network

MDPI

Article

Adaptive Decision-Level Intrusion Detection for Known and Zero-Day Attacks

Joseph P. Mchina *, Neema Mduma and Ramadhani S. Sinde

School of Computational and Communication Science and Engineering, Nelson Mandela African Institution of Science and Technology, Arusha 447, Tanzania; neema.mduma@nm-aist.ac.tz (N.M.); ramadhani.sinde@nm-aist.ac.tz (R.S.S.)

* Correspondence: mchinaj@nm-aist.ac.tz

Abstract

Network Intrusion Detection Systems (NIDS) face increasing challenges from sophisticated cyber threats, particularly zero-day attacks that evade signature-based methods. While supervised learning is effective for known attack classification, it struggles with novel threats, whereas anomaly-based approaches suffer from high false positive rates and unstable thresholds. To address these limitations, this paper proposes a decision-level adaptive intrusion-detection framework combining hierarchical CNN-based closed-set classification with autoencoder-based zero-day detection in a cascade architecture. The framework enables deployment-time adaptation by dynamically adjusting class-specific confidence thresholds and fusion parameters without model retraining. Experiments on the CSE-CIC-IDS2018 dataset demonstrate strong closed-set performance, achieving 98.98% accuracy and a macro-F1-score of 0.9342, with improved recall for minority attack classes under adaptive thresholding. Under a zero-day evaluation protocol in which Web_Attacks and Infiltration are excluded from training and validation, the proposed approach achieves an F1-score of 0.9319 while maintaining a low false positive rate of 0.0019. The framework is further evaluated on the Simulated University Network Environment (SUNE) dataset representing campus network traffic, achieving 96.18% closed-set accuracy and 97.54% accuracy in the integrated cascade setting. These results demonstrate that the proposed framework effectively balances minority attack detection, zero-day identification, and false-alarm control in dynamic and resource-constrained network environments.

Keywords: network intrusion detection system; deep learning; zero-day attack detection; open-set intrusion detection; adaptive thresholding; CSE-CIC-IDS2018

Academic Editor: Stavros Shiaeles

Received: 18 February 2026
Revised: 26 March 2026
Accepted: 30 March 2026
Published: 9 April 2026

1. Introduction

In recent years, the increasing reliance on digital infrastructures has made Network Intrusion Detection Systems (NIDS) a critical component of modern cybersecurity architectures. NIDS continuously monitor network traffic to detect and prevent malicious activities that threaten the confidentiality, integrity, and availability of information systems [1,2]. However, traditional signature-based NIDS are inherently limited to recognizing previously known attack patterns and therefore fail to detect novel or zero-day intrusions, which represent previously unseen or evolving threats [3–5]. This limitation has become increasingly critical as the volume and sophistication of cyberattacks continue to escalate globally, with recent studies estimating that cybercrime would cost the world economy more than USD 10.5 trillion annually by 2025 [6,7].

These challenges are particularly acute in resource-constrained higher learning institutions, where rapid digital transformation through e-learning platforms, research portals, and online administrative systems has accelerated significantly, while investments in cybersecurity infrastructure often remain limited [8,9]. In developing countries, this imbalance between digital adoption and cybersecurity readiness is further amplified by budgetary and infrastructural constraints [10,11]. These challenges are particularly evident in Tanzanian higher learning institutions, which face increased exposure to diverse cyber threats, including web-based attacks, Denial-of-Service (DoS) attacks, and Botnet intrusions. The heterogeneity of institutional networks, combined with limited computational resources, necessitates lightweight, intelligent, and scalable intrusion detection frameworks capable of operating efficiently under constrained deployment conditions.

To address these challenges, researchers have increasingly adopted Machine Learning (ML) and Deep Learning (DL) techniques, which demonstrate strong potential for enhancing intrusion detection capabilities beyond traditional rule-based systems [12,13]. Despite these advances, ML-based NIDS still face two fundamental challenges that restrict their real-world effectiveness. First, severe class imbalance in network traffic datasets often leads to high overall accuracy while masking poor detection performance on minority attack classes, resulting in missed detection of rare but high-impact threats [14,15]. Second, zero-day attack detection remains difficult, as unsupervised anomaly detection methods frequently exhibit high false positive rates due to their limited ability to distinguish malicious anomalies from legitimate but uncommon network behaviors [16,17].

Existing intrusion detection approaches attempt to address these challenges using static decision thresholds or periodic model retraining. However, static thresholds cannot simultaneously optimize minority-class sensitivity and false-alarm control under changing traffic conditions, while retraining introduces computational overhead and deployment instability [18,19]. Consequently, there remains a need for lightweight intrusion detection architectures that enable deployment-time adaptation without modifying learned model parameters.

To address this need, this study proposes a decision-level adaptive intrusion detection framework that introduces a unified deployment-time calibration paradigm for coordinating supervised and unsupervised detection components. The framework confines adaptation to the decision layer by dynamically updating class-specific confidence thresholds and fusion parameters while keeping all learned model weights fixed after training. This design separates representation learning from operational adaptation, enabling the system to improve minority attack sensitivity and zero-day detection robustness under changing traffic conditions without retraining overhead. The proposed approach integrates hierarchical CNN-based closed-set classification with autoencoder-based anomaly detection within a unified adaptive decision layer, designed for practical deployment environments where computational efficiency, stability, and false-alarm control are critical.

The main contributions of this work are summarized as follows:

- We propose a decision-level adaptive intrusion detection paradigm in which deployment-time adaptation is performed through dynamic calibration of class-specific confidence thresholds and fusion parameters while all learned neural network parameters remain fixed.
- We introduce a class-specific adaptive thresholding mechanism that improves detection sensitivity for minority attack classes under severe class imbalance through deployment-time calibration at the decision layer.
- We develop a confidence-aware adaptive fusion mechanism that integrates supervised CNN outputs with multiple anomaly signals to support coordinated detection of both known and previously unseen attacks within a unified framework, while demonstrat-

ing a parameter-efficient deployment strategy in which only decision thresholds and fusion parameters are updated during inference, avoiding costly retraining.

- We additionally evaluate the proposed framework on a simulated university network environment (SUNE) by re-applying the full training and evaluation pipeline under SUNE's native traffic conditions.

The remainder of this paper is organized as follows. Section 2 reviews related work on intrusion detection and adaptive cybersecurity systems. Section 3 describes the proposed adaptive intrusion detection framework and experimental methodology. Section 4 presents the experimental results and analysis. Section 5 discusses the findings and their implications. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

The evolution of intrusion detection systems from traditional signature-based approaches to machine learning and deep learning methods has been extensively studied in recent years [20]. Existing machine learning approaches, including Support Vector Machines (SVM), k-Nearest Neighbors (KNN), and Decision Trees, established foundational baselines for automated threat detection but demonstrated limited robustness against evolving and previously unseen attack patterns [21,22]. Deep learning architectures have shown substantial improvements in intrusion detection, with specific CNN-based implementations achieving over 99% accuracy on benchmark datasets [23]. Attention-enhanced recurrent models have demonstrated strong performance across multiple datasets [24]. Further, Liu and Liu [24] proposed HAGRU, which combines bidirectional GRU with hierarchical attention mechanisms to improve temporal dependency capture and minority attack detection, achieving F-scores of 96.71% on CIC-IDS2017 and 93.95% on CSE-CIC-IDS2018. However, these approaches primarily focus on architectural improvements and rely on fixed decision thresholds determined during offline validation.

Despite architectural advances, supervised approaches continue to face challenges in minority attack detection under severe class imbalance. Song et al. [25] reported that autoencoder-based methods showed poor detection of minority attacks with R2L TPR as low as 0.38 and U2R TPR around 0.63. These findings indicate that even sophisticated deep learning models may struggle to maintain balanced sensitivity across highly skewed class distributions.

Zero-day attack detection has emerged as an equally challenging problem in cybersecurity, with researchers primarily focusing on unsupervised anomaly detection approaches due to the absence of labeled data for previously unseen threats [26]. Autoencoder-based methods have dominated this research area due to their ability to learn normal behavior patterns without requiring labeled attack data. While autoencoder-based methods achieve accuracy ranging from 75 to 98% on benchmark datasets [27,28], they frequently suffer from high false positive rates. This limitation arises because anomaly detection models often struggle to distinguish malicious anomalies from legitimate but uncommon network behaviors. Moreover, most existing methods depend on statically defined reconstruction thresholds that require offline calibration and manual tuning, reducing robustness under dynamic operational conditions. Hybrid approaches by Dai et al. [29] integrated autoencoder-based anomaly detection with supervised classifiers to improve the detection of unseen attacks. However, these hybrid systems still rely on fixed calibration strategies determined prior to deployment.

These observations reveal a fundamental tension in intrusion detection system design. Optimizing for minority-class detection requires sensitive and finely tuned decision boundaries to capture rare attack behaviors, whereas robust zero-day detection demands broader anomaly boundaries capable of identifying previously unseen patterns. Integrating

supervised classification and anomaly-based detection within a unified framework remains a significant challenge in intrusion detection research [29]. While several studies address minority detection or zero-day detection independently, comparatively fewer works explicitly coordinate both through adaptive decision calibration mechanisms. The architectural requirements for each objective differ: minority attack detection benefits from discriminative supervised models capable of capturing subtle attack signatures, whereas zero-day detection requires anomaly detection models trained exclusively on Normal traffic. The central challenge therefore lies not in developing either capability independently, but in coordinating both through adaptive decision mechanisms that balance sensitivity and false-alarm control without incurring the computational cost of full model retraining.

Recent studies have emphasized the importance of decision threshold optimization in intrusion detection systems. Kim et al. [30] applied Gaussian KDE-based analysis to determine unified thresholds for autoencoder-based detection, achieving a 99.2% F1-score across multiple attack types in vehicular networks. Similarly, Lan et al. [31] employed ROC-based per-class threshold optimization in industrial control systems, significantly improving rare attack detection rates. Although effective, these approaches rely on offline threshold determination and assume stable traffic distributions, limiting their responsiveness to evolving operational conditions.

Adaptive approaches in cybersecurity have also explored ensemble selection and model-weighting strategies to improve robustness [32,33]. Zoppi et al. [32] investigated meta-learning-based ensembles for unsupervised intrusion detection, demonstrating reduced misclassification across multiple datasets. Alalhareth and Hong [33] proposed a performance-driven ensemble IDS that dynamically weights classifiers based on accuracy and confidence. However, these methods primarily operate at the model level, requiring multiple classifiers or dynamic model weighting schemes that increase computational overhead. In contrast, decision-level calibration where class-specific thresholds and fusion parameters are adjusted without modifying network weights has received comparatively limited attention in intrusion detection research.

The growing deployment of AI-based systems in security-critical applications has introduced sophisticated new adversarial threats beyond traditional network intrusions. Zhou et al. [34] surveyed backdoor threats in large language models demonstrating that adversaries can embed hidden malicious behaviors into AI models during training through data poisoning and parameter manipulation, causing models to produce adversary-controlled outputs when specific triggers are present while behaving normally otherwise. Although focused on large language models, this threat model is relevant to deep-learning-based intrusion detection systems more broadly, as CNN and autoencoder models trained on network traffic data face analogous risks of training-phase compromise. The proposed framework addresses this concern architecturally by fixing all neural network parameters after training and confining deployment-time adaptation exclusively to lightweight decision-layer parameters, namely class-specific thresholds and fusion weights, ensuring that learned model weights remain immutable during deployment.

These observations motivate the development of intrusion detection frameworks that perform adaptation at the decision layer rather than at the model level. The proposed framework addresses this need by introducing deployment-time calibration through dynamic adjustment of class-specific thresholds and fusion parameters while learned feature representations remain unchanged. This design enables coordinated handling of minority known attacks and previously unseen zero-day attacks within a unified architecture while preserving computational efficiency and operational stability. The adaptive coordination mechanism at the decision layer therefore represents the central methodological contribution of this work. Unlike prior hybrid IDS approaches such as Dai et al. [29] that

combine supervised and unsupervised components but rely on fixed offline calibration, and ensemble methods such as Zoppi et al. [32] and Alalhareth and Hong [33] that adapt at the model selection level requiring multiple classifiers, the proposed framework uniquely performs adaptation exclusively at the decision layer without modifying any learned model parameters during deployment.

3. Materials and Methods

This section presents the proposed decision-level adaptive intrusion detection framework, which enables parameter-efficient adaptation through class-specific threshold adjustment and adaptive fusion of supervised and unsupervised detection signals. Rather than retraining model weights, the framework performs deployment-time calibration by updating confidence thresholds and fusion weights using recent prediction statistics. All adaptation is confined to the decision layer, while feature representations and model parameters remain fixed after training, ensuring computational efficiency and stability in resource-constrained environments. The adaptive update rules follow a feedback-driven calibration principle in which operational performance metrics, specifically per-class recall and false positive rate, serve as error signals guiding dynamic adjustment of decision thresholds and fusion weights, providing a principled alternative to static, offline calibration without modifying any learned model parameters.

3.1. Proposed Adaptive Framework

Framework Architecture Overview

Figure 1 illustrates the architecture of the proposed AU-IDS framework, comprising two coordinated detection paths, the Gate CNN and Attack CNN for known attack classification and the autoencoder for zero-day detection, integrated through an adaptive decision controller that dynamically updates class-specific thresholds and fusion weights to produce final intrusion decisions via cascade logic. The system processes network traffic features through two coordinated but independently trained detection paths. The first path performs closed-set, supervised intrusion detection, targeting known attack classes including minority attacks [35]. The second path performs open-set anomaly detection for identifying previously unseen (zero-day) attacks [36]. A lightweight adaptive decision controller continuously monitors recent prediction behavior and dynamically updates class-specific decision thresholds and fusion weights to balance detection sensitivity and false positive control. All adaptation occurs at the decision level, without modifying or retraining the underlying neural network parameters.

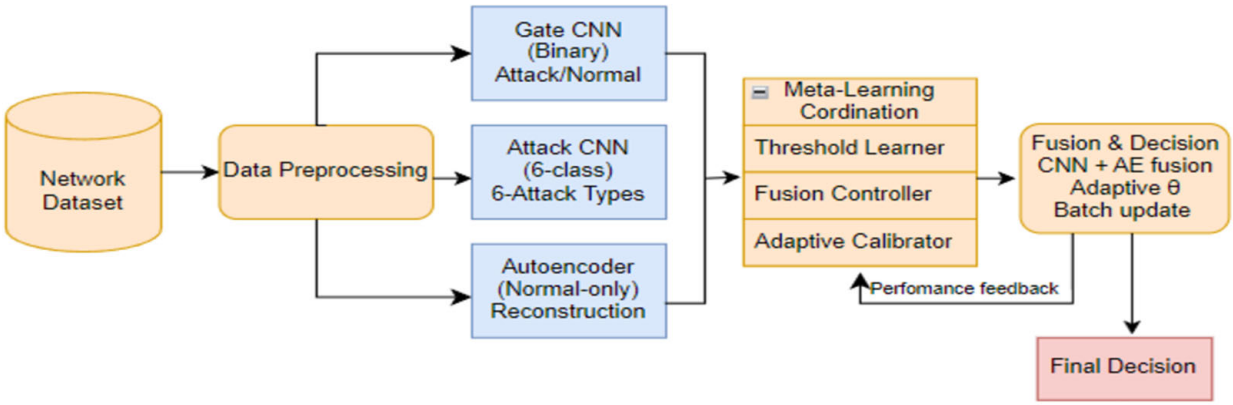

Figure 1. Overall system architecture.

The framework follows a dual-stage detection paradigm [37]. For known attacks, a hierarchical CNN-based classifier is used, consisting of a binary Gate CNN and a multi-class

Attack CNN. For zero-day detection, an autoencoder trained exclusively on Normal traffic identifies anomalous patterns. An adaptive controller integrates outputs from both paths through fusion weights and threshold adaptation, enabling robust detection of minority and novel attacks while maintaining operational stability.

3.2. Dataset and Data Preparation

This section describes the two datasets used for evaluation: CSE-CIC-IDS2018 and SUNE, together with the preprocessing, feature selection, class balancing, and normalization steps applied to each.

3.2.1. CSE-CIC-IDS2018 Dataset

This study employs the CSE-CIC-IDS2018 dataset, developed by the Canadian Institute for Cybersecurity through systematic intrusion traffic characterization over a 10-day capture period [38]. The dataset contains 16.1 million network flow records with 80 bidirectional flow-based features extracted across HTTP, HTTPS, SSH, SMTP, and POP3 protocols using CICFlowMeter v3 [39]. The underlying infrastructure comprised 50 attack machines, 420 victim PCs, and 30 servers, replicating enterprise-scale conditions. The dataset encompasses seven attack categories representing contemporary intrusion techniques: BruteForce, DoS, DDoS, Web_Attacks, Infiltration, Botnets, and port scans [39].

3.2.2. Data Preprocessing and Feature Preparation

Initial preprocessing addresses data quality issues present in the CSE-CIC-IDS2018 dataset. Missing values representing less than 0.1% of the dataset are removed, while infinite values are replaced with boundary values to maintain numerical stability. Non-predictive attributes such as flow IDs and timestamps are excluded to focus on analytically relevant features, reducing the feature space from 80 to 74 dimensions. Attack labels are Normalized and reorganized into seven principal classes: Normal, BruteForce, DoS, DDoS, Web_Attacks, Infiltration, and Botnets, enabling consistent multi-class evaluation.

Following initial preprocessing, feature selection is performed using a class-aware mutual information–based strategy applied exclusively to the training data to prevent data leakage. Balanced one-vs-rest mutual information scores are computed for each class to ensure adequate representation of minority attack categories. Final feature selection combines quota-based per-class ranking with global multi-class mutual information refinement. This process reduces the feature set from 74 to 22 selected features, which are used for all subsequent model training and evaluation stages.

3.2.3. Class Balancing Strategy

The dataset was partitioned using stratified sampling into 70% training, 15% validation, and 15% testing subsets to prevent data leakage. Class balancing is applied only to the training partition using SMOTE-Tomek [39,40]. This hybrid approach generates synthetic minority samples while cleaning the resulting dataset by removing overlapping instances [41]. As shown in Table 1, the balancing strategy successfully reduces the Normal class by 87.9% (from 1,169,625 to 141,653) while increasing the Web_Attacks class by 30.8 times from 650 to 20,000 samples, lowering the overall imbalance ratio from 1799.4:1 to a manageable 7.1:1.

Table 1. Dataset class distribution.

Class	Before Balancing	After SMOTE-Tomek
Normal	1,169,625	141,653
BruteForce	33,274	33,272

Table 1. *Cont.*

Class	Before Balancing	After SMOTE-Tomek
DoS	57,151	57,150
Web_Attacks	650	20,000
Infiltration	31,500	23,169
Botnet	24,998	25,000
DDoS	110,403	80,000

3.2.4. Data Partitioning and Normalization

After feature preparation and class balancing, the dataset is partitioned using stratified sampling into 70% training, 15% validation, and 15% testing sets to preserve class proportions across all subsets. Validation and test sets remain unbalanced to ensure realistic performance evaluation under operational conditions. Feature normalization is performed using StandardScaler with parameters derived exclusively from training data to maintain validation independence and prevent data leakage.

3.2.5. Simulated University Network Environment (SUNE)

To complement evaluation on the CSE-CIC-IDS2018 benchmark, an additional dataset was generated using a Simulated University Network Environment designed to reflect traffic conditions typical of higher learning institutions in Tanzania. Direct collection of live campus traffic was not feasible due to privacy and ethical constraints. Therefore, network activity was produced within a controlled physical network testbed incorporating Cisco Catalyst 2960 and 3560 switches, a Cisco Catalyst 2900 router, a D-Link DGS-1210-52MPT switch, and a Sophos XGS 2100 firewall interconnected across six VLANs. Server infrastructure ran on Windows Server 2022 hosting institutional services. A dedicated Kali Linux machine executed controlled attack scenarios. Attack scenarios included DoS attacks using hping3, Probe attacks using Nmap, and User-to-Root and Remote-to-Local attacks using Metasploit and Hydra. Normal traffic representing common academic network usage was generated alongside controlled attack scenarios, including User-to-Root, Remote-to-Local, Denial-of-Service, and Probe attacks. Traffic was collected continuously over five days to capture temporal variation in network activity. The simulated environment includes typical academic services such as web portals, email servers, file transfer services, DNS resolution, and SSH access generating realistic background traffic. Attack scenarios were executed under controlled conditions using dedicated attack machines with precise timing records maintained to ensure reliable ground-truth labeling. Flow labels were assigned based on attack execution timestamps to accurately distinguish between Normal and malicious traffic. The SUNE dataset is used as an independent evaluation dataset. The full training and evaluation pipeline applied to CSE-CIC-IDS2018 is re-executed on SUNE using its native class taxonomy, with models trained and evaluated separately for each dataset. The SUNE dataset contains 626,981 network flow records comprising 494,664 normal flows and 132,317 malicious flows distributed across four attack categories: DoS (60,567), U2R (27,250), Probe (23,250), and R2L (21,250), representing a normal-to-attack ratio of approximately 3.74:1. Traffic was generated across HTTP, HTTPS, SMTP, POP3, DNS, SSH, and FTP protocols capturing heterogeneous, application-level communication patterns representative of university network usage. Network flows were described using 45 flow-based features. The same preprocessing and feature selection pipeline described in Sections 3.2.2 and 3.2.3 was applied independently to SUNE, resulting in 18 selected features compared to 22 for CSE-CIC-IDS2018, reflecting domain-specific traffic characteristics rather than inconsistencies in methodology.

3.3. Base Detection Models

This section describes the two independently trained base detection components: the hierarchical CNN classifier for closed-set detection and the autoencoder for zero-day detection.

3.3.1. Closed-Set Hierarchical CNN Models

The closed-set intrusion detection component is implemented using a hierarchical convolutional neural network architecture composed of two independently trained CNN models, each serving a distinct role within the detection pipeline. The first model, referred to as the *Gate CNN*, performs binary classification to distinguish between Normal and Attack traffic. This model acts as an initial filtering stage, determining whether an observed network flow should be further analyzed for attack categorization. The second model, referred to as the *Attack CNN*, is trained exclusively on attack samples and performs multi-class classification across six known attack categories: BruteForce, DoS, DDoS, Web_Attacks, Infiltration, and Botnet. By separating binary detection from attack-type classification, the framework is designed to mitigate class confusion and improve sensitivity to minority attack classes.

Both CNN models employ the same architectural design to ensure consistent feature extraction. Each network processes normalized flow features through two one-dimensional convolutional layers with 32 and 64 filters, respectively, followed by max pooling and adaptive max pooling layers to capture both local and global temporal patterns in network traffic. The extracted features are flattened and projected into a 128-dimensional embedding space via a fully connected layer with ReLU activation. Finally, task-specific output layers are applied: The Gate CNN uses a single sigmoid output neuron optimized using binary cross-entropy loss with logits, while the Attack CNN employs a softmax output layer trained using categorical cross-entropy loss. Importantly, no class-weighted loss functions are used during training; instead, sensitivity to minority attack classes is achieved through adaptive confidence thresholding applied at inference time, ensuring stable training dynamics and parameter-efficient adaptation. The detailed layer-wise configuration and parameter specifications of the CNN architecture are summarized in Table 2.

Table 2. CNN architecture specifications.

Layer	Type	Parameters	Output Shape
Input	-	-	(batch, 22)
Reshape	-	-	(batch, 22, 1)
Conv1D	32 filters, kernel = 5, padding = 2	192	(batch, 22, 32)
MaxPool1D	Pool_size = 2	-	(batch, 11, 32)
Conv1D	64 filters, kernel = 3, padding = 1	6208	(batch, 11,64)
AdaptiveMaxPool1D	Output_size = 8	-	(batch, 8, 64)
Flatten	-	-	(batch, 512)
Dense	128, ReLU	65,664	(batch, 128)
Gate Head	1 unit, Sigmoid	129	(batch, 1)
Attack Head	6 units, Softmax	774	(batch, 6)

3.3.2. Autoencoder for Zero-Day Detection

Zero-day intrusion detection is performed using a fully connected autoencoder trained in an unsupervised manner on Normal traffic samples only [42]. The autoencoder consists of an encoder network that compresses input feature vectors into a 32-dimensional latent representation, followed by a symmetric decoder that reconstructs the original input features. Training is conducted by minimizing the mean squared reconstruction error

(MSE) between the input and reconstructed output, encouraging the model to learn a compact representation of legitimate network behavior [42].

During inference, reconstruction error serves as an anomaly score, with higher errors indicating a greater deviation from learned normal patterns and thus a higher likelihood of malicious activity. This design strictly enforces a zero-day-safe protocol, as samples from designated zero-day attack classes are explicitly excluded from both training and validation phases. As a result, the autoencoder remains unbiased toward unseen attack behaviors and provides a reliable anomaly signal for detecting previously unknown threats.

3.4. Adaptive Decision Mechanisms

This section describes the three adaptive decision mechanisms: class-specific threshold adaptation, adaptive fusion for zero-day detection, and integrated cascade decision logic, which together constitute the deployment-time calibration capability of the proposed framework

3.4.1. Class-Specific Threshold Adaptation

To improve detection performance for minority attack classes while preserving overall system stability, the framework assigns a class-specific confidence threshold, θ_c, to each known attack class, $c \in C$. These thresholds define the minimum classification confidence required for accepting an attack prediction. Initial threshold values are set conservatively, with lower thresholds assigned to minority classes (Web_Attacks = 0.30 and Infiltration = 0.25) to enhance early sensitivity. Thresholds are adapted online using a sliding window of recent predictions. This window-based update mechanism enables dynamic recalibration without modifying model weights. Adaptation follows the update rule:

$$\theta_c \leftarrow \mathbf{clip}(\theta_c - \eta_\theta(T_c - R_c),\ \mathbf{0.05},\ \mathbf{0.95}) \quad (1)$$

where R_c **denotes** a recall proxy estimated from recent predictions, T_c denotes a predefined target recall for class c, and η_θ is a small adaptation rate controlling update magnitude. In all reported experiments, the adaptation rate is set to $\eta_\theta = 0.01$; this value was selected to provide gradual stable threshold updates as larger values produced unstable oscillations during preliminary experiments while smaller values resulted in adaptation too slow to respond to genuine traffic distribution changes. The target recall is set to $T_c = 0.90$ for all attack classes reflecting a security-oriented operating point that prioritizes detection sensitivity. Sensitivity analysis across T_c values of 0.80, 0.90, and 0.95 confirmed stable performance with F1 varying only between 0.9296 and 0.9319. The calibration window size is set to 100 recent predictions; smaller windows produced noisy recall estimates while larger windows slowed adaptation to evolving traffic patterns. Fusion weights are initialized on the validation set under a fixed FPR budget. The update rule implements a proportional feedback mechanism in which the term (T_c and R_c) acts as a performance error signal driving threshold adjustment toward the desired operating point. When observed recall falls below the target, the threshold is reduced to increase detection sensitivity. When recall exceeds the target, the threshold is raised to control false positives. The adaptation rate η_θ governs the magnitude of each correction and the clip operation enforces hard bounds of [0.05, 0.95] ensuring stable and bounded adaptation throughout deployment. This formulation provides a direct and interpretable mechanism for controlling the precision–recall trade-off at the decision layer rather than relying on static or empirically chosen thresholds. In supervised evaluation mode true labels are used to compute the recall proxy, R_c, providing an offline upper-bound analysis of adaptation effectiveness; this mode is used exclusively for evaluation purposes to establish the maximum achievable adaptation gain. In unsupervised deployment mode which represents the intended real-world operation

adaptation relies solely on label-free proxy statistics derived from prediction confidence distributions ensuring practical applicability without requiring access to ground-truth labels during inference.

3.4.2. Adaptive Fusion for Zero-Day Detection

Zero-day detection decisions are derived by adaptively fusing multiple complementary signals extracted from the open-set detection pipeline. Specifically, the framework integrates four complementary signals each capturing a distinct perspective of anomalous behavior: the Gate CNN attack probability capturing supervised discriminative evidence of attack patterns; the autoencoder reconstruction error capturing unsupervised deviation from Normal traffic behavior; the GMM negative log-likelihood measuring embedding-space anomalies for samples falling outside known traffic distributions; and the energy-based score capturing classification uncertainty for samples not conforming to known class distributions. Together these signals combine supervised confidence, reconstruction-based anomaly detection, density estimation, and uncertainty measurement improving robustness against zero-day attacks that cannot be reliably detected by any single signal alone. Each signal is robustly normalized using percentile-based scaling to mitigate the influence of extreme values and ensure numerical stability. The fused anomaly score is computed as:

$$s_{fused} = w_{gate}\mathbf{p} + w_{ae}\mathbf{e} + w_{gmm}\mathbf{g} + w_{energy}\mathbf{s} \quad (2)$$

where p, e, g, and s denote the normalized component signals, and the corresponding weights are constrained to be non-negative and sum to one. The fusion weight update follows a performance-driven weighting strategy in which recent false positive rate statistics act as feedback signals indicating the operational reliability of each component detector. Signals associated with higher false-alarm rates are progressively down-weighted while more stable signals are reinforced. The non-negativity and sum-to-one constraints ensure a valid convex combination of anomaly signals at all times providing numerical stability while enabling dynamic rebalancing of supervised and unsupervised detection signals under changing traffic conditions. Fusion weights are initially calibrated on the validation set under a fixed false positive rate (FPR) budget, ensuring operational feasibility. During testing, these weights are further adapted online using recent FPR estimates in supervised mode, or pseudo-FPR proxies derived from high-confidence normal predictions in unsupervised mode. This adaptive fusion strategy enables the system to balance anomaly sensitivity and false-alarm control under dynamic conditions. The relative contribution of each signal is empirically validated through an ablation study presented in Section 4.3.

3.4.3. Integrated Cascade Decision Logic

Final intrusion decisions are produced using a cascade-based decision strategy that integrates closed-set and open-set detection outcomes. For each incoming sample, the closed-set classifier is evaluated first. If a known attack class is predicted with confidence exceeding the class-specific threshold, the corresponding attack label is immediately accepted. If the sample is classified as Normal, it is subsequently evaluated by the zero-day detector. Samples flagged as anomalous by the open-set detection module are labeled as NOVEL_ATTACK, while all remaining samples are classified as Normal. This cascade design preserves the high precision and interpretability of supervised multi-class detection for known attacks, while enabling reliable identification of previously unseen threats. By limiting zero-day evaluation to samples not confidently classified as known attacks, the framework avoids unnecessary false positives and achieves a balanced, deployment-oriented intrusion detection strategy.

3.5. Training and Evaluation Protocol

This section describes the four-stage training procedure, the performance metrics used for evaluation, and the experimental setup used in all reported experiments.

3.5.1. Training Procedure

The training procedure follows a four-stage approach designed to establish stable feature representations prior to deployment-time decision adaptation. Stage 1 involves training the Gate CNN to distinguish Normal from Attack traffic using binary cross-entropy loss for ten epochs with the Adam optimizer, a learning rate of $1 \times \mathbf{10^{-3}}$, and a batch size of 128. Stage 2 trains the Attack CNN exclusively on attack samples to perform multi-class classification using categorical cross-entropy loss under the same optimization settings. Stage 3 trains the autoencoder using only Normal traffic samples for thirty epochs with the Adam optimizer, a learning rate of $3 \times \mathbf{10^{-4}}$, a batch size of 256, and mean squared reconstruction error as the objective function. Stage 4 performs post-training calibration on the validation set, including temperature scaling and tuning of fusion parameters used in the decision-level detection stage. No class-weighted loss functions, sampling adjustments, or explicit minority-class boosting strategies are applied during model training. Instead, class imbalance is addressed exclusively through adaptive decision thresholds and fusion calibration applied during inference, ensuring that all adaptation is confined to the decision layer without modifying learned model parameters.

3.5.2. Performance Metrics

Performance evaluation employs comprehensive metrics covering both multi-class classification and zero-day detection capabilities [43]. The evaluation metrics considered in this study are organized as follows:

A. Multi-Class Classification Evaluation

(a) Accuracy: Accuracy measures the ratio of all examples correctly classified across all attack categories:

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

The metric provides an overall indication of classification correctness and serves as a baseline measure for comparing system effectiveness across different attack types.

(b) Precision: Precision quantifies the correctness of positive predictions, calculated as:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

This metric indicates how many of the instances predicted as attacks are actually attacks, which is crucial for minimizing false alarms in operational environments.

(c) Recall: Recall measures the system's ability to identify all actual attack instances:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

This metric represents the sensitivity of the classification system and is particularly important for ensuring comprehensive threat detection, especially for minority attack classes.

(d) F1-score: The F1-score is calculated as the harmonic mean between precision and recall:

$$\text{F1} = (2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

https://doi.org/10.3390/network6020023

This unified metric balances precision and recall, making it suitable for multi-class intrusion detection with severe class imbalance. Both macro-averaged and per-class F1-scores are computed to provide a comprehensive performance assessment.

B. Zero-Day Detection Evaluation

(a) Area Under the ROC Curve (AUC-ROC)

AUC-ROC measures the discriminative ability of the system to distinguish between Normal and anomalous traffic across all possible threshold settings. This metric provides a threshold-independent assessment of the fundamental zero-day detection capability, with values closer to 1.0 indicating better discriminative performance.

(b) Detection Rate: Detection Rate quantifies the proportion of zero-day attacks correctly identified:

$$\text{Detection Rate} = TP/(TP + FN)$$

This metric represents the system's sensitivity to previously unseen threats and is critical for evaluating the effectiveness of the autoencoder-based anomaly detection component.

(c) False Positive Rate (FPR): For zero-day detection, FPR measures the proportion of Normal traffic incorrectly flagged as anomalous:

$$FPR = FP/(FP + TN)$$

This metric is crucial for operational deployment, as high false positive rates in anomaly detection lead to alert fatigue and reduced system utility.

3.5.3. Experimental Setup

All experiments were conducted in the Google Colab environment using Python 3.12 with 12 GB RAM on a Linux x86_64 platform. Model development and evaluation were implemented using PyTorch for neural network training, Scikit-learn for preprocessing, evaluation metrics, and Gaussian mixture modeling, and imbalanced-learn for SMOTE-Tomek resampling. To ensure reproducibility, the random seed was fixed at 42 across all experimental stages. To validate the lightweight characterization of the proposed framework a runtime and computational cost analysis was conducted using the same experimental environment. The measured training times are 125.12 s for the Gate CNN, 62.64 s for the Attack CNN, and 48.03 s for the autoencoder, resulting in a total training time of 235.80 s across all components. For deployment-time operation, the complete cascade pipeline processes the entire test dataset in 4.12 s corresponding to an average inference latency of 0.0135 ms per network flow, confirming real-time capability under practical network traffic conditions. The adaptive decision controller introduces an overhead of only 124 microseconds per update since the threshold and fusion weight adjustments require only lightweight scalar operations on a small set of parameters without any gradient computation or model retraining. This difference of several orders of magnitude between full retraining time of 235.80 s and adaptation overhead of 124 microseconds directly supports the lightweight characterization of the proposed framework.

4. Results

This section presents experimental results across five configurations: closed-set detection, adaptive thresholding impact, zero-day detection, integrated cascade performance, and SUNE evaluation.

4.1. Closed-Set Multi-Class Detection Performance

This subsection evaluates the performance of the proposed hierarchical CNN-based intrusion detection system under a closed-set setting, where all attack classes observed during testing are known during training. Table 3 presents the per-class precision, recall, and F1-score across the seven traffic categories.

Table 3. Per-class closed-set classification performance.

Class	Precision	Recall	F1-Score
Normal	0.9933	1.0000	0.9966
BruteForce	0.8338	0.9414	0.8843
DoS	1.0000	0.7895	0.8824
Web_Attacks	0.6479	0.9928	0.7841
Infiltration	0.9996	0.9864	0.9929
Botnet	0.9996	0.9996	0.9996
DDoS	0.9999	0.9988	0.9994

The model achieves strong performance for the Normal class, with a recall of 1.000 and an F1-score of 0.9966. Botnet and DDoS attacks are detected with near-perfect performance, each achieving F1-scores above 0.999. For minority attack classes, Web_Attacks achieves a recall of 0.9928 and a precision of 0.6479, resulting in an F1-score of 0.7841. Infiltration achieves a precision of 0.9996 and a recall of 0.9864, yielding an F1-score of 0.9929. DoS achieves a recall of 0.7895 with a precision of 1.0000, indicating that no false positive DoS predictions were observed, while some DoS samples were misclassified into other attack categories. The normalized confusion matrix in Figure 2 shows minimal off-diagonal entries, indicating limited misclassification between attack classes, which is critical for maintaining operational security.

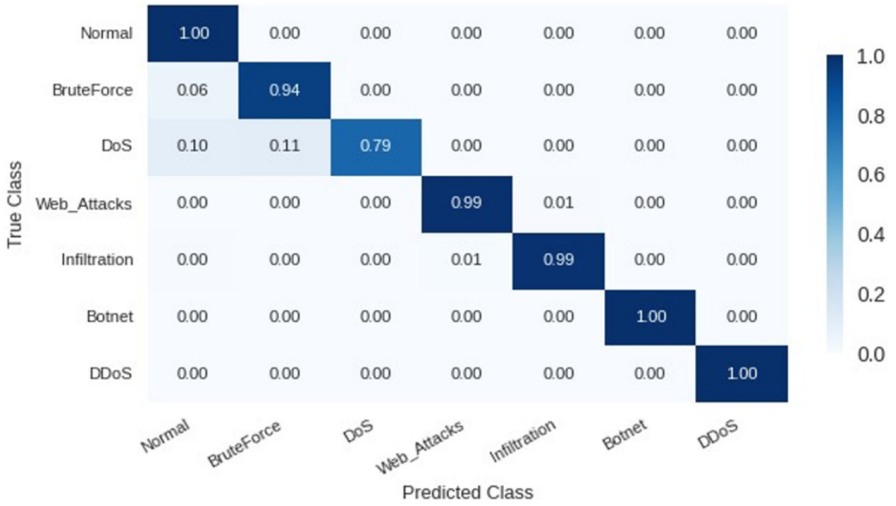

Figure 2. Normalized confusion matrix for closed-set multi-class intrusion detection.

Overall closed-set performance yields an accuracy of 98.98%, a macro-averaged F1-score of 0.9342, and a weighted F1-score of 0.9895. The disparity between macro and weighted metrics reflects differences in class frequency, while still indicating consistent performance across both majority and minority traffic classes. Figure 3 illustrates the validation F1-score progression of the Gate CNN and Attack CNN models across training epochs, providing additional insight into the closed-set training process.

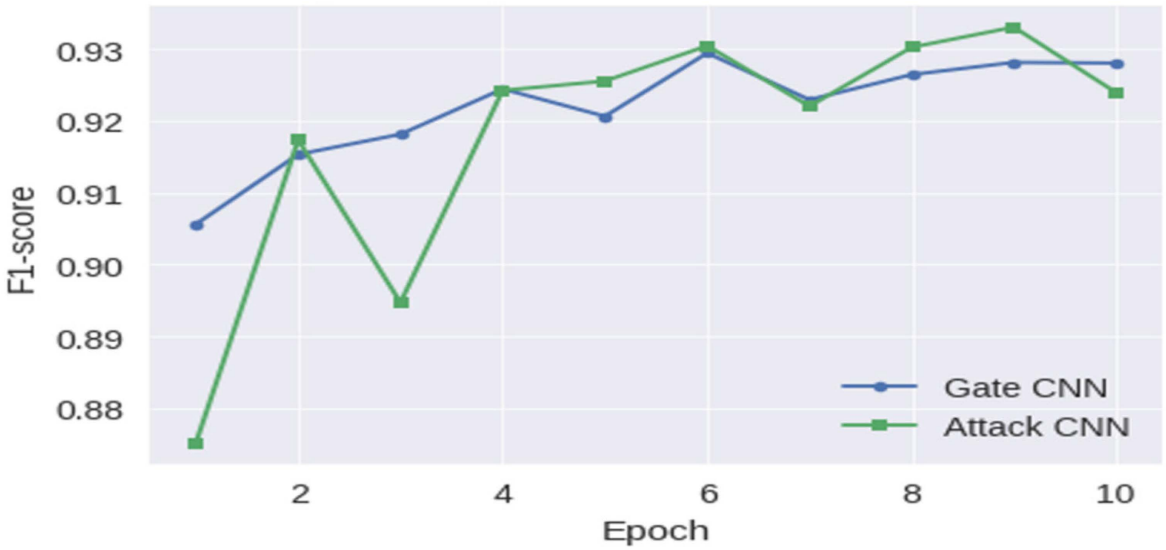

Figure 3. Training convergence of the Gate CNN and Attack CNN.

4.2. Impact of Adaptive Thresholding on Closed-Set Detection Performance

This subsection compares closed-set detection performance under static thresholding and adaptive thresholding, focusing on precision, recall, and overall accuracy. It is noted that the adaptive thresholding results reported in this subsection correspond to the supervised evaluation mode in which true labels are used to compute the recall proxy, representing an upper-bound on adaptation effectiveness. The unsupervised deployment mode, which operates without ground-truth labels, defines the actual deployment behavior of the framework. Using static thresholds, the model attains an overall accuracy of 98.33%, while the adaptive threshold configuration achieves an overall accuracy of 98.68%. Improvements in recall are observed for minority attack classes under adaptive thresholding. BruteForce recall increases from 0.6632 to 0.7983, Web_Attacks recall increases from 0.9137 to 0.9928, and Infiltration recall increases from 0.9858 to 0.9867. Changes in precision are also observed between the two configurations. For Web_Attacks, precision decreases from 0.8944 under static thresholding to 0.6479 under adaptive thresholding, while Infiltration precision remains consistently high at 0.9995 in both cases. Botnet and DDoS classes maintain near-perfect precision and recall across both configurations. Overall, adaptive thresholding improves minority-class recall at the cost of reduced precision for Web_Attacks, while maintaining comparable overall accuracy.

4.3. Zero-Day Open-Set Detection Results

This subsection reports the performance of the proposed system under a zero-day, open-set evaluation protocol, where detection is formulated as a binary classification task distinguishing Normal traffic from Attack traffic. Under this protocol, the Web_Attacks and Infiltration classes are excluded from both training and validation and are used only during testing, ensuring that zero-day attacks remain unseen during model learning. These classes were deliberately selected for the primary evaluation protocol because they represent the most challenging zero-day scenario. Web_Attacks contains only 650 original training samples, and both classes exhibit stealthy traffic patterns that overlap with Normal traffic, providing a conservative lower-bound assessment of framework capability. To support the zero-day detection pipeline, an autoencoder is trained exclusively on Normal traffic to model baseline reconstruction behavior. Figure 4 presents the reconstruction loss of the autoencoder across training epochs.

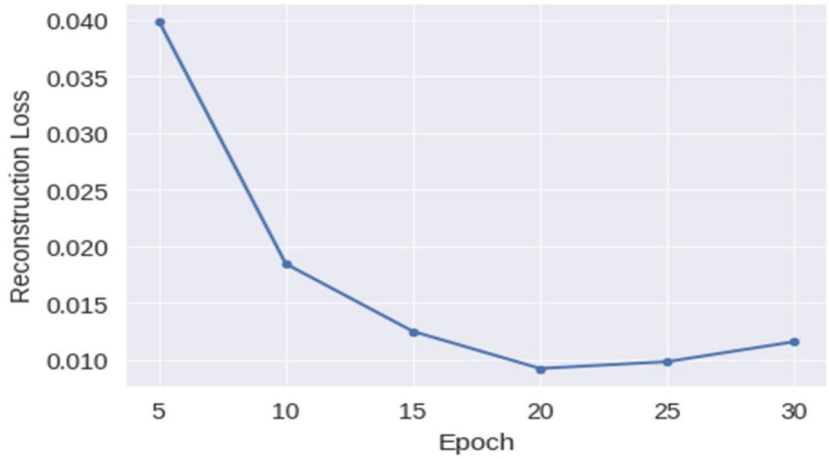

Figure 4. Autoencoder reconstruction loss on Normal traffic during training.

The zero-day detection module achieves a detection rate of 0.8747 at a false positive rate of 0.0019, with a corresponding precision of 0.9972 and an F1-score of 0.9319. The overall classification accuracy under this protocol is 0.9769. Figure 5 presents the receiver operating characteristic (ROC) curve for zero-day detection, with an area under the curve (AUC) of 0.9480, indicating strong separability between Normal and zero-day traffic across a wide range of decision thresholds. To assess sensitivity to the target recall parameter, Tc, the zero-day evaluation was repeated with Tc = 0.80, 0.90, and 0.95. At Tc = 0.80: Detection Rate 0.8895, FPR 0.0024, F1 0.9308. At Tc = 0.90 (our setting): Detection Rate 0.8747, FPR 0.0019, F1 0.9319. At Tc = 0.95: Detection Rate 0.8982, FPR 0.0029, F1 0.9296. The F1-score varies only between 0.9296 and 0.9319, confirming robustness to Tc variations while allowing operators to adjust the operating point without retraining. To assess the generalizability of the zero-day detection component across diverse attack categories, two additional exclusion protocols were evaluated on CSE-CIC-IDS2018. Under the second protocol, excluding Web_Attacks and BruteForce, the system achieves a detection rate of 0.9058, FPR of 0.0025, F1 of 0.8979, and AUC of 0.9519. Under the third protocol, excluding Botnet and DDoS, the system achieves a detection rate of 0.9701, FPR of 0.0019, F1 of 0.9807, and AUC of 0.9893. Across all three protocols, F1-scores range from 0.8979 to 0.9807 and AUC values remain above 0.9480, demonstrating consistent zero-day detection performance across diverse attack categories. Table 4 presents a comparison of the proposed framework against representative open-set and hybrid IDS methods from recent literature. Since different studies report heterogeneous metrics, detection rate serves as the primary comparable metric.

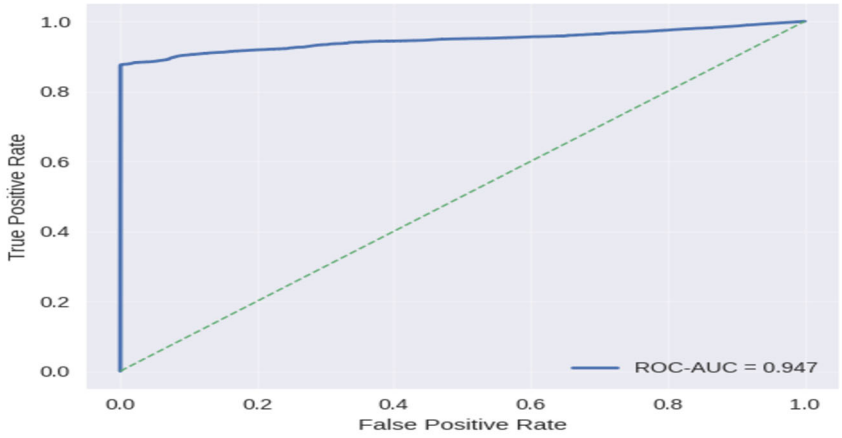

Figure 5. ROC curve for zero-day.

Table 4. Comparison with representative open-set intrusion detection methods.

Name	Dataset	Protocol Used	Detection Rate
Hindy et al. [27]	CICIDS2017	Normal-only training	0.2778–0.6688
Soltani et al. [5]	CSE-CIC-IDS2018	Leave-one-class-out	0.4006–0.4843
VAEMax [36]	CSE-CIC-IDS2018	Random exclusion	0.66
Farrukh et al. [44]	CICIDS2017	5-class exclusion	0.88
Al-Zewairi et al. [45]	CSE-CIC-IDS2018	Leave-one-class-out	0.88–0.95
Proposed	CSE-CIC-IDS2018	2-class exclusion	0.8747

Figure 6 shows the binary confusion matrix for zero-day detection, illustrating that most zero-day attack samples are correctly identified while the vast majority of Normal traffic remains correctly classified. To isolate the contribution of each anomaly signal in the fusion mechanism, an ablation study was conducted evaluating four progressively constructed fusion configurations under the same zero-day evaluation protocol. Table 5 summarizes the results.

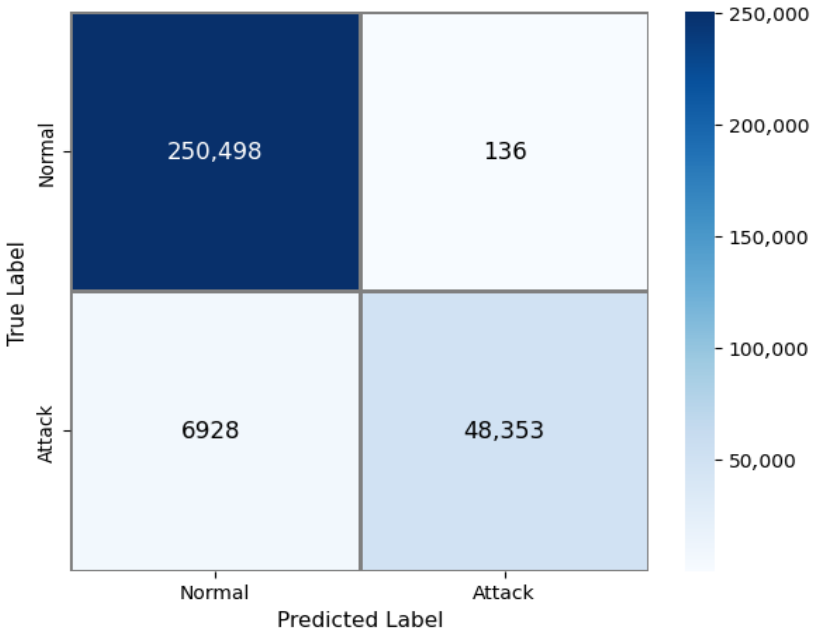

Figure 6. Confusion matrix for zero-day.

Table 5. Ablation study results.

Configuration	Detection Rate	FPR	F1
Gate CNN Probability only	0.81	0.004	0.88
Gate CNN + Autoencoder reconstruction error	0.85	0.003	0.90
Gate CNN + Autoencoder + GMM log-likelihood	0.87	0.0024	0.92
Full fusion (all four signals)	0.8747	0.0019	0.93

Each signal addition produces measurable improvement, confirming that all four signals contribute complementary information. The autoencoder reconstruction error provides the largest individual contribution, while the energy-based score achieves the final reduction in FPR, demonstrating that the complete fusion is necessary for optimal false-alarm control.

4.4. Integrated Cascade IDS Performance

This subsection reports the performance of the proposed system in its integrated cascade configuration, which represents deployment-oriented behavior where closed-set classification and zero-day detection operate jointly, and traffic is classified as either Normal or Attack. Under this configuration, the integrated IDS achieves an overall accuracy of 0.9837, with a precision of 0.9177, a recall of 0.9996, and an F1-score of 0.9569. The false positive rate is 0.0198. The near-perfect recall reflects the cascade's prioritization of comprehensive attack detection under operational conditions, where minimizing missed intrusions takes precedence over strict false-alarm suppression. Figure 7 presents the confusion matrix for the integrated cascade IDS, illustrating the resulting trade-off between detection sensitivity and false positive rate when both closed-set and open-set components are active simultaneously.

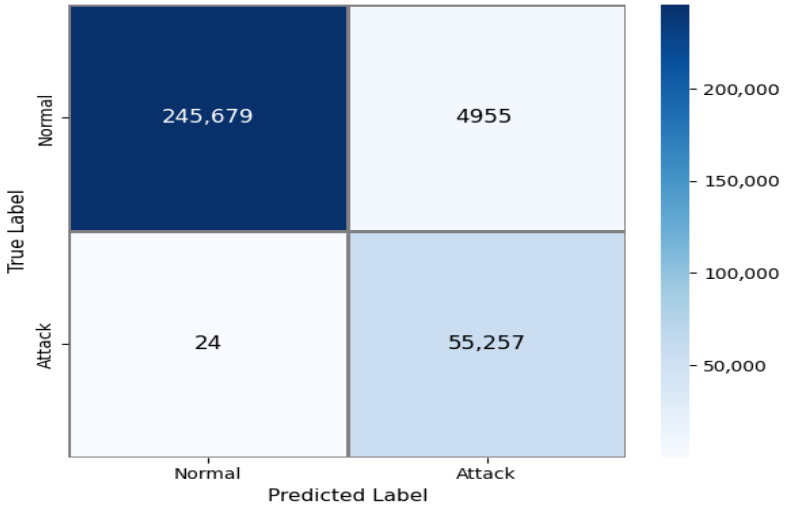

Figure 7. Confusion matrix for the integrated cascade IDS.

4.5. SUNE Performance Results

This section presents the evaluation results of the proposed framework on the SUNE dataset. The full training and evaluation pipeline is re-executed on SUNE using its native class taxonomy with models trained and evaluated separately from CSE-CIC-IDS2018. This experiment therefore evaluates architectural generalizability and reproducibility across independent datasets under dataset-specific retraining conditions rather than demonstrating direct cross-dataset transfer or domain-shift robustness.

4.5.1. Closed-Set Multi-Class Detection Performance (SUNE)

This subsection evaluates the performance of the proposed intrusion detection system under a closed-set multi-class classification. In this configuration, all traffic categories present during testing are known during training, and the model assigns each network flow to one of the predefined classes: Normal, U2R, R2L, DoS, or Probe. The model achieves an overall classification accuracy of 0.9618. The macro-F1-score is 0.8270, while the weighted F1-score is 0.9572, indicating strong overall classification performance under SUNE traffic conditions. Detection performance is highest for the Normal and DoS classes, both of which are classified with perfect precision and recall. The U2R class shows a very high recall of 0.9902 but a comparatively lower precision of 0.6209. The R2L class is detected with a balanced precision of 0.9035 and a recall of 0.8695. In contrast, the Probe class exhibits a lower recall of 0.3246, indicating that a substantial portion of Probe traffic is assigned to other attack categories. The distribution of classification outcomes across all classes is summarized in the normalized confusion matrix shown in Figure 8.

Figure 8. Closed-set multi-class confusion matrix for SUNE dataset.

4.5.2. Zero-Day Detection Performance

This subsection reports the performance of the proposed system under a zero-day, open-set evaluation protocol, where detection is formulated as a binary classification task distinguishing Normal traffic from Attack traffic. Under this protocol, the R2L and Probe classes are excluded from both training and validation and are used only during testing, ensuring that these attack categories remain unseen during model learning. On the test set, the zero-day detector achieved an overall classification accuracy of 0.9737, with a precision of 0.9838, a recall of 0.8964, and an F1-score of 0.9381. The model produced an ROC-AUC of 0.9955 and the false positive rate was 0.0042. The resulting classification outcomes are summarized in the normalized confusion matrix shown in Figure 9. To further assess zero-day generalizability on SUNE, two additional exclusion protocols were evaluated. Under the second protocol, excluding U2R and DoS, the system achieves accuracy of 0.9687, precision of 0.9990, recall of 0.8600, F1 of 0.9243, AUC of 0.9956, and FPR of 0.0025. Under the third protocol, excluding DoS and Probe, the system achieves accuracy of 0.9746, precision of 0.9817, recall of 0.9025, F1 of 0.9405, AUC of 0.9858, and FPR of 0.0048. Across all three SUNE protocols, F1-scores range from 0.9243 to 0.9405, demonstrating consistent zero-day detection performance across the available SUNE attack categories.

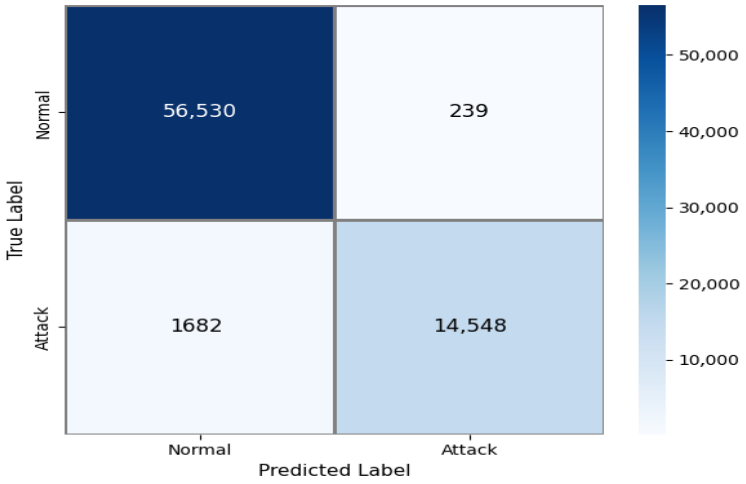

Figure 9. Confusion matrix for zero-day detection.

4.5.3. Integrated Cascade IDS Performance

This subsection reports the performance of the proposed system in its integrated cascade configuration, which represents deployment-oriented operation where closed-set,

multi-class classification and zero-day detection function jointly, and traffic is ultimately classified as either Normal or Attack. Under this configuration, the integrated IDS achieved an overall classification accuracy of 0.9754, with a precision of 0.9015, recall of 0.9986, and an F1-score of 0.9476. The false positive rate on Normal traffic was 0.0312. The resulting classification outcomes of the integrated cascade are summarized in the normalized confusion matrix shown in Figure 10.

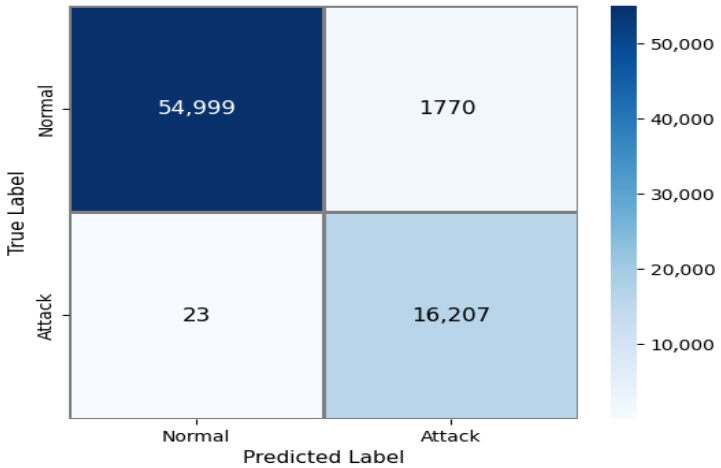

Figure 10. Confusion matrix for integrated cascade IDS.

5. Discussion

The closed-set multi-class evaluation indicates that the proposed hierarchical CNN with adaptive thresholds achieves performance that is competitive with recent intrusion detection approaches [43–45]. As summarized in Table 6, the proposed method attains high accuracy, precision, recall, and F1-score relative to prior studies.

Table 6. Comparison with related work.

Study	Accuracy	Precision	Recall	F1-Score
Alzugaib et al. [46]	98.41	99.55	98.85	99.20
Khan and Kim [47]	-	98.35	89.20	98.27
Harini et al. [48]	98.30	98.42	98.46	98.18
Hewapathirana et al. [49]	93.00	-	-	-
Our proposed	98.68	99.67	98.88	98.62

The results demonstrate that the proposed approach maintains balanced detection performance across both majority and minority attack classes, with differences in experimental protocols and preprocessing strategies noted where relevant. Notably, Hewapathirana [49] reports 93% overall accuracy on CSE-CIC-IDS2018 but highlights significant difficulty detecting the Infiltration class where many instances are misclassified as benign traffic. In contrast, the proposed framework achieves substantially higher overall accuracy while maintaining strong recall for minority attack classes including Infiltration, demonstrating the advantage of decision-level adaptive calibration. While improvements in recall are achieved for minority attack classes, a precision reduction is observed for Web_Attacks under adaptive thresholding from 0.8944 under static thresholding to 0.6479 under adaptive thresholding. This precision–recall trade-off warrants operational interpretation rather than purely metric-based assessment. The precision reduction for Web_Attacks arises because web-based attacks such as SQL injection and cross-site scripting exhibit flow-level characteristics similar to legitimate HTTP traffic resulting in greater feature overlap with Normal traffic. When the decision threshold is lowered to improve recall, borderline

samples are accepted increasing false positives and reducing precision. In contrast, Infiltration produces more distinctive traffic patterns at the flow level and therefore maintains high precision of 0.9995 even under adaptive thresholds. Web-based attacks represent high-severity threats capable of causing data breaches and system compromise. In this operational context, the cost of missing a Web_Attack substantially outweighs the cost of investigating an additional false alarm, justifying the recall-prioritizing behavior of adaptive thresholding. Furthermore, given the naturally low base rate of Web_Attacks in the test set (only 97 samples out of approximately 2.4 million total instances), the absolute increase in false alerts remains operationally manageable. The target recall parameter, T_c, is operator-configurable, allowing the precision–recall operating point to be adjusted to match specific alert budget requirements without retraining. Unlike methods that primarily rely on deeper architectures, extensive ensemble designs, or aggressive loss reweighting, the observed performance gains are achieved through lightweight threshold adaptation at the decision layer. This observation reinforces the central premise of this study that performance improvements arise primarily from the adaptive coordination mechanism at the decision layer, enabling minority-class sensitivity and zero-day robustness within a unified lightweight framework.

For zero-day intrusion detection, comparison across studies requires careful interpretation given differences in evaluation protocols, unknown-class definitions, and reported metrics. The following comparison uses detection rate as the primary comparable metric as it is most consistently reported across open-set and hybrid IDS studies. Existing anomaly-based and open-set IDS approaches, such as AutoIDS [50] and Open-CNN, primarily focus on detecting deviations from Normal traffic or recognizing unknown classes, often reporting detection rates without explicit control of false positives. More recent studies, including VAEMax [36] and Farrukh et al. [44], adopt stricter open-set protocols and emphasize recall–false positive trade-offs, reflecting a growing awareness of deployment constraints.

In contrast, the proposed system evaluates zero-day detection under an explicit attack-exclusion protocol and reports both recall and false positive rate at a defined operating point. The achieved detection rate of 0.8747 at an FPR of 0.0019 demonstrates that competitive sensitivity can be maintained while preserving operational false-alarm control. The variation in detection performance observed across exclusion protocols is scientifically expected and reflects the behavioral distance of each excluded attack class from Normal traffic. Volumetric attacks such as Botnet and DDoS produce very high reconstruction errors, and are detected with high recall of 0.9701, while attacks such as BruteForce and U2R produce moderate reconstruction errors due to partial behavioral overlap with Normal traffic resulting in slightly lower recall of 0.9058 and 0.8600, respectively. Rather than relying solely on anomaly scores with static thresholds, the framework integrates supervised classification and unsupervised anomaly detection through adaptive decision-level calibration. As shown in the zero-day comparison Table 4 in Section 4.3, Soltani et al. [5] report zero-day detection accuracies of only 40.06% for Web_Attacks and 48.43% for Infiltration on CSE-CIC-IDS2018 under a comparable protocol. The proposed framework improves these by more than 40 percentage points on the same dataset. VAEMax [36] achieves approximately 66% recall for unknown attacks on CSE-CIC-IDS2018; the proposed framework exceeds this by more than 21 percentage points. Al-Zewairi et al. [45] achieve a high recall of 0.96 but at the cost of low precision of 0.40 and an F1 of 0.45. In contrast, the proposed framework achieves an F1 of 0.9319 with an FPR of 0.0019. This integration enables coordinated control of sensitivity and robustness within a unified architecture. The ablation study results in Table 5 of Section 4.3 confirm that each fusion signal contributes distinct complementary information justifying the multi-signal fusion design of Equation (2).

The integrated cascade configuration further illustrates the deployment-oriented characteristics of the proposed system. By prioritizing supervised closed-set classification and invoking zero-day detection only for samples classified as Normal, the cascade achieves high overall recall of 0.9996 while maintaining an acceptable false positive rate of 0.0198. This behavior reflects the intended design trade-off: prioritizing comprehensive attack coverage while limiting excessive false alarms.

Evaluation on the independent SUNE dataset demonstrates that the proposed framework maintains consistent detection performance across independent datasets when retrained under domain-specific conditions, confirming architectural generalizability rather than cross-dataset transfer capability. Although overall performance remains strong, a modest reduction in macro-level performance and lower recall for Probe traffic indicate sensitivity to domain-specific traffic characteristics. Future work will address this limitation through domain-aware adaptation, expanded traffic diversity during training, and true cross-dataset evaluation without retraining.

Although the proposed system assumes short-term stability in traffic statistics for adaptive calibration, future work will focus on long-term robustness under concept drift, online recalibration strategies, and validation on live network deployments to further enhance operational reliability.

6. Conclusions

This study proposes a unified decision-level adaptive intrusion detection framework that integrates hierarchical CNN-based closed-set classification with autoencoder-based zero-day detection in a cascade architecture. By confining adaptation to class-specific confidence thresholds and fusion parameters, the framework enables deployment-time calibration without retraining model weights, providing a parameter-efficient and operationally stable detection strategy. Runtime analysis confirms an average inference latency of 0.0135 ms per network flow and an adaptation overhead of only 124 microseconds per update, directly supporting the lightweight characterization of the proposed framework. Experimental evaluation on the CSE-CIC-IDS2018 dataset demonstrated strong closed-set performance, achieving 98.98% accuracy with a macro-F1-score of 0.9342, alongside effective zero-day detection with an F1-score of 0.9319 and a false positive rate of 0.0019 under a strict attack-exclusion protocol. Validation on the independent SUNE dataset confirmed that the proposed framework maintains consistent detection performance across heterogeneous traffic conditions when retrained under domain-specific settings, achieving 96.18% closed-set accuracy and 97.54% accuracy in the integrated cascade setting, demonstrating architectural generalizability across independent network environments rather than cross-dataset transfer capability. However, the reduced recall observed for certain SUNE attack categories indicates sensitivity to domain-specific traffic characteristics, highlighting the need for improved cross-environment generalization. Overall, the results demonstrate that decision-level adaptive calibration can effectively balance minority attack detection, zero-day sensitivity, and false-alarm control in resource-constrained network environments. Future work will focus on domain-aware adaptation, long-term stability under evolving traffic distributions, validation on live operational networks, and true cross-dataset evaluation in which models trained on one dataset are applied directly to another without retraining to assess domain-shift robustness under distribution shift.

Author Contributions: All authors contributed equally to the conceptualization, methodology, model development, experimentation, analysis, and manuscript preparation. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The CSE-CIC-IDS2018 dataset is publicly available via Kaggle. The SUNE dataset is not publicly available but may be made available upon reasonable request from the corresponding author for non-commercial research purposes.

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

References

1. Ayo, F.E.; Folorunso, S.O.; Abayomi-Alli, A.A.; Adekunle, A.O.; Awotunde, J.B. Network intrusion detection based on deep learning model optimized with rule-based hybrid feature selection. *Inf. Secur. J. Glob. Perspect.* **2020**, *29*, 267–283. [CrossRef]
2. Göcs, L.; Johanyák, Z.C. Identifying Relevant Features of CSE-CIC-IDS2018 Dataset for the Development of an Intrusion Detection System. *arXiv* **2023**, arXiv:2307.11544. [CrossRef]
3. Ahmed, U.; Nazir, M.; Sarwar, A.; Ali, T.; Aggoune, E.-H.M.; Shahzad, T.; Khan, M.A. Signature-based intrusion detection using machine learning and deep learning approaches empowered with fuzzy clustering. *Sci. Rep.* **2025**, *15*, 1726. [CrossRef]
4. Liu, H.; Patras, P. NetSentry: A deep learning approach to detecting incipient large-scale network attacks. *Comput. Commun.* **2022**, *191*, 119–132. [CrossRef]
5. Soltani, M.; Ousat, B.; Siavoshani, M.J.; Jahangir, A.H. An Adaptable Deep Learning-Based Intrusion Detection System to Zero-Day Attacks. *arXiv* **2021**, arXiv:2108.09199. [CrossRef]
6. Cobos, E.V.; Cakir, S. *A Review of the Economic Costs of Cyber Incidents*; World Bank Group: Washington, DC, USA, 2024. Available online: https://documents1.worldbank.org/curated/en/099092324164536687/pdf/P17876919ffee4079180e81701969ad0a18.pdf (accessed on 3 January 2026).
7. Arroyabe, M.F.; Arranz, C.F.A.; De Arroyabe, I.F.; De Arroyabe, J.C.F. Revealing the realities of cybercrime in small and medium enterprises: Understanding fear and taxonomic perspectives. *Comput. Secur.* **2024**, *141*, 103826. [CrossRef]
8. Khan, N.F.; Ikram, N.; Saleem, S. Effects of socioeconomic and digital inequalities on cybersecurity in a developing country. *Secur. J.* **2024**, *37*, 214–244. [CrossRef] [PubMed]
9. Siphambili, N. Exploring Cybersecurity Implications in Higher Education. *Eur. Conf. Cyber Warf. Secur.* **2024**, *23*, 526–531. [CrossRef]
10. Kifaru, F.; Kavuta, K.; Semlambo, A. Assessment of the Impacts of Cyber Security on Student Information Management Systems: A Case of Ruaha Catholic University. *J. Inform.* **2023**, *3*, 51–67. [CrossRef]
11. Ntembo, F.N.; Casmir, R. The Driving Forces for the Involvement of Higher Learning Institution's Students in Cybercrime Acts. A Case of Selected Higher Learning Institutions in Tanzania. *Eur. J. Theor. Appl. Sci.* **2023**, *1*, 911–922. [CrossRef]
12. Ngo, V.-D.; Vuong, T.-C.; Luong, T.V.; Tran, H. Machine Learning-Based Intrusion Detection: Feature Selection versus Feature Extraction. *arXiv* **2023**, arXiv:2307.01570. [CrossRef]
13. 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]
14. Ngan, T.; Haihua, C.; Janet, J.; Jay, B.; Junhua, D. Effect of Class Imbalance on the Performance of Machine Learning-based Network Intrusion Detection. *Int. J. Perform. Eng.* **2021**, *17*, 741. [CrossRef]
15. Shanmugam, V.; Razavi-Far, R.; Hallaji, E. Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches. *Electronics* **2024**, *14*, 69. [CrossRef]
16. Li, E.; Shang, Z.; Gungor, O.; Rosing, T. SAFE: Self-Supervised Anomaly Detection Framework for Intrusion Detection. *arXiv* **2025**, arXiv:2502.07119. [CrossRef]
17. Sharma, N.; Arora, B.; Ziyad, S.; Singh, P.K.; Singh, Y. A Holistic review and performance evaluation of unsupervised learning methods for network anomaly detection. *Int. J. Smart Sens. Intell. Syst.* **2024**, *17*, 20240016. [CrossRef]
18. Abdelhamid, M.; Desai, A. Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification. *arXiv* **2024**, arXiv:2409.19751. [CrossRef]
19. Leevy, J.L.; Johnson, J.M.; Hancock, J.; Khoshgoftaar, T.M. Threshold optimization and random undersampling for imbalanced credit card data. *J. Big Data* **2023**, *10*, 58. [CrossRef]
20. Issa, M.M.; Aljanabi, M.; Muhialdeen, H.M. Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. *J. Intell. Syst.* **2024**, *33*, 20230248. [CrossRef]
21. Khorram, T. Network Intrusion Detection using Optimized Machine Learning Algorithms. *Eur. J. Sci. Technol.* **2021**, *25*, 463–474. [CrossRef]
22. Al-Saleh, A. A balanced communication-avoiding support vector machine decision tree method for smart intrusion detection systems. *Sci. Rep.* **2023**, *13*, 9083. [CrossRef] [PubMed]
23. Hnamte, V.; Hussain, J. DCNNBiLSTM: An Efficient Hybrid Deep Learning-Based Intrusion Detection System. *Telemat. Inform. Rep.* **2023**, *10*, 100053. [CrossRef]

24. Liu, X.; Liu, J. Malicious traffic detection combined deep neural network with hierarchical attention mechanism. *Sci. Rep.* **2021**, *11*, 12363. [CrossRef] [PubMed]
25. Song, Y.; Hyun, S.; Cheong, Y.-G. Analysis of Autoencoders for Network Intrusion Detection. *Sensors* **2021**, *21*, 4294. [CrossRef]
26. Lotfi, Z.; Lotfi, M. A Comprehensive Study of Supervised Machine Learning Models for Zero-Day Attack Detection: Analyzing Performance on Imbalanced Data. *arXiv* **2025**, arXiv:2512.07030. [CrossRef]
27. Hindy, H.; Atkinson, R.; Tachtatzis, C.; Colin, J.-N.; Bayne, E.; Bellekens, X. Utilising Deep Learning Techniques for Effective Zero-Day Attack Detection. *Electronics* **2020**, *9*, 1684. [CrossRef]
28. Babaey, V.; Faragardi, H.R. Detecting Zero-Day Web Attacks with an Ensemble of LSTM, GRU, and Stacked Autoencoders. *Computers* **2025**, *14*, 205. [CrossRef]
29. Dai, Z.; Por, L.Y.; Chen, Y.-L.; Yang, J.; Ku, C.S.; Alizadehsani, R.; Pławiak, P. An intrusion detection model to detect zero-day attacks in unseen data using machine learning. *PLoS ONE* **2024**, *19*, e0308469. [CrossRef]
30. Kim, D.; Im, H.; Lee, S. Adaptive Autoencoder-Based Intrusion Detection System with Single Threshold for CAN Networks. *Sensors* **2025**, *25*, 4174. [CrossRef]
31. Lan, M.; Luo, J.; Chai, S.; Chai, R.; Zhang, C.; Zhang, B. A Novel Industrial Intrusion Detection Method based on Threshold-optimized CNN-BiLSTM-Attention using ROC Curve. In *Proceedings of the 2020 39th Chinese Control Conference (CCC)*; IEEE: Shenyang, China, 2020; pp. 7384–7389. [CrossRef]
32. Zoppi, T.; Gharib, M.; Atif, M.; Bondavalli, A. Meta-Learning to Improve Unsupervised Intrusion Detection in Cyber-Physical Systems. *ACM Trans. Cyber-Phys. Syst.* **2021**, *5*, 1–27. [CrossRef]
33. Alalhareth, M.; Hong, S.-C. Enhancing the Internet of Medical Things (IoMT) Security with Meta-Learning: A Performance-Driven Approach for Ensemble Intrusion Detection Systems. *Sensors* **2024**, *24*, 3519. [CrossRef]
34. Zhou, Y.; Ni, T.; Lee, W.-B.; Zhao, Q. A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluations. *arXiv* **2025**, arXiv:2502.05224. [CrossRef]
35. Abdallah, E.E.; Eleisah, W.; Otoom, A.F. Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey. *Procedia Comput. Sci.* **2022**, *201*, 205–212. [CrossRef]
36. Qiu, Z.; Zhou, D.; Zhai, Y.; Liu, B.; He, L.; Cao, J. VAEMax: Open-Set Intrusion Detection based on OpenMax and Variational Autoencoder. In *Proceedings of the 2024 5th Information Communication Technologies Conference (ICTC)*; IEEE: Nanjing, China, 2024; pp. 98–105. [CrossRef]
37. Hari, V.; Narang, P.; Alladi, T.; Chamola, V. FAST-IDS: A Fast Two-Stage Intrusion Detection System with Hybrid Compression for Real-Time Threat Detection in Connected and Autonomous Vehicles. *arXiv* **2025**, arXiv:2512.24391. [CrossRef]
38. Songma, S.; Sathuphan, T.; Pamutha, T. Optimizing Intrusion Detection Systems in Three Phases on the CSE-CIC-IDS-2018 Dataset. *Computers* **2023**, *12*, 245. [CrossRef]
39. Thakkar, A.; Lohiya, R. A Review of the Advancement in Intrusion Detection Datasets. *Procedia Comput. Sci.* **2020**, *167*, 636–645. [CrossRef]
40. Yousefimehr, B.; Ghatee, M.; Seifi, M.A.; Fazli, J.; Tavakoli, S.; Rafei, Z.; Ghaffari, S.; Nikahd, A.; Gandomani, M.R.; Orouji, A.; et al. Data Balancing Strategies: A Survey of Resampling and Augmentation Methods. *arXiv* **2025**, arXiv:2505.13518. [CrossRef]
41. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
42. Torabi, H.; Mirtaheri, S.L.; Greco, S. Practical autoencoder based anomaly detection by using vector reconstruction error. *Cybersecurity* **2023**, *6*, 1. [CrossRef]
43. Duhok Polytechnic University; Salih, A.A.; Abdulazeez, A.M. Evaluation of Classification Algorithms for Intrusion Detection System: A Review. *J. Soft Comput. Data Min.* **2021**, *2*, 31–40. [CrossRef]
44. Farrukh, Y.A.; Wali, S.; Khan, I.; Bastian, N.D. Detecting Unknown Attacks in IoT Environments: An Open Set Classifier for Enhanced Network Intrusion Detection. *arXiv* **2023**, arXiv:2309.07461. [CrossRef]
45. Al-Zewairi, M.; Almajali, S.; Ayyash, M.; Rahouti, M.; Martinez, F.; Quadar, N. Multi-Stage Enhanced Zero Trust Intrusion Detection System for Unknown Attack Detection in Internet of Things and Traditional Networks. *ACM Trans. Priv. Secur.* **2025**, *28*, 1–28. [CrossRef]
46. Alzughaibi, S.; El Khediri, S. A Cloud Intrusion Detection Systems Based on DNN Using Backpropagation and PSO on the CSE-CIC-IDS2018 Dataset. *Appl. Sci.* **2023**, *13*, 2276. [CrossRef]
47. Khan, M.A.; Kim, J. Toward Developing Efficient Conv-AE-Based Intrusion Detection System Using Heterogeneous Dataset. *Electronics* **2020**, *9*, 1771. [CrossRef]
48. Harini, R.; Maheswari, N.; Ganapathy, S.; Sivagami, M. An effective technique for detecting minority attacks in NIDS using deep learning and sampling approach. *Alex. Eng. J.* **2023**, *78*, 469–482. [CrossRef]

49. Hewapathirana, I.U. A Comparative Study of Two-Stage Intrusion Detection Using Modern Machine Learning Approaches on the CSE-CIC-IDS2018 Dataset. *Knowledge* **2025**, *5*, 6. [CrossRef]
50. Gharib, M.; Mohammadi, B.; Dastgerdi, S.H.; Sabokrou, M. AutoIDS: Auto-encoder Based Method for Intrusion Detection System. *arXiv* **2019**, arXiv:1911.03306. [CrossRef]