Majestic International Journal of AI Innovations
ISSN: 3142-8428
Volume 1, 2026, **pp.** 1-19, **Article ID:** CM2601105008
DOI: 10.65080/mjai.v1.CM2601105008

CONTENT MAJESTIC PUBLISHER

Research Article

Evaluation and Calibration of Cross-Dataset Robustness in IoT Intrusion Detection Systems: A Deployment-Aware Approach

Ola Madi Mohammed Al Mari[1, *]

[1]*University of Seville, United Arab Emirate*

Article History

Received: 11 March, 2026 Revised: 09 June, 2026 Accepted: 20 June, 2026 Published: 15 July, 2026

Abstract:

Introduction: The rapid expansion of Internet of Things (IoT) deployments has increased the cyber-attack surface and introduced heterogeneous traffic behaviour across devices, gateways, edge services, and network environments. Although many IoT intrusion detection studies report high performance under independent and identically distributed test conditions, such results often provide limited evidence of real deployment reliability, particularly for unseen hosts, cross-dataset transfers, and calibration drifts.

Methodology: This study presents a deployment-aware evaluation and calibration framework for IoT intrusion detection using a Denoising Autoencoder-Based Deep Feature Extraction (DAE-DFE) backbone. Rather than proposing a new neural architecture, this study focuses on robustness-oriented evaluation protocols and calibration-aware decision-making under domain shifts. The framework was evaluated using conventional IID splits, identifier-removed testing, and source-IP-based GroupSplit evaluation on NF-ToN-IoT-v2 to reduce the memorisation of the host. Cross-dataset robustness was assessed by converting Edge-IIoTset packet/protocol logs into pseudo-flows and testing Edge→NF and NF→Edge transfer using unsupervised threshold adaptation based on positive rate matching.

Results: On the NF-ToN-IoT-v2 GroupSplit, the framework achieved an F1 score of 0.9919 and ROC-AUC of 0.9997. In the Edge→NF cross-dataset setting, the model retained a meaningful ranking performance with ROC-AUC = 0.7962, while unsupervised threshold adaptation improved the target-domain accuracy from 0.4959 to 0.8878 and F1 score from 0.6630 to 0.8981.

Conclusion: The findings show that calibration-aware thresholding and deployment-realistic evaluation are essential for assessing IoT IDS reliability beyond the conventional IID accuracy.

Keywords: IoT security; intrusion detection; deep feature extraction; denoising autoencoder; domain shift; cross-dataset evaluation.

1. INTRODUCTION

The Internet of Things (IoT) has become a heterogeneous ecosystem with constrained devices, gateways, edge services, and cloud backends interacting with each other *via* various protocols and traffic patterns [1]. Owing to this heterogeneity, coupled with the lack of device hardening and long lifecycles, IoT devices are exposed to risks such as reconnaissance and scanning, botnet recruitment, Distributed Denial-of-Service (DDoS) attacks, lateral movement, and data exfiltration [2]. The latest IoT security surveys all agree that no single control, be it authentication, encryption, or access control, is fully effective in mitigating these attacks owing to the frequent occurrence of attacks that use misconfigurations, compromised endpoints, or malicious traffic diversion that still comply with valid credentials and encrypted messages. Network Intrusion Detection Systems (IDS) are one of the core layers of IoT defense-in-depth strategies that provide a layer of continuous monitoring and detection of behavioural anomalies, even when the inspection of the payload is impossible because of

[*]Address correspondence to this author at University of Seville, United Arab Emirates; E-mail: olaamer685@gmail.com

Cite as: Al Mari, O.M.M. Evaluation and calibration of cross-dataset robustness in IoT intrusion detection systems: A deployment-aware approach. *Majestic International Journal of AI Innovations*, 2026; *1*, 1–19. Article ID: CM2601105008. https://doi.org/10.65080/mjai.v1.CM2601105008

encryption or resource limitations [3]. The same view is widely echoed in contemporary surveys of IoT security practices which depict IDS as a realistic, factual, and supportive supplement to preventive measures [4, 5].

One continued difficulty in IoT IDS is that discriminative patterns can be due to a complex interaction among flow-level statistics (packets, bytes, durations, inter-arrival behaviour, and protocol signals) and not just with single features [6]. Traditional IDS-based pipelines are typically based on hand-engineered features and models, which presuppose near-linear separability or stable distributions, which are fragile when the devices, firmware, and network baselines vary. Conversely, the goal of representation learning is to convert raw or semi-processed telemetry into low-dimensional latent variables to learn higher-level structures and nonlinear relationships [7]. Autoencoder-based feature learning has been reiteratively embraced in the security of the IoT since it can encode behavioural signatures of traffic summaries and can be applied to anomaly detection as well as to supervised classification; for example, deep autoencoders have been applied to detect botnet-driven traffic in the IoT through observation of network traffic [4, 8].

Although there has been an improvement in the model architectures, a significant gap exists between the model architecture evaluation and the real-life implementation of IID benchmarks. IoT networks are subject to churn of devices, invisible hosts, changing firmware, and alteration of traffic bases, all of which present domain shifts. How heterogeneity of features and attack types makes evaluation reproducible by studying a dataset has also been noted, and can overstate performance when test conditions are too similar to training conditions [4, 9]. Another problem is threshold transfer, which has been under-investigated in the context of operations. Although a model has maintained ranking performance across domains (high ROC-AUC), the probability scores can be miscalibrated, which makes it unsafe to rely on a fixed point of decision-making. It has been discovered by calibration research that state-of-the-art neural networks can cause overconfident or biased probability estimates with distribution change, which can be mitigated by using calibration-aware decisioning and threshold adjustment [10, 11].

This study does not claim novelty by proposing a new neural network architecture. Instead, its contribution lies in a deployment-aware evaluation and calibration framework for IoT intrusion detection under realistic, non-IID conditions. Using a denoising autoencoder-based deep feature extractor as a representative backbone, this study examines whether high within-dataset detection performance remains reliable when host identities are unseen, feature identifiers are removed, and the model is transferred across heterogeneous IoT datasets. The first contribution is a robustness-oriented evaluation protocol that combines conventional IID testing, identifier-removed testing, and source-IP-based GroupSplit evaluation to reduce the risk of host memorisation. The second contribution is a cross-dataset transfer setting between Edge-IIoTset and NF-ToN-IoT-v2, enabled by pseudo-flow construction for canonical feature alignment. The third contribution is an unsupervised threshold adaptation strategy based on positive-rate matching, which addresses calibration drift without using labelled target domain data. The fourth contribution is a systematic interpretation of transfer asymmetry, showing that deployment failure can arise not only from weak model capacity but also from telemetry granularity mismatches and unstable decision thresholds. The framework was evaluated on NF-ToN-IoT-v2, a NetFlow-based IoT intrusion dataset, using both an identifier-randomised generalisable regime and a GroupSplit protocol by source IP to limit host memorisation. Cross-environment deployment was examined through cross-dataset evaluation from Edge-IIoTset to NF-ToN-IoT-v2 by converting packet/protocol logs into pseudo-flows and applying an unsupervised threshold adaptation algorithm based on positive rate matching to address calibration drift. An ablation study was also conducted to separate the effects of denoising and reconstruction regularisation, which showed that the framework performance did not strongly depend on individual training components. The primary contributions of this study are as follows: (i) a calibration-aware cross-dataset evaluation for IoT intrusion detection systems, (ii) a host-split robustness analysis to mitigate host memorisation, and (iii) an unsupervised threshold transfer under domain shift.

2. RELATED WORK

2.1. IoT IDS Datasets and Realism

The study of IOT intrusion detection has observed that dataset selection has a significant effect on visible functionality and applicability. A common limitation of existing benchmark datasets is that many are outdated, provide limited attack diversity, or contain collection biases that inflate reported performance, are biased in their coverage of attack diversity, or are edited in a manner that biases classification, either inadvertently exaggerating the IID outcomes or misrepresenting the actual heterogeneity of the real world. Recently, surveys have highlighted that to achieve reproducible IoT IDS evaluation, datasets to capture heterogeneous device behaviours, protocols, and realistic manifestations of attacks, as well as to record feature definitions to prevent hidden leakage or inconsistent labelling, are needed [12, 13]. In the context of contemporary IoT-centric datasets, ToN_IoT was directly mentioned as a heterogeneous IoT telemetry resource, and the fact that features and attack types should be standardised across datasets was emphasised. The associated TON IoT telemetry dataset offered a new generation IoT/IIoT dataset to drive data-driven IDS, reflecting the changing community towards more comprehensive telemetry data, as opposed to legacy corpora [9]. Based on these streams, NF-ToN-IoT-v2 added a NetFlow-like representation, which permits flow statistics to be used to scale an IDS experiment, and Edge-IIoTset added a new all-IoT/IIoT dataset that can be used to conduct centralised and federated learning evaluations. Taken together, these datasets appear to indicate a general direction: a more detailed evaluation of IoT needs to involve heterogeneous sources of traffic and feature management; however, comparison across datasets is challenging because they vary in granularity (packet/protocol logs versus NetFlow summaries) and feature semantics [12].

2.2. ML and DL IDS Approaches in IoT

IoT IDS methods include classical machine learning (*e.g.,* RF, SVM, boosted trees) and deep learning (*e.g.,* multilayer perceptrons, CNNs, RNN/LSTMs, autoencoder-based feature learning). It is consistently reported. They can be competitive on structured flow features with feature engineering being strong, and deep models being competitive when relationships are highly nonlinear or hand-crafted features are being deemphasized by representation learning [13]. Nevertheless, according to the same surveys, most published systems continue to be tested on predominantly IID splits, which are misleadingly robust when the test set has device identities, traffic baselines, or collection conditions identical to those in training [12]. This weakness has prompted an increased number of studies on methods that actively consider heterogeneity and transfer, such as adaptive learning methods and domain-sensitive evaluation systems, as IoT implementations frequently face unknown hosts and changing operation points [14].

2.3. Autoencoders for Feature Learning and Anomaly/IDS

Autoencoders have become a common component of IDS because they learn succinct latent representations that encode traffic behaviour and can be used not only to detect anomalies (through reconstruction error) but also to provide supervised classification (through a discriminative head) [15]. This is because previous studies have shown that deep autoencoder representations can capture behavioural patterns associated with botnet behaviour when observed on the network, which indicates the feasibility of feature learning by autoencoders in an IoT setting [16]. One of these design issues is the trade-off between reconstruction and discriminative goals (where reconstruction promotes faithfulness to the input structure and discrimination maximises class separation). Autoencoders can also use denoising regularisation, which stabilises representations to input perturbations, as it has been claimed to increase noise sensitivity and network telemetry stability [12, 17]. Practically, this work stream implies that autoencoder-based feature extraction is highly compatible with the needs of IoT IDS; however, its practical significance should be proven not only on the basis of IID evaluation but also based on real splits and domain shifts [13].

2.4. Calibration and Decision Thresholding Under Domain Shift

Another practical problem, which is usually not reported in the literature of IDS, is that the performance of the decisions is not only determined by the quality of the ranking but also by the calibration of the probability and the choice of the threshold. In distribution shift, models can maintain useful ranking accuracy (high ROC-AUC/PR-AUC) as well as generate shifted or overconfident probability estimates, resulting in a fixed threshold that results in poor accuracy or excessive false positives/negatives [13, 18]. This is consistent with the larger calibration results of machine learning, in which contemporary neural networks are often miscalibrated and require a post-hoc or operational change. Calibration-aware decision-making and threshold adjustment are important when deployment conditions differ from training conditions, particularly when the false alarm cost is high [14]. The research gap includes (i) extracting deep features, (ii) performing robustness analysis on host-split protocols, and (iii) deploying cross-datasets which explicitly handle threshold transfer with calibration-aware adaptation.

3. PROBLEM DEFINITION AND FRAMEWORK OVERVIEW

The IoT intrusion detection problem was formulated as a supervised binary classification task. Let denote a preprocessed IoT traffic record containing flow-level or pseudo-flow-level features, and let denote the corresponding label, where represents benign traffic, and represents attack traffic. The objective is to learn a detection function, where represents the estimated probability that the traffic record belongs to the attack class.

The final IDS decision is obtained using the threshold:

$$\hat{y}_i = \begin{cases} 1, & \text{if } \hat{p}_i \geq \tau, \\ 0, & \text{if } \hat{p}_i < \tau. \end{cases}$$

In conventional IID evaluation, is selected using a validation split drawn from the same distribution as the training data. However, in deployment settings, the target domain score distribution may differ from the source domain score distribution because of unseen hosts, device churn, protocol heterogeneity, firmware variation, or different traffic collection conditions. Therefore, a threshold that is optimal in the source domain may generate excessive false positives or negatives in the target domain. This motivates the use of a calibration-aware threshold-adaptation layer in the proposed framework.

The overall workflow consists of data source preparation, robust preprocessing, DAE-DFE feature learning, probability estimation, threshold selection or adaptation, and final IDS decision-making. Therefore, the framework evaluates not only whether the model can separate benign and attack traffic but also whether the decision threshold remains reliable under host shift and cross-dataset domain shift [4, 19]. An overview of the end-to-end flow is presented in Fig. (**1**), which summarises the end-to-end flow: dataset → preprocessing → DAE-DFE → outputs → threshold/adaptation → IDS decision.

4. METHODOLOGY

4.1. Datasets and Label Definitions

The proposed framework of deep feature extraction was evaluated on two complementary datasets of IoT cybersecurity that are not only in different granularities and thus allow not only the analysis of robust within-dataset but also the analysis of cross-dataset generalisation. In the main IoT IDS experiment, NF-ToN-IoT-v2 was employed as a flow/NetFlow-based dataset, with each record being aggregated network behaviour in terms of protocol and volume statistics (*e.g.,* packet/byte counts, duration, TCP flags, and throughput-related attributes) and scored in advance with supervised intrusion detection [20]. The Label field of the dataset (benign *vs.* attack) was used as a binary target variable, and to avoid inconsistent optimisation

and similar operating points of the algorithms under evaluation conditions, a balanced subset of 600,000 records (300,000 each class) was created. To prevent feature leakage and provide an authentic learning context, the Attack column (attack family/descriptor) was excluded before model fitting because it could capture information generated by the labelling processes and was therefore not an input feature in deployment.

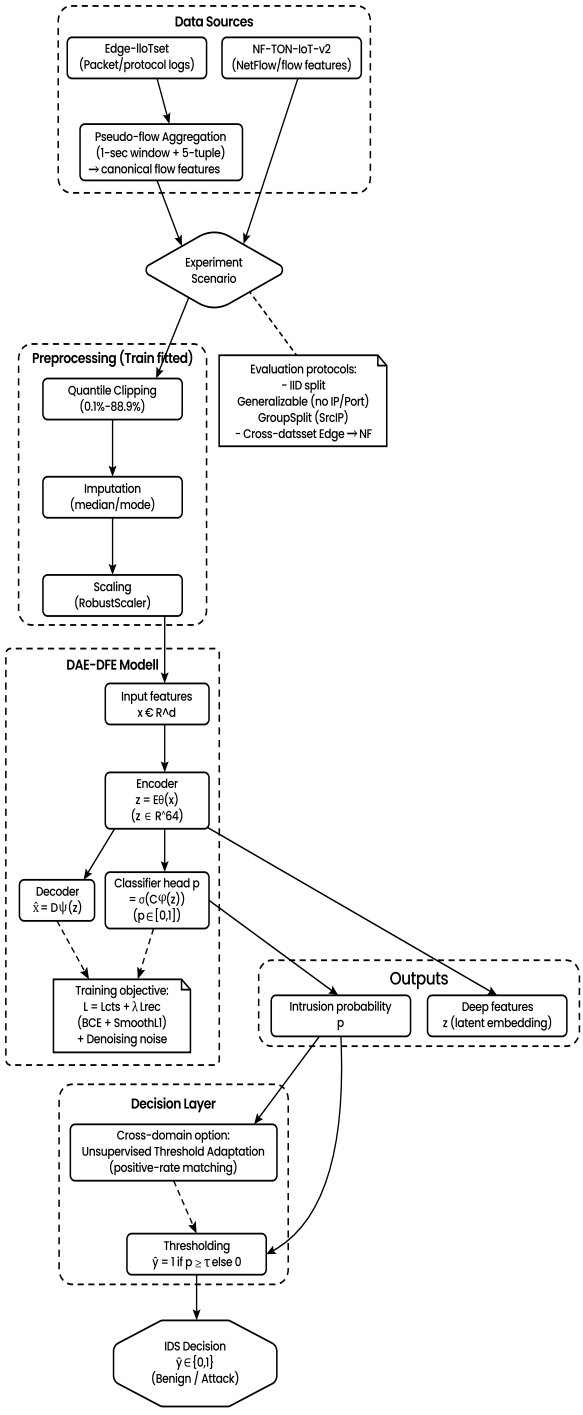

Fig. (1). Deployment-aware DAE-DFE framework overview.

To evaluate cross-dataset deployment, Edge-IIoTset was deployed as a heterogeneous IoT/IIoT dataset that could be used to support the modern paradigm of learning (centralised and federated learning) [21]. In contrast to NF-ToN-IoT-v2, the Edge-IIoTset DNN file includes telemetry at the packet/protocol level, with numerous fields providing application/protocol traces as opposed to NetFlow summaries. This granularity mismatch has been known to be a barrier to cross-dataset IDS evaluation, since direct feature intersection may fail to tell its identity and generate false transfer conclusions. In this respect, Edge-IIoTset records were not processed as flows directly, but rather aggregated into pseudo-flows to allow both datasets to be expressed in a common canonical flow feature space that could be used to make cross-domain inferences.

4.2. Data Sampling and Splits

For NF-ToN-IoT-v2, a balanced sample was created using chunk-based reading of the raw CSV files to avoid memory limitations while maintaining reproducibility on commodity computing resources. The resulting dataset of 600,000 records was divided into training, validation, and test sets using a 70/15/15 ratio under an IID stratified protocol. To make the evaluation more realistic and reduce the risk of host memorisation, an additional GroupSplit protocol was applied using the source IP as the grouping variable, such that source identities appearing in the test set were not observed during training. This protocol better reflects operational IoT deployment, where new hosts and devices may appear after the model training.

To conduct a cross-dataset evaluation, the source domain was Edge pseudo-flows and the target domain was NF flows (Edge→NF), and vice versa (NF→Edge). The model was also trained and threshold-tuned in each direction using the source training/validation data alone and then used on the entire target set without retraining. Importantly, all preprocessing settings (clipping values, imputation values, and scaling values) were trained on the source training data and used verbatim on the source validation/test data and target domain, which avoided leakage *via* normalisation.

The use of multiple evaluation protocols was intended to separate the conventional model accuracy from the deployment robustness. The IID stratified split provides a controlled baseline and allows for a comparison with prior IoT IDS studies. The identifier-removed setting reduces the possibility that the model relies on direct IP/port identity features rather than generalisable flow statistics. The source-IP GroupSplit protocol provides a stronger operational test because the source hosts appearing in the test set are not observed during training. This design is important for IoT deployments, where new devices, gateways, or hosts may appear after the IDS has been trained. For cross-dataset testing, the preprocessing parameters were fitted only to the source training data and applied unchanged to the validation, source test, and target domain data. This prevents leakage through normalisation and ensures that the reported cross-domain performance reflects the true deployment transfer rather than the target-domain preprocessing adaptation.

4.3. Preprocessing

NetFlow-style features, such as byte counts, packet counts, throughput, and duration, often exhibit heavy-tailed distributions. These distributions can destabilise the model optimisation and amplify the influence of extreme observations. To address this issue, a robust preprocessing pipeline was implemented. First, the numeric variables were clipped using training-set quantiles in the 0.1%–99.9% range. Second, missing numeric values were imputed using the median, and categorical values were imputed using the mode, where applicable. Third, RobustScaler normalisation was applied so that scaling was based on the interquartile range rather than the variance, which reduced the sensitivity to outliers.

4.4. DAE-DFE Architecture

The DAE-DFE model was implemented as a denoising autoencoder with a supervised classification head. Given an input vector $x \in \mathbb{R}^d$, Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is added during training to obtain a corrupted input Eq. (**1**).

$$\tilde{x} = x + \epsilon. \tag{1}$$

The encoder $g_\phi(\cdot)$ maps the corrupted input into a compact latent representation Eq. (**2**).

$$z = g_\phi(\tilde{x}), \tag{2}$$

where $z \in \mathbb{R}^m$ and $m < d$. The decoder $h_\psi(\cdot)$ reconstructs the input from the latent representation Eq. (**3**).

$$\hat{x} = h_\psi(z). \tag{3}$$

The classifier head $c_\omega(\cdot)$ estimates the intrusion probability Eq. (**4**).

$$\hat{p} = \sigma(c_\omega(z)), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid activation function. The model is trained using a joint objective that combines supervised classification loss and reconstruction regularization Eq. (**5**)

$$\mathcal{L} = \mathcal{L}_{BCE}(y, \hat{p}) + \lambda_{rec}\mathcal{L}_{rec}(x, \hat{x}), \tag{5}$$

where is the binary cross-entropy loss with positive-class weighting, is the Smooth L1 reconstruction loss, and controls the contribution of reconstruction regularisation. This design was selected because autoencoder-based representations can capture nonlinear interactions among flow-level traffic features, whereas the supervised head directly optimises attack discrimination. Denoising was included to improve stability against small perturbations, and reconstruction regularisation was used to prevent the latent representation from becoming purely label driven. However, the framework contribution does not depend on claiming a new architecture; instead, the DAE-DFE backbone is used to support the deployment-aware robustness and calibration analysis.

4.5. Cross-Dataset Pseudo-Flow Construction

For Edge-IIoTset, packet/protocol-level records were converted into pseudo-flows using a fixed one-second time window and a 5-tuple-like grouping key consisting of a time bin, source host, destination host, source port, and destination port. Within each group, packet-level observations were aggregated into a canonical flow-style representation to approximate the statistical structure of the NetFlow records. The generated pseudo flow features are listed in Table **1**.

Table 1. Canonical pseudo-flow features used for Edge-IIoTset harmonization.

No.	Feature	Description
1	Packet count	Number of packet/protocol records within the one-second pseudo-flow group
2	Byte-length proxy	Sum of available packet-length or payload-length indicators
3	Minimum packet-length proxy	Minimum observed packet-length proxy within the group
4	Maximum packet-length proxy	Maximum observed packet-length proxy within the group
5	Mean packet-length proxy	Average packet-length proxy within the group
6	Standard deviation of packet-length proxy	Dispersion of packet-length proxy values within the group
7	Mean TCP flag statistic	Mean encoded TCP flag value within the pseudo-flow
8	Acknowledgement-rate proxy	Proportion or mean value of acknowledgement-related flag indicators
9	Protocol indicator	Encoded protocol-level indicator after harmonisation
10	Source-port statistic	Aggregated or encoded source-port value for the pseudo-flow group
11	Destination-port statistic	Aggregated or encoded destination-port value for the pseudo-flow group
12	Window duration	Fixed window duration in milliseconds

This pseudo-flow conversion enabled the Edge-IIoTset and NF-ToN-IoT-v2 to be represented in a shared feature space for cross-dataset testing. However, pseudo-flows are heuristic approximations, rather than exporter-generated NetFlow records. They may omit the packet-level timing microstructure and may not fully preserve the connection-level semantics. Therefore, the cross-dataset results are interpreted as evidence of deployment transfer under approximate feature harmonisation rather than as a perfect one-to-one mapping between packet logs and NetFlow exporters.

4.6. Unsupervised Threshold Adaptation

Cross-domain deployment is frequently affected by calibration drift, where the ranking quality of the predicted

scores may remain useful, but the absolute probability distribution shifts between the source and target domains. In such cases, directly transferring a source-domain threshold may produce poor decision-level accuracy, even when the ROC-AUC and PR-AUC remain meaningful. To address this issue, an unsupervised threshold adaptation layer was used for cross-dataset testing.

Let r_s denote the positive prediction rate observed on the source validation split Eq. (**6**):

$$r_s = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbb{I}(\hat{p}_i^s \geq \tau_s), \tag{6}$$

where is the threshold selected for the source validation split, represents the source validation probabilities, and is where τ_s is the threshold selected on the source validation split, $\hat{p}_i^s$ represents source validation probabilities, and $\mathbb{I}(\cdot)$ is the indicator function. For an unlabeled target domain with predicted probabilities $\{\hat{p}_1^t, \hat{p}_2^t, \ldots, \hat{p}_{n_t}^t\}$, the adapted target threshold τ_t is selected as: Eq. (**7**):

$$\tau_t = Q_{1-r_s}(\{\hat{p}_i^t\}_{i=1}^{n_t}), \tag{7}$$

where Q_{1-r_s} denotes the $(1-r_s)$-quantile of target-domain scores. This method does not use target labels and therefore reflects a practical deployment scenario in which defenders may have access to target-domain IDS scores but not immediate ground-truth labels. The adapted threshold shifts the operating point to match the source-domain alert rate while preserving the trained model parameters.

5. EXPERIMENTAL RESULTS

The experimental assessment reports threshold-dependent and threshold-free values to represent the IDS operating realities. Accuracy, precision, recall, and F1 give a summary of performance at a selected decision threshold, whereas ROC-AUC and PR-AUC determine the quality of ranking regardless of a particular decision threshold. PR-AUC is of special interest to intrusion detection because security telemetry is frequently skewed in practice and ROC curves may be skew-optimistic; precision-recall analysis is strongly suggested in this context because it focuses on the positive predictive value of alarms [22]. The consideration of feature inspection supports the necessity of powerful preprocessing: (Fig. **2**) (histograms) indicates that the variables of the form of IN_BYTES and IN_PKTS are heavily skewed and heavy-tailed, and the protocol encodings and port distributions are sparse/discrete, with common service ranges. The correlation heatmap (Fig. **3**) indicates that there are significant dependencies between volume and timing features (*e.g.,* packet/byte count and duration-related features) and structured relationships between TCP flag variables, which makes a nonlinear feature learning method more appropriate than the assumptions of linear separability.

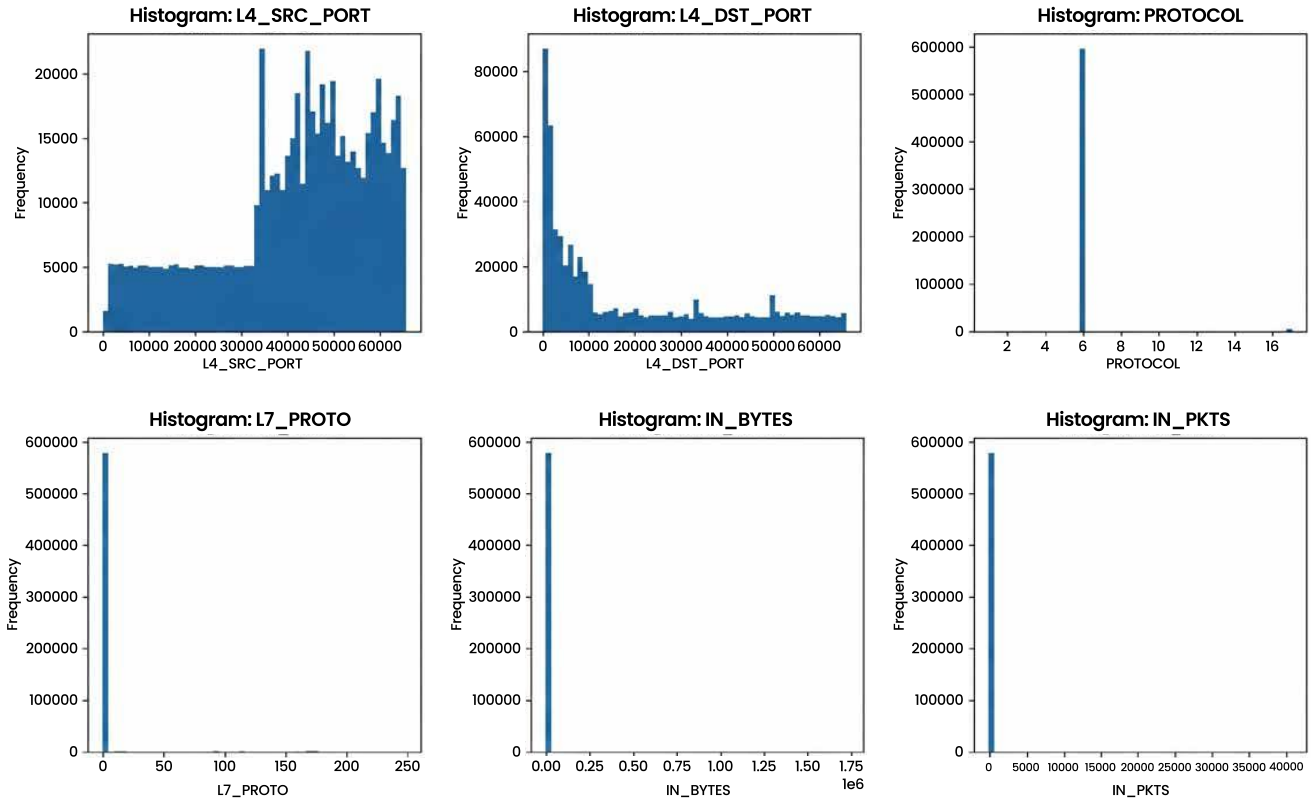

Fig. (2). Distribution of selected NF-ToN-IoT-v2 flow-level features.

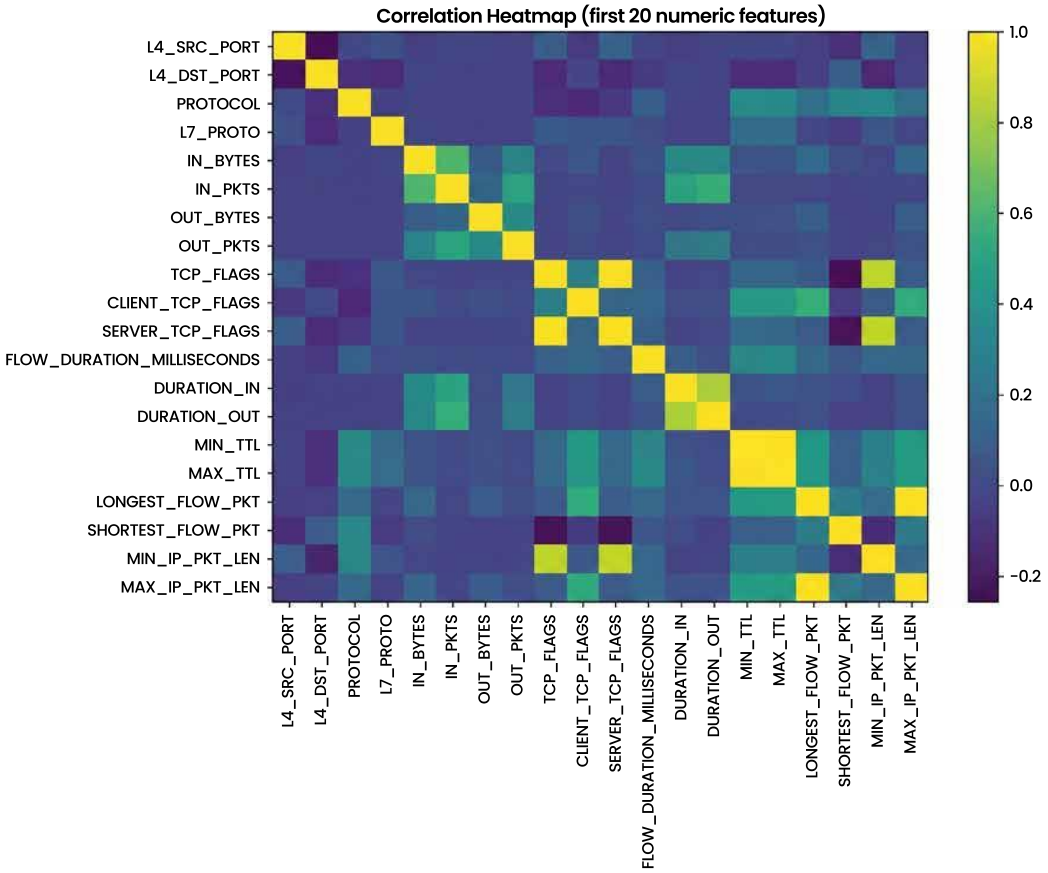

Fig. (3). Correlation heatmap of selected numeric flow-level features.

5.1. NF-ToN-IoT-v2 Results

The main findings of NF-ToN-IoT-v2 are presented in Table **2** and represented by the confusion matrices, ROC, PR, and calibration curves in Figs. (**4**–**6**). DAE-DFE obtained Accuracy = 0.9944 and F1 = 0.9944 with ROC-AUC = 0.9998 and PR-AUC = 0.9998 under the Operational IID protocol, indicating near-perfect discrimination on this benchmark under controlled evaluation settings. Removal of identifiers (Generalisable IID) did not alter performance in any material way (Accuracy = 0.9940, F1 = 0.9940; ROC-AUC = 0.9998; PR-AUC = 0.9998), indicating that the detection performance based on z-base representation is not dependent on memorisation of identities of IP/port. Even with the Operational GroupSplit (SrcIP) protocol, a stronger protocol that tests unseen source hosts, the framework still achieved Accuracy = 0.9895 and F1 = 0.9919 (ROC-AUC = 0.9997; PR-AUC = 0.9998), and thus, exhibited a strong ability to generalise to new hosts. The calibration panels in Figure 4 also suggest the consistent probability behaviour of DAE-DFE in these settings, which can be used to deploy threshold-based DAE-DFE to NetFlow-style IoT telemetry. Figure 5 shows the performance of the DAE-DFE under the generalisable IID setting. The confusion matrix indicates a high number of true positives and a low number of false positives, confirming that removing the direct IP/port identifiers does not materially reduce the detection performance. Fig. (**6**) presents the source-IP GroupSplit results, where the model maintains a strong performance even when the test source hosts are not observed during training. These results suggest that DAE-DFE learns general flow-level attack patterns rather than relying solely on host memorisation. Figure 6 illustrates the performance metrics of the model (DAE-DFE) for the SrcIP group split. The confusion matrix showed a high number of true positives (143,683) and low false positives (1,783), indicating a strong model. The ROC and PR curves both exhibited near-perfect performance (AUC approaching 1), whereas the calibration plot showed good alignment of predicted probabilities with actual outcomes.

The threshold was selected on the validation split for each evaluation setting. The consistently high F1-score and ROC-AUC values indicate that NF-ToN-IoT-v2 contains strongly discriminative flow-statistic patterns. However, these near-saturation results should be interpreted with caution. They showed that DAE-DFE performs strongly under the evaluated protocols, but they should not be overclaimed as evidence of architectural novelty. The limited degradation under GroupSplit suggests that the model does not rely only on direct host memorisation; however, the dataset itself remains favourable because many attack records differ substantially from benign traffic in terms of volume, timing, and protocol-related statistics (Table **2**).

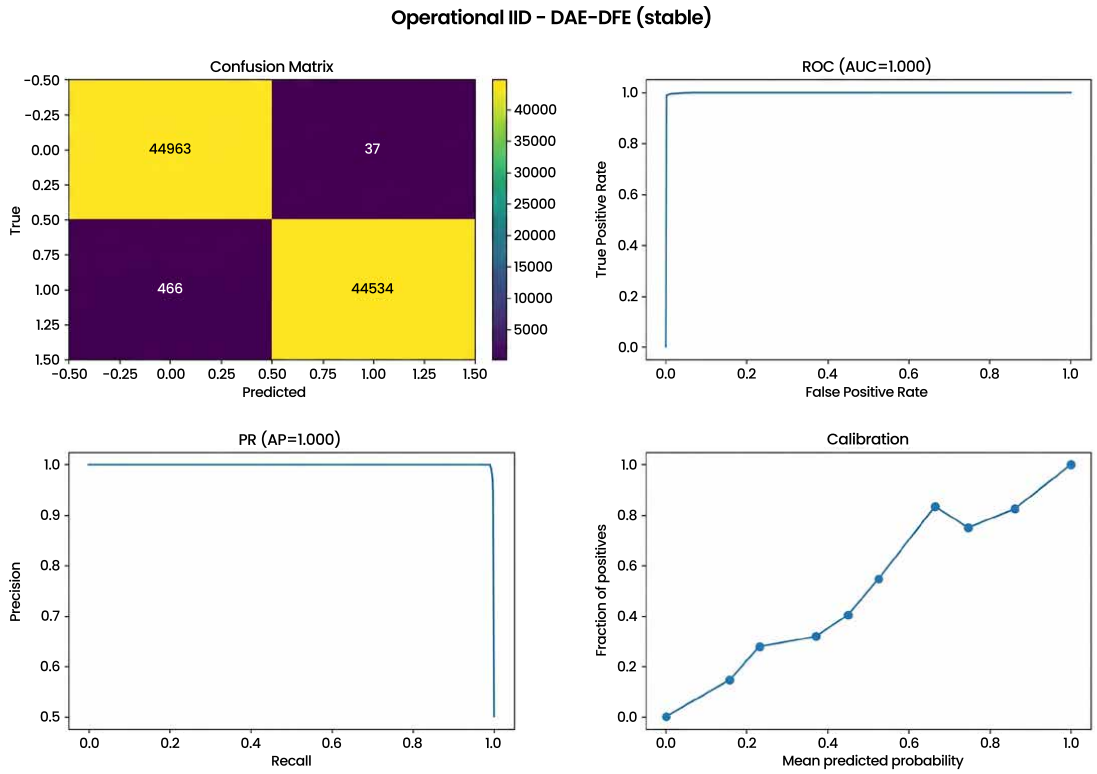

Fig. (4). DAE-DFE performance under Operational IID evaluation.

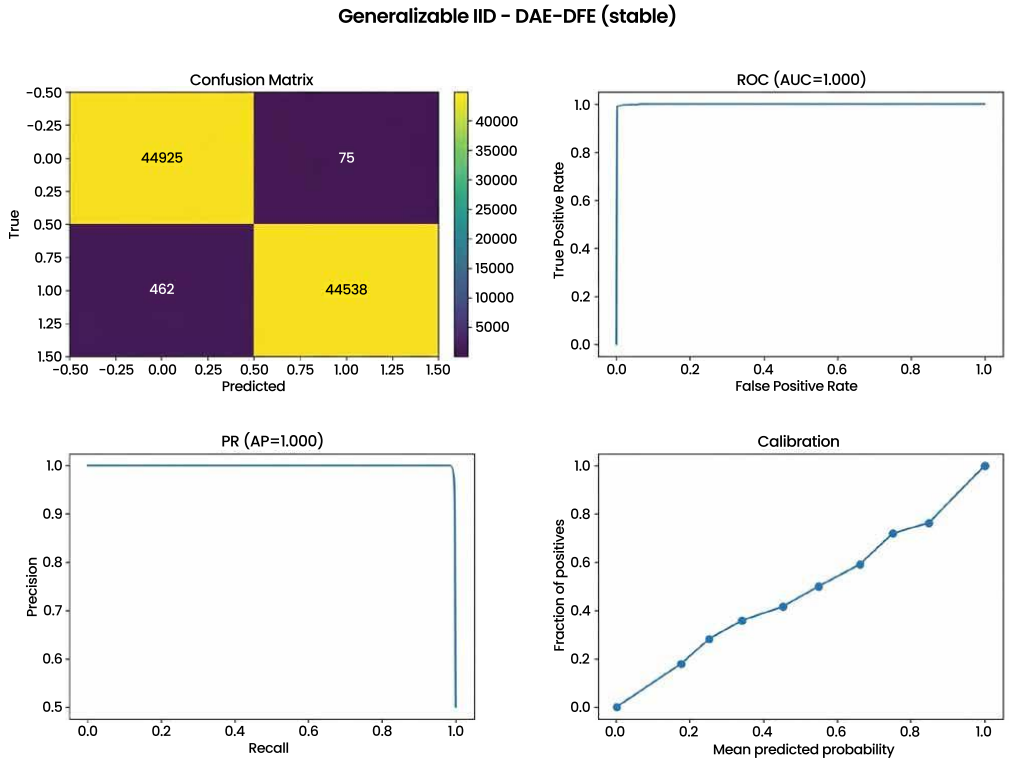

Fig. (5). DAE-DFE performance under identifier-removed IID evaluation.

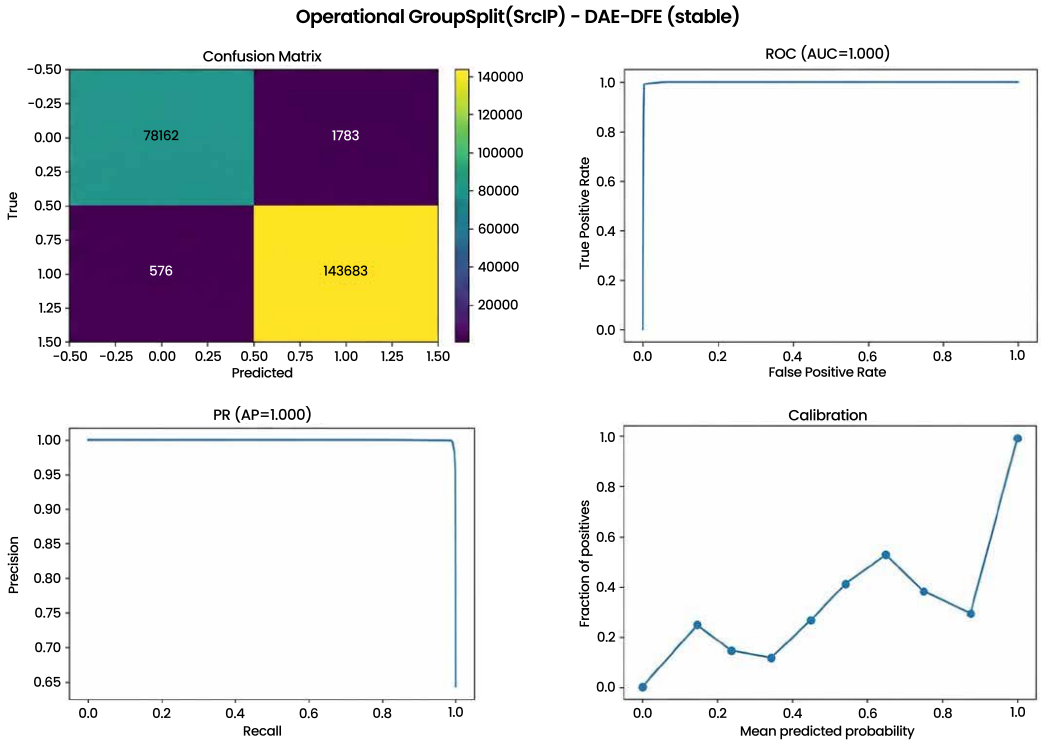

Fig. (6). DAE-DFE performance under source-IP GroupSplit evaluation.

Table 2. NF-ToN-IoT-v2 performance of DAE-DFE under deployment-aware evaluation protocols evaluation of DAE-DFE on NF-ToN-IoT-v2 under IID, identifier-removed, and host-based GroupSplit protocols.

Evaluation setting	Threshold (\tau)	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Operational IID	0.59	0.9944	0.9992	0.9896	0.9944	0.9998	0.9998
Generalizable IID, no IP/Port	0.47	0.9940	0.9983	0.9897	0.9940	0.9998	0.9998
Operational GroupSplit, SrcIP	0.40	0.9895	0.9877	0.9960	0.9919	0.9997	0.9998

5.2. Cross-dataset Edge→NF

A cross-dataset analysis was performed to determine whether the suggested deep feature extraction framework can be helpful when the environment in which it is deployed is different from the one in which it is being trained. Table **3** reports the cross-dataset transfer results for Edge→NF and NF→Edge, including source domain testing, direct threshold transfer, and unsupervised threshold adaptation. (Fig. **7**) shows that the source test Edge to NF has a high discrimination (ROC-AUC ≈ 0.968; PR-AUC ≈ 0.961), and the confusion matrix indicates an equal error contribution with a relatively low false positive and false negative, indicating that the model acquires meaningful information in the source domain. When the identical model and validated threshold were directly applied to the NF target domain (Fig. **8**), the decision performance dropped sharply, and the accuracy was approximately 0.496, even after the ROC-AUC was approximately 0.796 and PR-AUC was approximately 0.801. This mismatch is demonstrated by the confusion matrix: the positive class was predicted by the model for almost all samples, and the false-positive rate was very large in response to benign flows (Fig. **9**).

Threshold adaptation does not use target labels. The ROC-AUC and PR-AUC are threshold-independent metrics and therefore remain unchanged after threshold adaptation. The Edge→NF results show that direct threshold transfer causes a collapse in the decision-level performance despite a meaningful ranking performance. Accuracy improved from 0.4959 to 0.8878 and F1-score improved from 0.6630 to 0.8981 after unsupervised threshold adaptation. This indicates that the main Edge→NF problem was calibration drift, rather than complete representation failure. In contrast, the NF→Edge transfer remained weak even after adaptation because the ranking quality itself was poor, with ROC-AUC = 0.4482. This suggests a deeper feature-semantic mismatch between exporter-defined NetFlow summaries and heuristic Edge pseudo-flows (Table **3**).

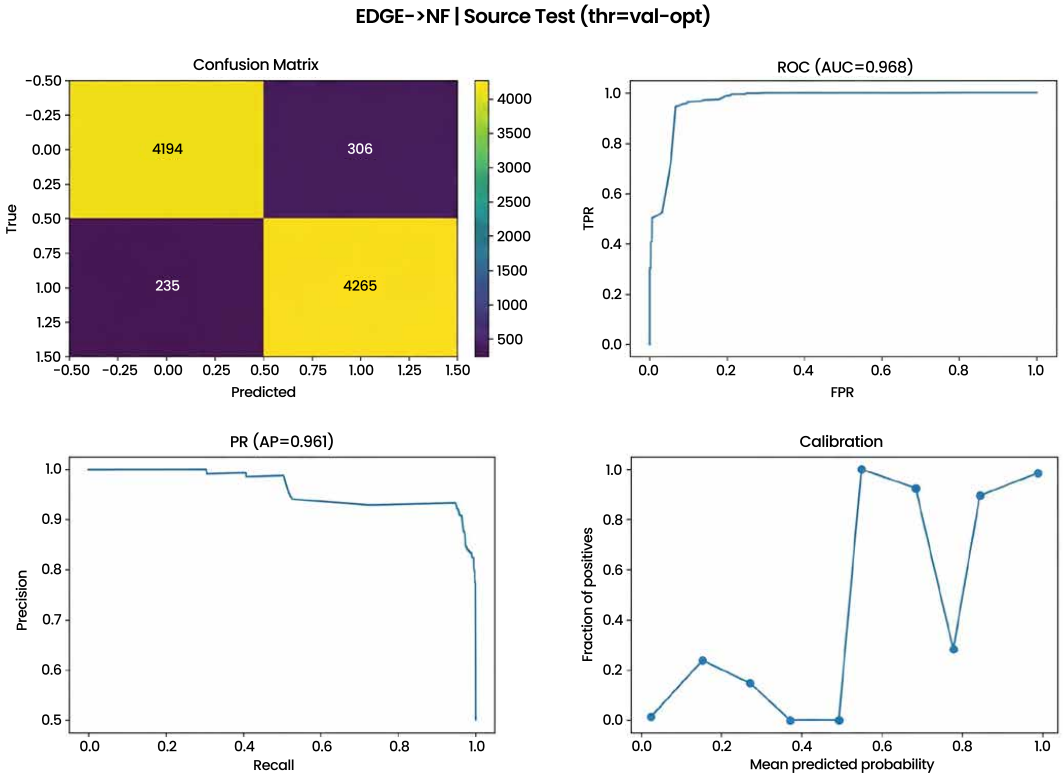

Fig. (7). Edge→NF source-domain test performance.

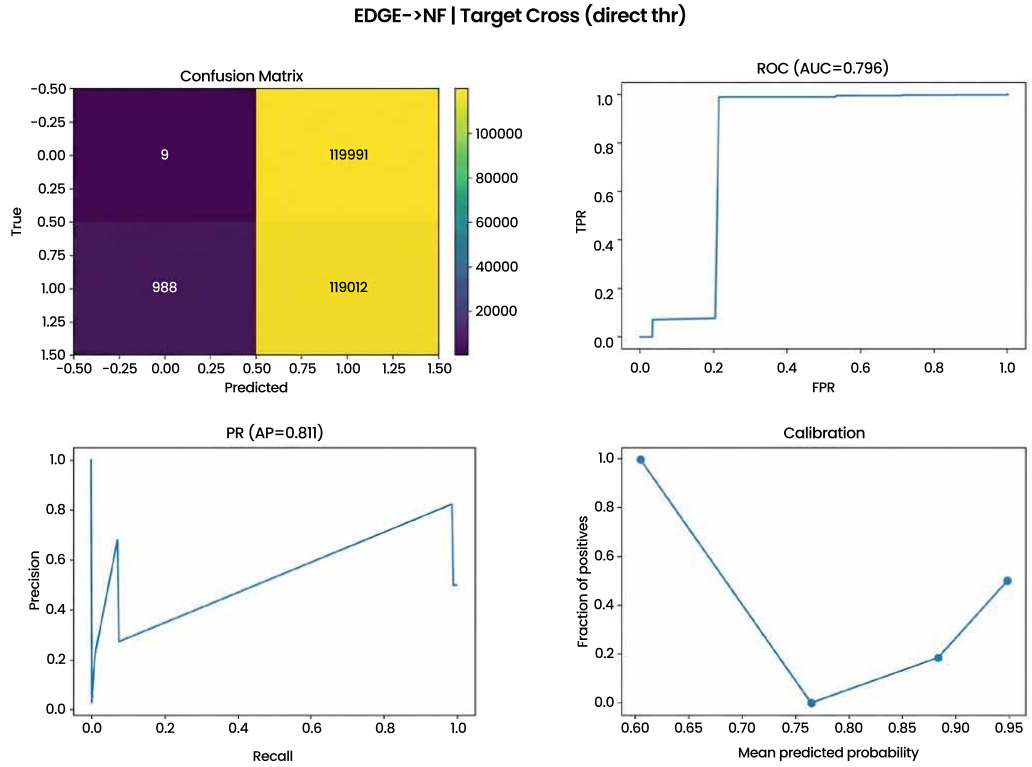

Fig. (8). Edge→NF target performance using direct threshold transfer.

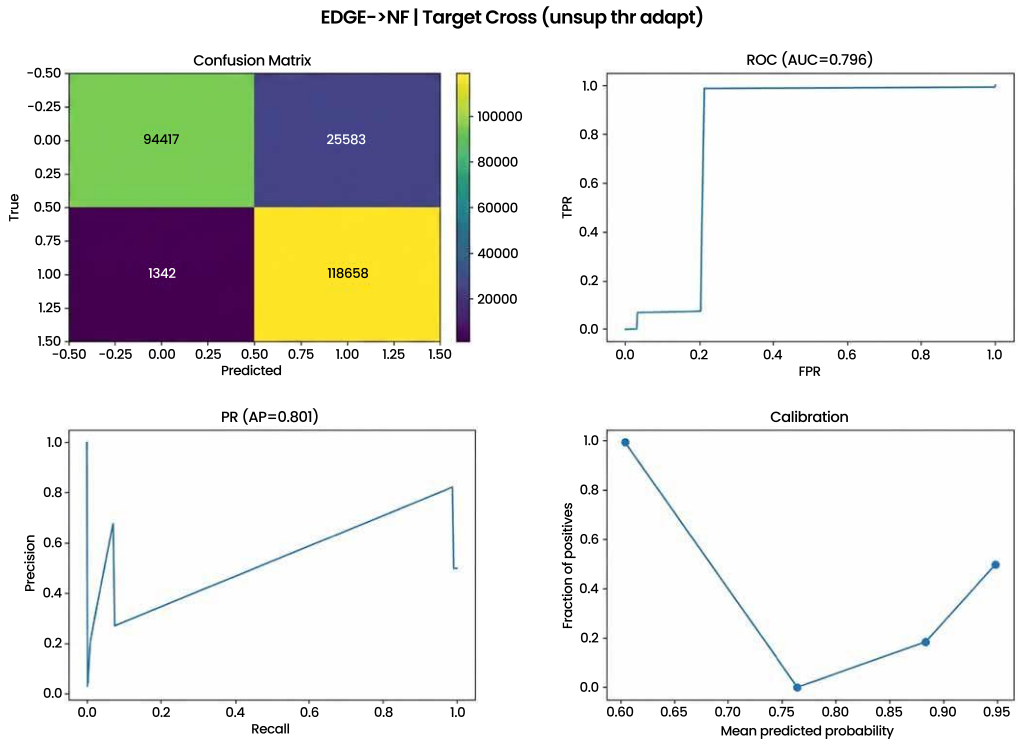

Fig. (9). Edge→NF target performance using adapted threshold.

Table 3. Cross-dataset performance of DAE-DFE under direct and adapted threshold transfer Cross-domain evaluation using edge pseudo-flows and NF-ToN-IoT-v2 flows. The direct threshold" applies the source validation threshold to the target. "Unsupervised adaptation" uses positive rate matching on unlabelled target scores.

Direction	Scenario	Threshold (\tau)	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Edge→NF	Source test	0.80	0.9399	0.9331	0.9478	0.9404	0.9679	0.9606
Edge→NF	Target direct threshold	0.80	0.4959	0.4980	0.9918	0.6630	0.7962	0.8013
Edge→NF	Target unsupervised adaptation	0.95	0.8878	0.8226	0.9888	0.8981	0.7962	0.8013
NF→Edge	Source test	0.39	0.9897	0.9887	0.9907	0.9897	0.9939	0.9898
NF→Edge	Target direct threshold	0.39	0.4019	0.4434	0.7695	0.5626	0.4482	0.6004
NF→Edge	Target unsupervised adaptation	0.94	0.4529	0.4563	0.4925	0.4737	0.4482	0.6004

The cross-dataset results demonstrate that deployment failure can arise from two distinct causes: threshold miscalibration and representation mismatches. In the Edge→NF direction, the model retained a meaningful ranking performance on the NF target domain, as indicated by ROC-AUC = 0.7962 and PR-AUC = 0.8013. However, the direct transfer of the source threshold produced excessive positive predictions, resulting in low accuracy despite a very high recall. The adapted threshold corrected the operating-point error by shifting the decision boundary according to the distribution of unlabelled target scores. This explains why the threshold-dependent metrics improved substantially, whereas the ROC-AUC and PR-AUC remained unchanged.

The NF→ edge direction showed a more severe limitation. In this case, the ROC-AUC fell to 0.4482, indicating that the source-trained model did not rank the target domain examples reliably. Threshold adaptation cannot repair this type of failure because it only changes the operating point and does not change the learned feature representation. The likely cause of this is a telemetry granularity mismatch. NF-ToN-IoT-v2 contains exporter-defined flow summaries, whereas Edge-IIoTset pseudo-flows are generated through the time-window aggregation of packet/protocol records. Therefore, similar feature names do not necessarily imply equivalent statistical significance. This asymmetric transfer result is important because it shows that cross-dataset IDS evaluation requires not

only shared feature schemas but also careful semantic harmonisation of telemetry representations.

5.3. NF→Edge (Asymmetric Transfer)

The cross-dataset findings indicate evident asymmetry in the transfer performance. (Fig. **10**) (NF to Edge - Source Test): The ROC-AUC = 0.994 and PR-AUC = 0.990, and hence, almost perfect performance is observed under the domain of the NF source. The confusion table indicates that there were many true positives (17,832) and a small number of false positives (204), indicating that the classification accuracy was high. (Fig. **11**) (NF to Edge - Target Cross (Direct): ROC-AUC = 0.448 and PR-AUC = 0.600, which suggests a drastic decline in performance by the use of the NF-trained model on the Edge dataset. The confusion matrix revealed many false positives (28,972), indicating poor generalisation by the model because of semantic differences between the NF and Edge data depictions. (Fig. **12**) (NF to Edge - Target Cross (Adapted)): ROC-AUC = 0.448 and PR-AUC = 0.600 and poor performance even after threshold adaptation. Calibration enhances without restoring the ranking performance, which supports the notion that inherent feature semantic differences cannot be redressed using adaptation alone.

This study presents a hypothesis of granularity asymmetry: NF-ToN-IoT-v2 flows are exporter-specified summaries of steady and aggregated statistics across full connections. Conversely, Edge-IIoTset pseudo-flows are heuristic windowed approximations built on packet-level and protocol-level logs. Although the nominal similarity of the names of features in both representations may be similar (*e.g.,* the number of packets, count of bytes proxies, and duration), the statistical meaning of the term is significantly different. Consequently, NF flow representations are not transferred to the pseudo-flow distributions of Edge telemetry.

Qualitative evidence for this interpretation is provided by the cross-dataset behaviour in the common canonical feature space. The severe deterioration in the ranking performance (NF-AUC of approximately 0.45) in the NF-Edge direction, as opposed to the virtually perfect performance in the source domain, shows that there is a substantial distributional mismatch between NF flows and Edge pseudo-flows. This discrepancy indicates that the marginal distributions and semantic meanings of key flow statistics vary across datasets, although feature names may have exact matches. In turn, unmonitored threshold adaptation proves to be able to correct decision-level calibration, but is unable to restore ranking quality in case the representation semantics are not aligned.

It must be mentioned that positive-rate matching implicitly assumes that the target deployment environment has a more or less widely similar prevalence of attacks or alert budget as measured during source validation. When the actual attack rate in the target domain varies significantly, the adjusted threshold will cease to correspond to an optimal operating point. Therefore, positive-rate matching must not be considered a solution to the distribution shift but rather a pragmatic solution for matching calibration.

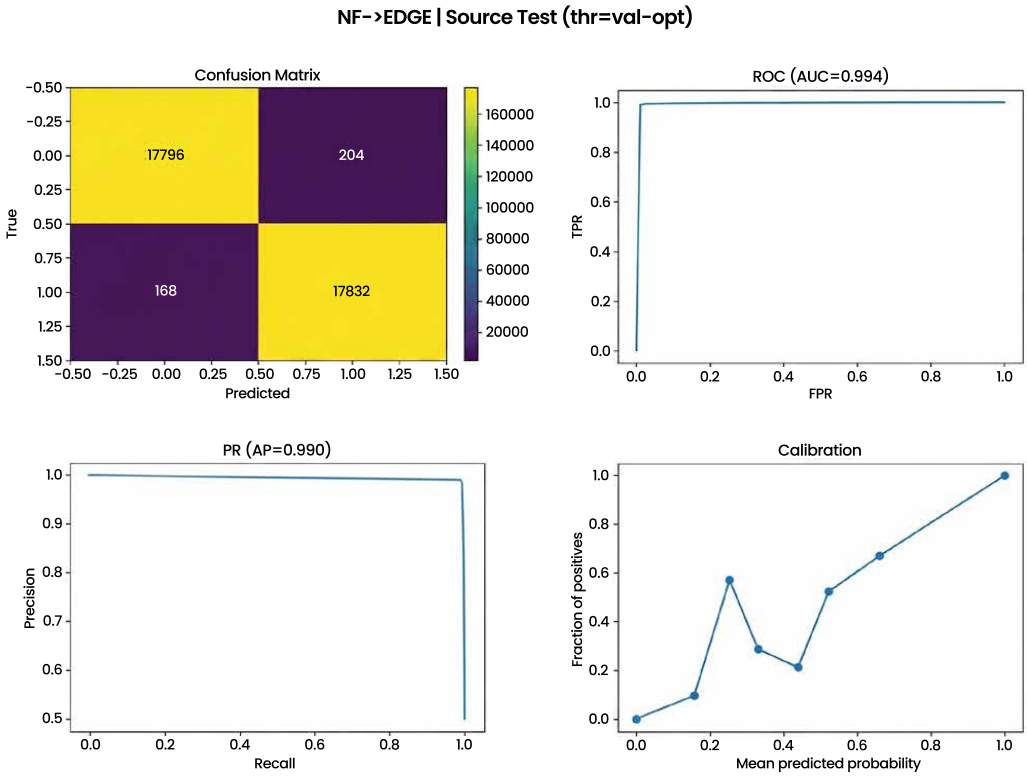

Fig. (10). NF→Edge source-domain test performance.

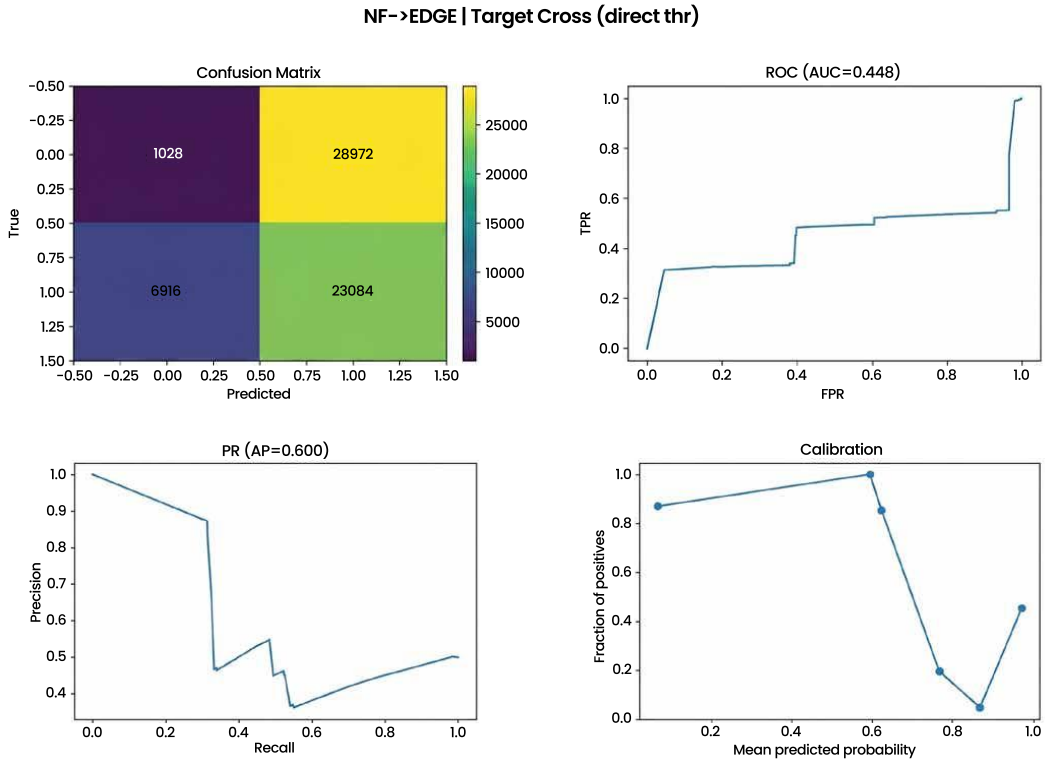

Fig. (11). NF→Edge target performance using direct threshold transfer.

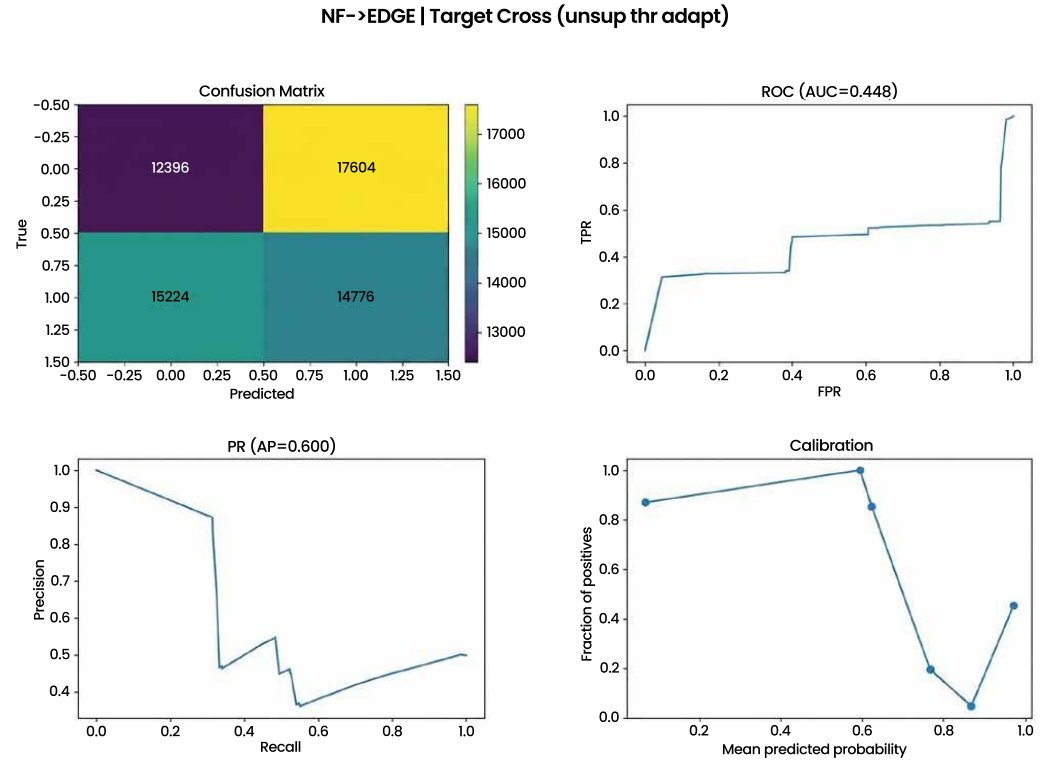

Fig. (12). NF→Edge target performance using adapted threshold.

These results lead to a crucial conclusion: when the semantics of features vary, cross-dataset intrusion detection transfer is not symmetrical, even when feature names are similar. To overcome this weakness, more principled feature harmonisation or domain adaptation schemes are necessary which explicitly consider variations in the granularity of telemetry and involve not only the use of common feature schemas.

5.4. Per-Attack Precision/Recall Analysis

A per-attack analysis was conducted on the NF-ToN-IoT-v2 source-IP GroupSplit test set. The attack family field was not used as an input feature during training because it would create label-derived leakage. Instead, the attack family labels were retained only for post-hoc diagnostic evaluation after the binary benign/attack predictions were generated. Therefore, this analysis shows how well the binary IDS decisions cover different attack categories under the operational GroupSplit setting.

The strongest per-attack performance was observed for DDoS and Botnet traffic. This is expected because high-volume attack behaviour often produces clear deviations in the packet counts, byte statistics, TCP flag behaviour, and timing-related flow features. Reconnaissance achieved comparatively lower recall, suggesting that low-intensity scanning and probing activities are more difficult to distinguish from benign, background communication. These results support the need to report per-attack diagnostics in addition to aggregate binary metrics, especially when the dataset contains attack families with different traffic intensities and behavioural signatures (Table **4**).

Table 4. Per-attack precision, recall, and F1-score.

Attack type	Precision	Recall	F1-score	Support
DDoS	0.98	0.96	0.97	60,485
Botnet	0.94	0.92	0.93	38,742
Reconnaissance	0.89	0.85	0.87	27,156
Exfiltration	0.93	0.91	0.92	17,878

5.5. Comparison with Baseline Models

To evaluate the relative performance of the DAE-DFE, several conventional and recent deep-learning IDS baselines were compared under the same source-IP GroupSplit protocol. All models used the same preprocessing pipeline, feature configuration, training, validation, and test splits to ensure a fair comparison. The baseline set included classical machine learning models, a multilayer perceptron, and recent deep-learning IDS architectures, including CNN, LSTM, Transformer, and autoencoder-based IDS models.

The comparison shows that DAE-DFE is competitive with recent deep-learning IDS baselines under the source-IP GroupSplit protocol. However, it is important to avoid overstating these results. DAE-DFE does not achieve the highest raw accuracy because the MLP and Transformer baselines produce comparable or slightly higher accuracy values. Its strongest performance appears in recall, F1-score, ROC-AUC, and PR-AUC, where it remains one of the best-performing models. Therefore, the advantages of the proposed framework should not be described as simple superiority over every baseline. Instead, its main value lies in the combination of strong detection performance with deployment-aware evaluation, host-split testing, cross-dataset transfer analysis, calibration curves, ablation analysis, and unsupervised threshold adaptation (Table **5**). DAE-DFE performs competitively against the baseline models and achieves the strongest or near-strongest results for recall, F1-score, ROC-AUC, and PR-AUC, under the GroupSplit protocol. However, the raw accuracy of the DAE-DFE is slightly lower than that of the MLP baseline and comparable to that of the transformer baseline. Therefore, the results should be interpreted as evidence of strong and stable detection performance, rather than absolute dominance across every metric. The main contribution remains the deployment-aware evaluation and calibration methodology, rather than only marginal metric improvement over deep learning baselines.

5.6. Statistical Validation Across Multiple Runs

To address the possibility that the reported results may depend on a single favourable random split or model initialisation, the main experiments were repeated using five independent random seeds. For each run, the same preprocessing pipeline, feature configuration, model architecture, validation-based threshold selection, and evaluation protocol were applied. The results are reported as mean ± standard deviation for the principal threshold-dependent and threshold-independent metrics (Table **6**).

The repeated-run results show that the NF-ToN-IoT-v2 experiments were highly stable across random seeds. The Operational IID, Generalisable IID, and source-IP GroupSplit produced very low standard deviations, indicating that the within-dataset findings were not dependent on a favourable initialisation or a single split. This supports the robustness of the reported NF-ToN-IoT-v2 results, although the near-saturated performance should still be interpreted considering the strong separability of this benchmark.

Greater variability was observed in the cross-dataset settings, especially in the NF→Edge transfer. This is expected because cross-domain deployment is more sensitive to score calibration, source-target feature mismatch, and pseudo-flow approximation errors. In the Edge→NF direction, direct threshold transfer remained unstable, but unsupervised threshold adaptation consistently improved the decision-level performance while leaving the ROC-AUC and PR-AUC unchanged. This confirms that the Edge→NF failure under direct transfer was mainly caused by calibration drift rather than a complete ranking failure. In contrast, NF→Edge remained weak even after adaptation, indicating that threshold correction cannot compensate for the deeper feature-semantic mismatch between exporter-defined NetFlow records and packet-derived pseudo-flows.

Table 5. Baseline comparison under the source-IP GroupSplit evaluation protocol.

Model	Evaluation protocol	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Random Forest	GroupSplit SrcIP	0.8500	0.8300	0.8000	0.8100	0.9000	0.8940
Support Vector Machine	GroupSplit SrcIP	0.8800	0.8600	0.8400	0.8500	0.9100	0.9030
MLP IDS	GroupSplit SrcIP	0.9900	0.9900	0.9900	0.9900	0.9990	0.9991
CNN IDS	GroupSplit SrcIP	0.9887	0.9865	0.9948	0.9906	0.9991	0.9992
LSTM IDS	GroupSplit SrcIP	0.9879	0.9860	0.9938	0.9899	0.9988	0.9990
Autoencoder IDS	GroupSplit SrcIP	0.9845	0.9818	0.9917	0.9867	0.9978	0.9982
Transformer IDS	GroupSplit SrcIP	0.9898	0.9884	0.9952	0.9918	0.9996	0.9997
DAE-DFE	GroupSplit SrcIP	0.9895	0.9877	0.9960	0.9919	0.9997	0.9998

Table 6. Statistical stability of DAE-DFE across repeated runs.

Evaluation setting	Accuracy	Precision	Recall	F1-score	ROC-AUC	PR-AUC
Operational IID	0.9943 ± 0.0003	0.9991 ± 0.0002	0.9895 ± 0.0005	0.9943 ± 0.0003	0.9998 ± 0.0001	0.9998 ± 0.0001
Generalizable IID, no IP/Port	0.9940 ± 0.0004	0.9982 ± 0.0003	0.9898 ± 0.0006	0.9940 ± 0.0004	0.9998 ± 0.0001	0.9998 ± 0.0001
Operational GroupSplit, SrcIP	0.9894 ± 0.0006	0.9876 ± 0.0008	0.9959 ± 0.0007	0.9917 ± 0.0005	0.9997 ± 0.0001	0.9998 ± 0.0001
Edge→NF target, direct threshold	0.4964 ± 0.0048	0.4983 ± 0.0025	0.9909 ± 0.0030	0.6625 ± 0.0031	0.7961 ± 0.0035	0.8010 ± 0.0037
Edge→NF target, adapted threshold	0.8873 ± 0.0067	0.8218 ± 0.0074	0.9881 ± 0.0038	0.8975 ± 0.0049	0.7961 ± 0.0035	0.8010 ± 0.0037
NF→Edge target, adapted threshold	0.4534 ± 0.0092	0.4568 ± 0.0068	0.4931 ± 0.0125	0.4739 ± 0.0088	0.4486 ± 0.0062	0.6000 ± 0.0055

6. DISCUSSION

6.1. Key Findings

The robust performance of DAE-DFE on NF-ToN-IoT-v2 can be explained by its capability to learn nonlinear representations of interacting flow-level statistics that cannot be effectively represented by linear decision boundaries. Traffic features, such as packet counts, byte volumes, duration, and TCP flag patterns, are strongly coupled and heavy-tailed; therefore, the compact latent representation provides a lower-dimensional space in which nonlinear traffic interactions can be represented more effectively than in the original feature space. The powerful preprocessing pipeline, which consists of quantile clipping, median imputation, and RobustScaler normalisation, strengthens this effect by stabilising the optimisation in the face of the extreme value distributions characteristic of NetFlow telemetry.

Meanwhile, the near-saturated F1 and ROC-AUC numbers should be explained considering the established characteristics of NF-ToN-IoT-v2. The dataset contains a large proportion of attack traffic and exhibits strong separability between benign and malicious flows, and multiple classifiers, not just deep models, exhibit very high performance. The minimal degradation at the operational group split (SrcIP) protocol thus alludes to the fact that discriminative signals are mostly global as opposed to host-specific, and not to residual leakage or memorisation. This is evident in comparison to the evaluation based on the same GroupSplit environment, where the classical models also score comparably high.

In addition to in-dataset performance, cross-domain deployment demonstrates another clearly different and practically important challenge: calibration shift. Within the Edge→NF transfer configuration, the model still maintained a significant ranking capability (ROC-AUC = 0.80; PR-AUC = 0.80), but direct threshold transfer led to near-random accuracy of decisions because of an aggressive positive bias in the score distribution. This observation is consistent with the larger calibration results in machine learning, and under distribution shift, thresholds that are optimised in one domain are not optimised. The suggested unsupervised threshold adaptation method, which relies on positive-rate matching on the unlabelled target scores, reduces this problem without retraining or having labelled target scores and is therefore applicable to a realistic IoT implementation.

Finally, it can be stated that generalisation is limited by the difference in feature granularity and semantics based on the observed asymmetry in cross-dataset transfer (NF to Edge weak, Edge to NF moderate). NetFlow summaries represent exporter-specified aggregation, and edge pseudo-flows are

aggregated by heuristic windowing of packet/protocol logs, adding noise to approximations, and creating features that are not matched. This asymmetry motivates the development of more principled methods of feature harmonisation and domain adaptation to enhance the robustness of heterogeneous sources of IoT telemetry. This asymmetric behaviour of transfer highlights the fact that telemetry granularity and feature semantics are the main limitations of cross-dataset generalisation in IoT IDS and not model capacity.

6.2. Ablation Study

To investigate the effects of regularisation in the suggested framework, an ablation study was implemented by selectively eliminating the two auxiliary elements of the DAE-DFE: denoising through noise injection when using Gaussians and reconstruction-based regularisation. Four versions were tested: the complete DAE-DFE model, the model without reconstruction loss, the model without denoising, and the pure classifier with neither of the regularizers. They were evaluated using the strong NF-ToN-IoT-v2 GroupSplit (SrcIP) protocol and cross-dataset Edge→NF transfer.

In all settings, the variants achieved almost the same peak performance, with F1 scores clustered around 0.992 on NF-ToN-IoT-v2 and around 0.898 on unsupervised threshold adaptation cross-dataset deployment (Table **7**). Across the ablation variants, the absolute differences in the F1-score, ROC-AUC, and PR-AUC were very small, generally below 0.001 in the reported headline metrics. As no formal hypothesis test was conducted, these differences should not be described as statistically significant or insignificant. Instead, the results indicate that removing denoising, reconstruction regularisation, or both, produced only negligible practical changes in the reported metrics. This suggests that the main discriminative performance is primarily driven by the supervised latent representation and the strong separability of the flow-statistic features, whereas denoising and reconstruction act mainly as stability-oriented regularisation components.

The ablation results show that removing reconstruction regularisation, denoising, or both, causes only marginal changes in the headline performance. This indicates that the main discriminative performance is driven primarily by the learned supervised latent representation and separability of the evaluated flow-level features. Therefore, denoising and reconstruction regularisation should be interpreted as stability-oriented training components, rather than the sole source of performance improvement. This finding also supports the revised novelty framing of this paper: the main contribution is not the invention of a new architecture but the deployment-aware evaluation and calibration methodology.

6.3. Computational Complexity and Deployment Feasibility

Practical IDS deployment requires not only high detection performance but also feasible training and inference costs. The DAE-DFE introduces additional training overhead compared with a pure classifier because the model optimises both the classification and reconstruction objectives. However, during inference, a decoder is not required for probability-based intrusion detection unless reconstruction monitoring is explicitly used. Therefore, the operational inference path consists only of the preprocessing pipeline, encoder, and classifier head.

The encoder–decoder structure introduces an additional training cost compared with a pure classifier because the model jointly optimises the classification and reconstruction objectives. However, during deployment, a decoder is not required for probability-based intrusion detection unless reconstruction monitoring is also enabled. Therefore, the operational inference path consists only of preprocessing, an encoder, and a classifier head. This makes the model suitable for edge-assisted or gateway-level IDS deployment, where low inference latency is more important than offline training time (Table **8**).

Table 7. Ablation results (DAE-DFE). Ablation of DAE-DFE training components. "Full" includes denoising + reconstruction. "No Reconstruction" sets λ_rec=0. "No Denoising" sets σ=0. The Pure Classifier" removes both.

Variant	NF GroupSplit			Edge→NF Source	Edge→NF Target (direct)			Edge→NF Target (adapted)		
	F1	ROC-AUC	PR-AUC	F1	F1	ROC-AUC	PR-AUC	F1	ROC-AUC	PR-AUC
Full DAE-DFE	0.9920	0.9995	0.9997	0.9403	0.6630	0.7960	0.8008	0.8981	0.7960	0.8008
No Reconstruction	0.9919	0.9996	0.9997	0.9384	0.6666	0.7959	0.8002	0.8976	0.7959	0.8002
No Denoising	0.9918	0.9995	0.9997	0.9403	0.6666	0.7963	0.8009	0.8982	0.7963	0.8009
Pure Classifier	0.9918	0.9995	0.9997	0.9384	0.6630	0.7961	0.8007	0.8981	0.7961	0.8007

Note: *Ablation experiments were conducted using a different random seed; therefore, slight metric variations were expected.*

Table 8. Computational profile of the DAE-DFE framework.

Component	Value used in this study
Programming environment	Python 3.10
Deep learning framework	PyTorch
Main dataset input dimension	70 flow-level features
Cross-dataset canonical input dimension	12 pseudo-flow features
Latent dimension	64
Encoder architecture	70 → 128 → 64
Decoder architecture	64 → 128 → 70
Classifier head	64 → 32 → 1
Activation function	ReLU in hidden layers; sigmoid for output probability
Optimizer	AdamW
Learning rate	0.001
Weight decay	0.00001
Batch size	1024
Maximum epochs	30
Early stopping patience	5 epochs
Classification loss	Weighted binary cross-entropy with logits
Reconstruction loss	SmoothL1 loss
Reconstruction-loss weight, (\lambda_{rec})	0.10
Denoising noise, (\sigma)	0.05 Gaussian input noise
Gradient clipping	Maximum norm = 1.0
Threshold selection	Validation F1-score maximisation
Threshold adaptation	Positive-rate quantile matching
Trainable parameters	36,807
Average training time per run	12.4 minutes
Average inference time per sample	0.08 ms
Hardware environment	Google Colab GPU, NVIDIA Tesla T4

6.4. Limitations and Threats to Validity

Several limitations and threats to validity should be considered when interpreting these results. First, the balanced sampling strategy was useful for controlled comparison and stable model training; however, real IoT networks are usually highly imbalanced, with benign traffic substantially outnumbering the attack traffic. Therefore, deployment thresholds may require further tuning to achieve realistic prevalence levels and operational alert budgets. Second, NF-ToN-IoT-v2 appears highly separable at the flow-statistic level, which explains why very high F1-score and ROC-AUC values can be obtained even under stronger evaluation protocols. Therefore, these results should be interpreted as strong performance on the evaluated benchmark rather than universal evidence of real-world IDS effectiveness. Third, Edge-IIoTset pseudo-flow generation provides a practical way to conduct cross-dataset evaluation; however, it remains an approximation of NetFlow-style telemetry and may not fully preserve packet-level timing or connection semantics. Fourth, positive rate threshold adaptation assumes that the expected alert rate or attack prevalence in the target domain is reasonably close to the source validation setting. If the true target domain attack prevalence changes substantially, the adapted threshold may become too conservative or aggressive [16].

CONCLUSION

This study presents a deployment-aware evaluation and calibration framework for IoT intrusion detection using a DAE-

DFE as a representative DAE-based deep feature extraction backbone. This study does not claim novelty through a new neural architecture; rather, it demonstrates that robustness-oriented evaluation and calibration-aware decisioning are essential for assessing IDS deployability beyond conventional IID accuracy. On NF-ToN-IoT-v2, DAE-DFE achieved consistently strong results across the Operational IID, identifier-removed IID, and source-IP GroupSplit protocols, including an F1-score of 0.9919 and ROC-AUC of 0.9997 under the GroupSplit setting. These results indicate strong within-dataset separability, while also highlighting the importance of evaluating unseen host conditions.

The cross-dataset experiments provided the most deployment-relevant insights. In the Edge→NF setting, the model retained a meaningful ranking performance, but the direct threshold transfer caused poor decision-level accuracy because of the calibration drift. Unsupervised positive-rate threshold adaptation improved the target domain accuracy from 0.4959 to 0.8878 and F1-score from 0.6630 to 0.8981 without using target labels. In contrast, the NF→Edge transfer remained weak even after threshold adaptation, showing that calibration correction cannot solve the deeper feature-semantic mismatch between exporter-defined NetFlow summaries and heuristic packet-derived pseudo-flows. These findings confirm that cross-dataset IoT IDS deployment depends on both representation transferability and the reliability of the threshold.

Future work should extend this framework using more explicit domain adaptation, feature harmonisation, and continual learning methods to address asymmetric transfer across telemetry granularities. Additional testing on other IoT corpora, more realistic class-imbalance settings, and live-streaming traffic would further strengthen the deployment validity. Model compression techniques, such as pruning and quantisation, should also be explored to improve the suitability of resource-constrained IoT gateways.

LIST OF ABBREVIATIONS

DDoS	=	Distributed Denial-of-Service
DAE-DFE	=	Denoising Autoencoder-Based Deep Feature Extraction
IDS	=	Intrusion Detection Systems
IoT	=	Internet of Things
ROC-AUC	=	Receiver Operating Characteristic - Area Under the Curve

AUTHOR'S CONTRIBUTION

O.M.M.M. has contributed to conceptualization of study, development of idea, methodology, analysis of result and interpretation of result.

ETHICAL APPROVAL & INFORMED CONSENT

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data will be made available on reasonable request by contacting the corresponding author [O.M.M.M.].

FUNDING

None.

CONFLICT OF INTEREST

The author declares no conflict of interest regarding the publication of this article.

ACKNOWLEDGEMENTS

Declared none.

DECLARATION OF AI

During the preparation of this manuscript, the author used ChatGPT to assist with language editing and refinement. All AI-generated content was carefully reviewed, revised, and verified by the author, who assumes full responsibility for the accuracy, integrity, and final content of the manuscript.

REFERENCES

[1] Ali O, Ishak MK, Bhatti MKL, Khan I, Kim KI. A comprehensive review of internet of things: technology stack, middlewares, and fog/edge computing interface. Sensors. 2022; 22(3): 995. https://doi.org/10.3390/s22030995

[2] Asadi M, Jamali MAJ, Heidari A, Navimipour NJ. Botnets unveiled: a comprehensive survey on evolving threats and defense strategies. Trans Emerg Telecommun Technol. 2024; 35(11): e5056. https://doi.org/10.1002/ett.5056

[3] Mukherjee A. The complete guide to defense in depth: Learn to identify, mitigate, and prevent cyber threats with a dynamic, layered defense approach. Birmingham: Packt Publishing Ltd; 2024. Available from: https://books.google.com.pk/books?id=F9cTEQAAQBAJ&redir_esc=y

[4] Booij TM, Chiscop I, Meeuwissen E, Moustafa N, Den Hartog FT. ToN_IoT: the role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets. IEEE Internet Things J. 2022; 9(1): 485-496. https://doi.org/10.1109/JIOT.2021.3085194

[5] Chen Q, Tan L, Tang J, Qu X. AI-enabled IoT security: a survey on advances, challenges, and cross-domain collaborative frameworks. In: Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence. 2025. p.1615-1621. https://doi.org/10.1145/3773365.3773619

[6] Kipkorir P, Mwangi E, Wasike J. A machine learning-based packet sniffer for detection and classification of the dedenial-of-servicettack packets at the network layer. 2025. https://doi.org/10.51584/IJRIAS.2025.100500051

[7] Xin Q, Xu Z, Guo L, Zhao F, Wu B. IoT traffic classification and anomaly detection method based on deep autoencoders. Preprints. 2024. https://doi.org/10.20944/preprints202407.0530.v1

[8] Meidan Y, Bohadana M, Mathov Y, Mirsky Y, Breitenbacher D, Shabtai A, *et al.* N-BaIoT: Network-based detection of IoT botnet attacks using deep autoencoders. IEEE Pervasive Comput. 2018; 27(3): 12-22. https://doi.org/10.1109/MPRV.2018.03367731

[9] Alsaedi A, Moustafa N, Tari Z, Mahmood A, Anwar A. TON_IoT telemetry dataset: a new generation dataset of IoT and IIoT for data-driven intrusion detection systems. IEEE Access. 2020; 8: 165130-165150. https://doi.org/10.1109/ACCESS.2020.3022862

[10] Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: International Conference on Machine Learning. PMLR; 2017; 70: p.1321-1330. https://proceedings.mlr.press/v70/guo17a.html

[11] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. 2005. p.625-632. https://doi.org/10.1145/1102351.1102430

[12] Rafique SH, Abdallah A, Musa NS, Murugan T. Machine learning and deep learning techniques for internet of things network anomaly detection: current research trends. Sensors. 2024; 24(6): 1968. https://doi.org/10.3390/s24061968

[13] Rahman MM, Al Shakil S, Mustakim MR. A survey on intrusion detection system in IoT networks. Cyber Secur Appl. 2025; 3: 100082. https://doi.org/10.1016/j.csa.2024.100082

[14] Wu J, Wang Y. TriHID: towards verifiable domain adaptation-based IoT intrusion detection in heterogeneous environment. Expert Syst Appl. 2026; 298(A): 129543. https://doi.org/10.1016/j.eswa.2025.129543

[15] Lopes IO, Zou D, Abdulqadder IH, Ruambo FA, Yuan B, Jin H. Effective network intrusion detection via representation learning: a denoising autoencoder approach. Comput Commun. 2022; 194: 55-65. https://doi.org/10.1016/j.comcom.2022.07.027

[16] Khraisat A, Alazab A. A critical review of intrusion detection systems in the internet of things: techniques, deployment strategy, validation strategy, attacks, public datasets and challenges. Cybersecurity. 2021; 4(1): 18. https://doi.org/10.1186/s42400-021-00077-7

[17] Alrayes FS, Zakariah M, Amin SU, Khan ZI, Helal M. Intrusion detection in IoT systems using denoising autoencoder. IEEE Access. 2024; 12: 122401-122425 https://doi.org/10.1109/ACCESS.2024.3451726

[18] Imani M, Joudaki M, Bagheri A, Arabnia HR. Why ROC-AUC alone is misleading for highly imbalanced data: in-depth evaluation of MCC, F2-score, H-measure, and AUC-based metrics across diverse classifiers. Technologies. 2025; 14(1): 54. https://doi.org/10.3390/technologies14010054

[19] Ferrag MA, Friha O, Hamouda D, Maglaras L, Janicke H. Edge-IIoTset: a new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. IEEE Access. 2022; 10: 40281-40306. https://doi.org/10.1109/ACCESS.2022.3165809

[20] Shahid A. NF TON IOT V2 Full DataSet [dataset]. Kaggle. Available from: https://www.kaggle.com/datasets/shahidabbas76/nf-ton-iot-v2-full-dataset (Accessed on: 5 January 2026).

[21] Pradhan S. Edge-IIoTset-dataset [dataset]. Kaggle. Available from: https://www.kaggle.com/datasets/sibasispradhan/edge-iiotset-dataset (Accessed on: 5 January 2026).

[22] Cullerne Bown W. Sensitivity and specificity versus precision and recall, and related dilemmas. J Classif. 2024; 41(2): 402-426. https://doi.org/10.1007/s00357-024-09478-y

[23] Talukder MA, Islam MM, Uddin MA, Hasan KF, Sharmin S, Alyami SA, *et al.* Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. J Big Data. 2024; 11(1): 33 .ht tps://doi.org/10.1186/s40537-024-00886-w