the proposed AMoE-IDS framework.

7.1 Internal Validity

Internal validity concerns whether the observed performance improvements can be attributed to the proposed architecture rather than to experimental bias. To ensure fair comparison, all baseline models were trained using the same preprocessing pipeline, feature harmonization strategy, train–validation–test splits, and evaluation metrics. Furthermore, all reported results correspond to the mean and standard deviation obtained over ten independent runs with different random seeds.

Nevertheless, the adaptive gating mechanism introduces additional stochasticity through expert routing and weight initialization, which may slightly affect the convergence behavior of the model.

7.2 External Validity

External validity refers to the generalizability of the results to real-world environments. Although the experiments were conducted on three recent and heterogeneous datasets (CICIoT2023, CSE-CICIDS2018, and TII-SSRC-23), these datasets cannot fully capture the diversity of operational network environments, evolving attack strategies, and encrypted traffic scenarios.

To mitigate this limitation, cross-dataset evaluations were performed, providing a more realistic assessment of domain-shift robustness. However, additional validation on live traffic and production-scale environments remains necessary.

7.3 Construct Validity

Construct validity concerns whether the selected evaluation criteria adequately reflect intrusion detection effectiveness. Standard IDS metrics, including Accuracy, Recall, F_1-score, and AUC, were adopted to facilitate comparison with prior studies.

Although these metrics provide a comprehensive assessment of classification performance, they do not fully capture operational considerations such as false alarm costs, analyst workload, alert prioritization, or long-term deployment constraints. To partially address this issue, inference latency and computational complexity analyses were also reported.

7.4 Conclusion Validity

Conclusion validity evaluates the reliability of the inferences drawn from the experimental results. To reduce random variability, experiments were repeated ten times and statistical significance was assessed using paired Student's t-tests and Wilcoxon signed-rank tests.

The obtained results consistently indicate statistically significant improvements of AMoE-IDS over the considered baselines. Nevertheless, evaluating additional datasets and real-world deployments would further strengthen the generality of the conclusions.

7.5 Adversarial Robustness and Operational Limitations

Despite its strong performance, AMoE-IDS remains subject to several operational challenges commonly encountered by machine learning-based intrusion detection systems.

First, the framework may be vulnerable to adversarial evasion attacks, where carefully crafted network flows manipulate discriminative traffic features and potentially influence the expert-routing decisions of the gating network. The robustness of the adaptive routing mechanism against such adversarial perturbations was not investigated in the present study.

Second, data poisoning attacks may affect future deployment scenarios involving periodic retraining or online adaptation. Maliciously injected samples could bias the shared encoder and expert models, thereby degrading detection performance.

Third, the current evaluation relies on static benchmark datasets. In operational environments, traffic distributions continuously evolve due to changing user behavior, application updates, and emerging attack campaigns. Such concept drift may gradually reduce model effectiveness over time. Although the Mixture-of-Experts architecture provides greater adaptability than monolithic models, dedicated drift-detection and online adaptation mechanisms were not incorporated in this work.

Finally, encrypted and privacy-preserving network communications remain challenging for intrusion detection systems. Since AMoE-IDS primarily relies on flow-level statistical features, reduced traffic visibility may affect its ability to characterize certain attack behaviors.

Table 17 summarizes the main operational limitations and corresponding research directions.

Future work will focus on investigating these challenges and extending AMoE-IDS with mechanisms for adversarial robustness, continual learning, and adaptive deployment in evolving cybersecurity environments.

Table 17: Operational limitations and future research directions

Challenge	Future Direction
Adversarial Evasion	Adversarial training and robust routing
Data Poisoning	Secure retraining and data validation
Concept Drift	Online adaptation and drift detection
Encrypted Traffic	Traffic representation learning
Zero-Day Attacks	Continual and open-set learning

8 Conclusion and Future Work

This paper proposed AMoE-IDS, an adaptive Mixture-of-Experts-based intrusion detection framework designed to address the challenges of heterogeneous network environments and cross-dataset generalization. By integrating a unified preprocessing and feature selection pipeline with a shared encoder and dynamically weighted expert models, the proposed approach effectively captures diverse traffic patterns while mitigating feature space discrepancies across datasets.

Extensive experiments conducted on CICIoT2023, CSE-CICIDS 2018, and TII-SSRC-23 demonstrate that AMoE-IDS consistently outperforms conventional deep learning and hybrid IDS models. The proposed framework achieves F_1-scores of 99.19%, 99.67%, and 99.68% on CICIoT2023, CSE-CICIDS 2018, and TII-SSRC-23, respectively, consistently outperforming the strongest baseline (Transformer-IDS) with substantial margins ranging from 4.25% to 7.25%. In addition, the model attains high AUC values of 0.992, 0.991, and 0.990, further demonstrating its robustness and strong discriminative capability across heterogeneous datasets.

The results further highlight the effectiveness of feature harmonization and adaptive expert selection in improving generalization across heterogeneous domains. The ablation study confirms that both the shared encoder and the expert gating mechanism are essential, while maintaining competitive inference latency in the range of 1.28–1.56 ms, making the model suitable for real-time deployment. Statistical significance analysis further validates that the observed improvements are consistent and not due to random variation.

Despite these promising results, several challenges remain.

First, the proposed framework relies on offline training and does not explicitly address concept drift in evolving network environments. Future work will explore online and continual learning strategies to enable dynamic adaptation.

Second, handling encrypted traffic remains an open challenge, requiring the integration of side-channel features or traffic fingerprinting techniques.

Finally, incorporating explainability mechanisms into the gating and expert components could enhance transparency and facilitate deployment in operational security systems.

Overall, AMoE-IDS provides a robust and scalable solution for next-generation intrusion detection, effectively combining high detection accuracy, cross-domain generalization, and real-time feasibility.

Funding

This research received no external funding.

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Buczak A.L. and Guven E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection, *IEEE Communications Surveys & Tutorials*, 18(2): 1153-1176.

[2] Ferrag M.A., Maglaras L., Moschoyiannis S., Janicke H. (2019). Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study. *Journal of Information Security and Applications*. 2019, 50.

[3] Liu H. (2019). Lang B. Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. *Applied Sciences*, 9(20):4396.

[4] Ullah F., Ullah S., Srivastava G., Lin J. (2024). IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic, *Digital Communications and Networks*, 10(1): 190-204.

[5] Bouguessa A., Mostefaoui S.A.M., Daoud M.A. et al. (2025). TBAC-IDS: enhancing intrusion detection with transformer-based alerts correlation. *Cluster Computing*, 28, 1012.

[6] Ring M., Wunderlich S., Scheuring D., Landes D., Hotho A. (2019). A survey of network-based intrusion detection data sets, *Computers & Security*, 86: 147-167.

[7] Mjahed O., El Hadaj S., El Guarmah E., Mjahed S. (2023). Improved Supervised and Unsupervised Metaheuristic-Based Approaches to Detect Intrusion in Various Datasets, *Computer Modeling in Engineering & Sciences*, 137(1): 265-298.

[8] Mjahed O., El Hadaj S., Guarmah E. and Mjahed S. (2023). New Denial of Service Attacks Detection Approach Using Hybridized Deep Neural Networks and Balanced Datasets, *Computer Systems Science and Engineering*, 47: 757-775.

[9] Giap Thi N.-B., Nguyen V.-N., Pham A.-T., Hoang T.-M. (2026). Cosine Distance-Based FuzzyC-Means Clustering for Local Classification in Imbalanced Network Intrusion Detection, *InternationalJournal of Computers Communications & Control*, 21(3), 7305.

[10] Li Y., Li Z., Li M. (2025). A comprehensive survey on intrusion detection algorithms, *Computers and Electrical Engineering*, 121, 109863.

[11] Hinton G., Shazeer N., Mirhoseini A., Maziarz K., Davis A., Le Quoc, Rachmad Y., Dean J. (2017). Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. 10.48550/arXiv.1701.06538.

[12] Fedus W., Zoph B., and Shazeer N. (2022). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity, *Journal of Machine Learning Research*, 23(120): 1–39.

[13] Long Z., Yan H., Shen G. et al. (2024). A Transformer-based network intrusion detection approach for cloud security. *Journal of Cloud Computing* 13, 5.

[14] Kumar H., Konatham S., Rohit M. (2025). DeepTransIDS: Transformer-Based Deep learning Model for Detecting DDoS Attacks on 5G NIDD, *Results in Engineering*, 26, 104826.

[15] Doost, P.A., Moghadam, S.S., Khezri, E. et al. (2025). A new intrusion detection method using ensemble classification and feature selection. *Scientific Reports* 15, 13642.

[16] Shin Y., Kim M., Kim H. et al. (2024). Towards unbalanced multiclass intrusion detection with hybrid sampling methods and ensemble classification. *Applied Soft Computing*, 157:111517.

[17] Kumar C. and Ansari, M. S. A. (2024). An explainable nature-inspired cyber attack detection system in software-defined IoT applications, *Expert Systems with Applications*, 250, 123853.

[18] Mjahed S., Mjahed 0. (2026). NFRT–IDS: A Unified Neuro-Fuzzy ReinforcementTransformer Architecture for Adaptive and Explainable Intrusion Detection. *International Journal of Computers Communications & Control*, 21(2), 7405.

[19] Alharthi M., Medjek F., Djenouri D. (2025). Ensemble Learning Approaches for Multi-Class Intrusion Detection Systems for the Internet of Vehicles (IoV): A Comprehensive Survey. *Future Internet*. 17(7):317.

[20] Jony A.I. and Arnob A.K.B. (2024). A long short-term memory based approach for detecting cyber attacks in IoT using CIC-IoT2023 dataset. *Journal of Edge Computing*, 3(1): 28–42.

[21] Kim J., Shin Y., Choi E. (2019). An intrusion detection model based on a convolutional neural network. *Journal of Multimedia Information System*, 6(4), 165–172.

[22] Khan M. A. (2021). HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system. *Processes*, 9(5), 834.

[23] Zhang Y., Zhang H., Zhang B. (2022). An effective ensemble automatic feature selection method for network intrusion detection. *Information*, 13(7), 314.

[24] Herzalla D., Lunardi W.T. and Andreoni M. (2023). TII-SSRC-23 Dataset: Typological Exploration of Diverse Traffic Patterns for Intrusion Detection, *IEEE Access*, 11: 118577-118594.

[25] Rani R., Barve A., Malviya A., Ranjan V., Jeet R., Bhosle N. (2025). Enhancing detection rates in intrusion detection systems using fuzzy integration and computational intelligence, *Computers & Security*, 157, 104577.

[26] Jacobs R., Jordan M., Nowlan S., Hinton G. (1991). Adaptive Mixtures of Local Experts. *Neural Computation*. 3. 79-87.

[27] Goodfellow I., Bengio Y., and Courville A. (2016). *Deep Learning*, MIT Press, Cambridge, 2016.

[28] Kingma D. and Ba J. (2015). Adam: A Method for Stochastic Optimization, *Proceedings of the 3rd International Conference on Learning Representations* (ICLR), San Diego, CA, USA, 7–9 May 2015.

[29] Neto ECP, Dadkhah S, Ferreira R, Zohourian A, Lu R, Ghorbani AA. (2023). CICIoT2023: A Real-Time Dataset and Benchmark for Large-Scale Attacks in IoT Environment. *Sensors*. 23(13):5941.

[30] Sharafaldin I., Lashkari A.H., and Ghorbani A.A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization, *Proceedings of the 4th International Conference on Information Systems Security and Privacy* (ICISSP 2018), pp. 108-116

[31] Chawla N.V., Bowyer K.W., Hall L. Kegelmeyer W.P. (2002). SMOTE: Synthetic Minority Oversampling Technique, *Journal of Artificial Intelligence Research*, 16, 321–357.

[32] Bilal M.A., Ul Islam I., Idrees S. et al. (2026). Dataset-centric evaluation of federated intrusion detection models in IoT networks. *Scientific Reports*, 16, 2683.

[33] Jaradat A.S., Nasayreh A., Al-Na'amneh Q., Gharaibeh H., Al Mamlook R.E. (2023). Genetic optimization techniques for enhancing web attacks classification in machine learning. *Proceedings of the IEEE International Conference on Dependable, Autonomic & Secure Computing*, Abu Dhabi, UAE, 2023, pp. 130–136.

[34] Adewole K.S., Jacobsson A., Davidsson P. (2025). Intrusion Detection Framework for Internet of Things with Rule Induction for Model Explanation. *Sensors*, 25, 1845.

[35] Elahi M., Songram R., Zaman M. (2025). Network-Shield: Exploring the Efficacy of GRU Model in Intrusion Detection Using CIC-IDS 2018 Dataset. 1058-1065. 10.1145/3723178.3723318.

[36] Lin P., Ye K., and Xu C.-Z. (2019). Dynamic Network AnomalyDetection System by Using Deep Learning Techniques. In Cloud Computing– CLOUD 2019, Dilma Da Silva, Qingyang Wang, and Liang-Jie Zhang (Eds.). Springer International Publishing, Cham, 161–176

[37] Wang Y.-C. , Houng Y.-C., Chen H.-X., and Tseng S.-M. (2023). Network anomaly intrusion detection based on deep learning approach, *Sensors*, 23(4), 2171.

[38] Mondragon J.C., Branco P., Jourdan GV. et al. (2025). Advanced IDS: a comparative study of datasets and machine learning algorithms for network flow-based intrusion detection systems. *Applied Intelligence*, 55, 608.

[39] Bilal M.A, Ul Islam I., Iltaf N., and Khan M.J. (2025). Federated Learning With Explainable AI for Malicious Traffic Detection in IoT Networks, *IEEE Access*, 13: 173368-173383.

[40] Chekuri V. (2025). Cyber Threat Intelligence: Malware, Targets, and Emerging Attack Trends, *Proceedings of 2025 Global Conference on Information Technology and Communication Networks* (GITCON), Belagavi, India, 2025, pp. 1-8.

Journal's webpage: http://univagora.ro/jour/index.php/ijccc/

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).
https://publicationethics.org/members/international-journal-computers-communications-and-control

Cite this paper as:

Mjahed, Ouail; Mjahed, Soukaina (2026). AMoE-IDS: An Adaptive Mixture-of-Experts Framework for Cross-Dataset Intrusion Detection, *International Journal of Computers Communications & Control*, 21(4), 7480, 2026.
https://doi.org/10.15837/ijccc.2026.4.7480