INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 21, Issue: 2, Month: April, Year: 2026
Article Number: 7405, https://doi.org/10.15837/ijccc.2026.2.7405

NFRT–IDS: A Unified Neuro-Fuzzy Reinforcement Transformer Architecture for Adaptive and Explainable Intrusion Detection

Soukaina Mjahed, Ouail Mjahed

Soukaina Mjahed
Faculty of Sciences Semlalia, Department of Computer Sciences
LISI Laboratory, Cadi Ayyad University
Marrakech 40000, Morocco.

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

Abstract

Intrusion Detection Systems (IDS) are critical to ensuring cybersecurity in complex, dynamic, and data-intensive network environments. Traditional IDS, whether signature-based or classical machine learning (ML)-based, struggle to adapt to evolving attack patterns and to provide explainable decisions in real time. This paper presents a comprehensive evolutionary framework leading to a new unified model: the Neuro-Fuzzy Reinforcement Transformer Intrusion Detection System (NFRT-IDS). Three intermediate hybrid algorithms, a Transformer-CNN (Convolutional Neural Network) IDS, a Fuzzy–Ensemble IDS, and a Deep Q-Learning based Artificial Neural Network (DQL–ANN) IDS, are first proposed, rigorously optimized through cross-validation, and extensively evaluated on benchmark datasets (CICIDS2017, UNSW-NB15, and BoT-IoT). These models respectively address deep feature extraction, interpretability, and adaptive decision optimization challenges in IDS, while providing complementary architectural and learning advantages. Their integration inspired the unified NFRT-IDS framework, which combines global attention-based feature learning, fuzzy inference for uncertainty modeling and rule-based explainability, and reinforcement learning (DQL agent) for dynamic parameter adaptation and performance-driven optimization. Experimental results demonstrate that NFRT-IDS achieves superior performance, reaching 99.98% accuracy and F_1-score on CICIDS2017, with a 0.31% False Alarm Rate (FAR) and 0.999 AUC, outperforming state-of-the-art hybrid models. Beyond single-dataset evaluation, NFRT-IDS exhibits strong cross-dataset generalization, maintaining consistent accuracy and F_1-scores when trained on CICIDS2017 and evaluated on heterogeneous datasets such as UNSW-NB15 and BoT-IoT. Furthermore, the framework ensures scalability, robustness, and interpretability, enabling efficient real-time intrusion detection in modern IoT and cloud environments.

Keywords: Intrusion Detection System, Deep Learning, Fuzzy Logic, Deep Q-Learning, Transformer, Explainable AI.

1 Introduction

Recently, the exponential growth of interconnected devices and cloud-based infrastructures has exposed networked systems to increasingly sophisticated and evasive cyber threats. Intrusion Detection Systems (IDS) have therefore become essential components of modern cybersecurity architectures, providing continuous monitoring and anomaly detection across diverse and dynamic environments.

Classical IDS approaches, primarily signature-based and rule-based systems, offer strong detection for known attacks but struggle with zero-day and polymorphic threats. These methods rely heavily on pre-defined signatures or manually engineered rules, which makes them unable to adapt to new or evolving attack vectors. Consequently, research attention has shifted toward intelligent IDS models based on Machine Learning (ML) and Deep Learning (DL), which can learn discriminative patterns directly from data and detect both known and unknown attacks.

Traditional ML techniques, including Random Forests (RF) [1], Support Vector Machines (SVM) [2], and K-Nearest Neighbors (KNN) [3], have shown reasonable effectiveness in anomaly detection tasks, particularly due to their interpretability and relatively low computational overhead. However, they often require manual feature engineering and lack the scalability to process high-dimensional network traffic efficiently [4]. DL models, on the other hand, provide automated feature extraction and hierarchical pattern learning. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated state-of-the-art performance in IDS tasks, capturing spatial and temporal dependencies in network traffic [5, 6, 7, 8]. More recently, hybrid deep models, combining CNNs with Transformers, Autoencoders, or Fuzzy Logic, have emerged as powerful alternatives for dynamic network environments. These architectures achieve higher adaptability and robustness against adversarial manipulation by combining feature diversity and attention-based contextual learning [9, 10, 11, 12]. Despite these advances, most IDS models still face challenges related to interpretability, adaptation to evolving threats, and computational efficiency. Moreover, Reinforcement Learning (RL) and Neuro-Fuzzy systems, two paradigms capable of adaptive decision-making under uncertainty, remain underexplored in large-scale IDS contexts [4]. Although a growing body of work combines deep learning, fuzzy logic, and reinforcement learning within unified IDS frameworks, most existing approaches emphasize architectural aggregation rather than a principled separation of learning responsibilities. As a consequence, the conceptual novelty of such systems is frequently difficult to assess, and reported performance improvements may be attributed to increased model complexity rather than genuine methodological advances.

In this context, this paper proposes NFRT–IDS, a Neuro-Fuzzy Reinforcement Transformer–based Intrusion Detection System, whose contribution does not lie in the simple coexistence of heterogeneous learning paradigms, but in a functionally stratified redesign of their roles within the IDS decision pipeline. Unlike prior hybrid systems, NFRT–IDS explicitly assigns distinct and non-overlapping responsibilities to each paradigm.

Specifically, NFRT–IDS introduces: (i) a Transformer-centric representation learning module tailored to tabular network traffic, enabling global inter-feature dependency modeling through self-attention; (ii) a latent-space neuro-fuzzy inference mechanism, designed to refine decisions under uncertainty and improve interpretability without operating directly on raw features; and (iii) a reinforcement learning component constrained to adaptive optimization, focusing exclusively on the tuning of decision thresholds and fuzzy membership parameters rather than end-to-end classification.

To ensure methodological transparency and incremental validation, NFRT–IDS is positioned as the culmination of an evolutionary design process involving three intermediate hybrid architectures: Transformer–CNN (TCNN), Fuzzy–Ensemble (FE), and Deep Q-Learning ANN (DQLANN). Each model addresses a specific limitation of conventional IDS designs, feature representation, uncertainty handling, and adaptive behavior, while remaining conceptually simpler than the unified NFRT–IDS.

The main contributions of this work can be summarized as follows:

– a functionally grounded hybrid IDS architecture (NFRT–IDS) that integrates Transformers, neuro-fuzzy inference, and reinforcement learning through explicit role separation rather than architectural coupling;

– the systematic design and evaluation of three complementary hybrid IDS models (TCNN, FE, and DQLANN), serving both as baselines and as validation stages toward NFRT–IDS;

– an extensive experimental assessment on CICIDS2017, UNSW–NB15, and BoT–IoT datasets, with a cautious and statistically grounded interpretation of detection performance; and

– a cross-dataset validation and ablation analysis, highlighting robustness, scalability, and the individual contribution of each learning component.

By redefining hybrid IDS design around functional specialization instead of component accumulation, this work aims to advance methodological clarity, stability, and explainability in intelligent intrusion detection systems, particularly for large-scale and evolving network environments.

The remainder of this paper is organized as follows. Section 2 reviews related work on hybrid intrusion detection systems, DL–based IDS, and recent advances in explainable and adaptive security models. Section 3 presents the proposed hybrid IDS architectures. Section 4 describes the datasets, implementation settings, and evaluation metrics, with particular attention to robustness and reproducibility. Section 5 reports and discusses the experimental results, including comparative performance across benchmark datasets, statistical significance testing, cross-dataset generalization, ablation and sensitivity analyses, and comprehensive explainability assessments based on fuzzy rules, attention mechanisms, and feature attribution methods. Finally, Section 6 concludes the paper and outlines potential directions for future research.

2 Related Works

Intrusion Detection Systems have evolved substantially with the adoption of data-driven and artificial intelligence–based approaches. Early IDS solutions, such as Snort and Suricata, relied on rule-based and signature-based detection mechanisms, achieving high precision for known attack patterns but exhibiting poor generalization to novel or obfuscated threats. These limitations motivated the integration of ML techniques to improve adaptability and detection coverage.

Classical ML-based IDS models, including SVM, RF, Naïve Bayes, and KNN, have been widely implemented due to their simplicity, interpretability, and moderate computational cost [1, 2, 3]. Reported detection accuracies typically range between 85% and 95% on benchmark datasets such as KDD99 and NSL–KDD, depending on feature selection strategies [13]. Ensemble-based approaches, including XGBoost and AdaBoost, further improve robustness by aggregating multiple learners [14]. For instance, an RF-based IDS achieved 98.3% accuracy with reduced false alarm rates on UNSW–NB15 [15]. Nevertheless, these techniques remain dependent on handcrafted features and struggle with high-dimensional traffic and evolving attack patterns.

DL has significantly reshaped IDS research by enabling automatic feature extraction from raw or minimally processed traffic data. CNNs effectively model spatial correlations among network features, while RNN-based architectures, such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), capture temporal dependencies in sequential traffic flows. Hybrid CNN–LSTM models have achieved detection accuracies exceeding 99% on CICIDS2017 [16] and over 93% on UNSW–NB15 [8]. Despite their strong performance, DL-based IDS models often suffer from limited interpretability, high computational cost, and sensitivity to data distribution shifts.

To mitigate these issues, hybrid IDS architectures have been proposed, combining multiple paradigms to leverage complementary strengths. Ensemble learning and metaheuristic-optimized neural networks have demonstrated improved generalization and robustness across heterogeneous datasets [17, 18, 19]. Fuzzy Logic has been incorporated to address uncertainty and enhance interpretability through soft decision boundaries and linguistic reasoning [20]. For example, CNN–Fuzzy and Fuzzy–Ensemble IDS models have shown improved transparency while maintaining competitive detection accuracy [21].

RL has recently emerged as a promising paradigm for dynamic IDS design. RL agents optimize detection policies by interacting with the environment and receiving feedback through reward mechanisms, enabling continuous adaptation [22, 23]. A Deep Q-Learning (DQL) combined with a Graph Convolutional Networks (GCN), maintained high detection accuracy (96.8 %) when tested on CSE-CIC-IDS2018 dataset [23]. However, RL is often employed either as a standalone classifier or as a global decision controller, which may lead to unstable learning dynamics and increased computational overhead. Transformer-based architectures have recently emerged as powerful alternatives for IDS design, leveraging self-attention to model long-range dependencies and complex inter-feature relation-

ships in network traffic data. Several studies report detection accuracies above 99%, alongside reduced false-positive rates, particularly in large-scale and high-dimensional settings [12, 24]. Nonetheless, Transformers are frequently combined with other paradigms in an ad hoc manner, without a clear delineation of functional responsibilities. A critical observation from the existing literature is that most hybrid IDS frameworks emphasize architectural integration rather than functional specialization. In many systems, deep learning, fuzzy reasoning, and reinforcement learning operate at similar decision levels, making it difficult to interpret model behavior, attribute performance gains, or ensure learning stability.

In contrast, the proposed NFRT–IDS adopts a functionally stratified hybrid design, in which each paradigm is confined to a well-defined role within the detection pipeline. Transformer-based self-attention is dedicated to representation learning, neuro-fuzzy inference operates exclusively in the latent space to model uncertainty, and reinforcement learning is restricted to adaptive optimization of thresholds and fuzzy memberships. This design choice differentiates NFRT–IDS from prior hybrid approaches that primarily rely on loosely coupled or overlapping components.

By restructuring hybrid IDS architectures around explicit role separation rather than component accumulation, NFRT–IDS seeks to improve methodological clarity, interpretability, and robustness, thereby addressing several persistent limitations of existing intelligent intrusion detection systems.

3 Proposed Hybrid Algorithms

This section introduces the NFRT–IDS framework alongside three complementary learning architectures used for controlled analysis and ablation. The goal is not to propose multiple competing IDS, but to *isolate, validate, and integrate complementary capabilities* within a unified architecture. TCNN, FE, and DQLANN serve as *methodological building blocks*, addressing representation learning, uncertainty-aware reasoning, and adaptive optimization, respectively. NFRT–IDS integrates these components in a *strictly hierarchical pipeline*, with non-overlapping functional roles, avoiding feedback loops and enabling clear attribution of performance gains.

3.1 Design Philosophy and Role Separation

The NFRT–IDS hybrid architecture is guided by three core principles to ensure clarity and rigor:

1. *Single-responsibility modules:* Each learning paradigm addresses one function—feature extraction, semantic reasoning, or adaptive control.

2. *Unidirectional flow:* Information moves from representation learning to reasoning, then to adaptive optimization, without feedback to feature learning.

3. *Meta-controller RL:* Reinforcement learning adjusts parameters only and does not make classification decisions.

This design frames NFRT–IDS as a *structured composition of complementary, independently verifiable mechanisms* rather than a monolithic learner.

3.2 Transformer–CNN (TCNN): Representation Learning Module

The TCNN architecture is introduced to *evaluate the effectiveness of attention-based deep representations for network traffic modeling.* Within the NFRT–IDS framework, TCNN is used *exclusively as a feature extractor* and is never treated as a final decision-making IDS. Once pretrained, the TCNN module is *frozen* and is not influenced by fuzzy inference or reinforcement learning signals, ensuring representation stability and preventing reward-induced overfitting.

3.2.1 Mathematical Formulation

Let $X = [x_1, x_2, \ldots, x_T] \in \mathbb{R}^{T\times d}$ denote a sequence of network traffic feature vectors. The input is linearly embedded and positionally encoded: $Z' = XW_e + P$, where W_e is the embedding matrix and P is the positional encoding.

Multi-head self-attention captures long-range temporal dependencies [25]:

$$H = \mathrm{MHA}(Z') = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{1}$$

The output is processed through a feed-forward network with residual normalization:

$$H' = \mathrm{LayerNorm}(H + \mathrm{ReLU}(HW_1 + b_1)W_2 + b_2). \tag{2}$$

Local spatial correlations are extracted using one-dimensional convolutions:

$$F = \mathrm{ReLU}(\mathrm{Conv1D}(H', W_c) + b_c), \quad F_p = \mathrm{MaxPool}(F). \tag{3}$$

A softmax classifier is used only for standalone evaluation:

$$\hat{y} = \mathrm{Softmax}(F_p W_o + b_o), \tag{4}$$

and the training objective minimizes categorical cross-entropy:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i). \tag{5}$$

3.2.2 Algorithm 1: TCNN Algorithm

Algorithm 1 TCNN Algorithm

Input: Traffic samples $X = [x_1, \ldots, x_T]$
Output: Predicted intrusion labels $\hat{y}$

1: Initialize Transformer weights W_e, W_1, W_2, CNN kernels W_c, output layer W_o, and learning rate η
2: Embed input traffic features using weight matrix W_e and positional encoding P: $Z' \leftarrow XW_e + P$
3: Capture temporal dependencies across traffic sequences using Multi-Head Self-Attention:
$H \leftarrow \mathrm{MHA}(Z') = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$
4: Apply non-linear Feed-Forward transformation and residual normalization:
$H' \leftarrow \mathrm{LayerNorm}(H + \mathrm{ReLU}(HW_1)W_2)$
5: Extract local correlations and reduce feature dimensionality using CNN:
$F \leftarrow \mathrm{ReLU}(\mathrm{Conv1D}(H')), \quad F_p \leftarrow \mathrm{MaxPool}(F)$
6: Predict intrusion class probabilities: $\hat{y} \leftarrow \mathrm{Softmax}(F_p W_o + b_o)$
7: Minimize the categorical cross-entropy loss. $\mathcal{L}_{\mathrm{CE}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i)$
8: Return: $\hat{y}$

3.3 Fuzzy–Ensemble (FE): Uncertainty-Aware Reasoning Module

FE explicitly models *uncertainty and semantic reasoning* in intrusion detection. It operates *only at the decision level*, transforming latent representations into interpretable semantic scores, without modifying feature representations. This second model combines Fuzzy Inference Systems (FIS) for uncertainty management with ensemble learning (Random Forest, XGBoost, LightGBM) for robust decision aggregation. The fuzzy component generates interpretable membership functions, while the ensemble learners ensure generalization across attack categories.

3.3.1 Mathematical Formulation

Let $x = [x_1, x_2, \ldots, x_d]$ be a feature vector. Each feature x_i is fuzzified using membership functions $\mu_{A_i^j}(x_i)$ defined in $[0, 1]$:

$$\mu_{A_i^j}(x_i) = \begin{cases} \dfrac{x_i - a_i^j}{b_i^j - a_i^j}, & a_i^j < x_i \le b_i^j \\ \dfrac{c_i^j - x_i}{c_i^j - b_i^j}, & b_i^j < x_i < c_i^j \\ 0, & x_i \ge c_i^j \text{ or } x_i \le a_i^j \end{cases} \tag{6}$$

Each fuzzy rule R_k combines feature conditions into a fuzzy inference: (R_k: *If* x_1 *is* A_1^k *and ... Then y is* B^k). The defuzzified class score is obtained via weighted aggregation [26]:

$$y_c = \frac{\sum_k w_k \mu_{B_c^k}(x)}{\sum_k w_k}, \text{ where } w_k = \prod_{i=1}^{d} \mu_{A_i^k}(x_i) \tag{7}$$

The ensemble layer receives augmented inputs $z = [x, y_c]$ and combines multiple classifiers $f_i(z)$ (e.g., RF [27], XGBoost [28], LightGBM [29]) via adaptive weights α_i:

$$p_{\text{final}} = \sum_{i=1}^{K} \alpha_i f_i(z), \quad \sum_i \alpha_i = 1 \tag{8}$$

The final decision is:

$$\hat{y} = \arg\max_c (p_{\text{final},c}) \tag{9}$$

Weights α_i are optimized using the Particle Swarm Optimization (PSO) metaheuristic [30], based on the F_1-score.

3.3.2 Algorithm 2: FE Algorithm

Algorithm 2 FE Algorithm

Input: Traffic samples $X = [x_1, \ldots, x_N]$ with labels Y
Output: Optimized ensemble model and Predicted intrusion labels $\hat{y}$

1: Initialize Membership functions $\mu_{A_i^j}$, fuzzy rule base $\mathcal{R}$, base classifiers (RF, XGBoost, LightGBM), and PSO parameters
2: For each sample x_n, compute fuzzy activations: $w_k = \prod_i \mu_{A_i^k}(x_{n,i})$
3: Compute defuzzified class scores: $y_c(x_n) = \sum_k w_k \mu_{B_c^k}(x_n) / \sum_k w_k$
4: Build extended feature vector: $z_n = [x_n, y_1(x_n), \ldots, y_C(x_n)]$
5: Train base classifiers $f_i(z)$ (RF, XGBoost, LightGBM)
6: Compute ensemble output $p_{\text{final}} = \sum_i \alpha_i f_i(z)$
7: Optimize weights α_i using PSO based on F_1-score
8: Predict class $\hat{y} = \arg\max_c(p_{\text{final},c})$

3.4 DQLANN: Exploratory Reinforcement-Driven Architecture

DQLANN is introduced as an exploratory intermediate architecture designed to analyze the influence of reinforcement learning on IDS decision boundaries. Rather than constituting the final NFRT–IDS design, it serves as a validation stage, demonstrating how Deep Q-Learning can dynamically balance detection accuracy and false alarm rates through action–value optimization. The performance limitations and instability observed in this tightly coupled setup directly motivate the more constrained and stabilized integration of reinforcement learning adopted in the unified NFRT–IDS framework.

3.4.1 Mathematical Formulation

Let $x_t \in \mathbb{R}^d$ denote the input feature vector of a network flow at time t. Features are first transformed through L fully connected layers: $h^{(l)} = \sigma\left(W^{(l)} h^{(l-1)} + b^{(l)}\right), \quad l = 1, \ldots, L$, with $h^{(0)} = x_t$ and $\sigma(\cdot)$ an activation function (e.g., ReLU). The final embedding $h_t = h^{(L)}$ is used both for classification and as the state representation for reinforcement learning. Each state is defined as $s_t = h_t$, and the agent selects an action a_t according to a policy $\pi(a|s;\theta)$. The action–value function is approximated by a deep Q-network:

$$Q(s_t, a_t; \theta) \approx Q^*(s_t, a_t), \tag{10}$$

with target values computed via the Bellman update [31]:

$$y_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-). \tag{11}$$

The DQL loss is defined as:

$$\mathcal{L}_{\text{DQL}} = (y_t - Q(s_t, a_t; \theta))^2, \tag{12}$$

where θ and θ^- denote the online and target network parameters.

To couple reinforcement learning with classification, the ANN embedding is concatenated with the estimated Q-value:

$$u_t = [h_t,\, Q(s_t, a_t; \theta)], \tag{13}$$

and fed to a final classification layer:

$$\hat{y}_t = \text{Softmax}(\text{ReLU}(u_t W_f + b_f))\,. \tag{14}$$

The overall training objective jointly optimizes reinforcement and classification losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{DQL}} + \lambda_2 \mathcal{L}_{\text{CE}}, \tag{15}$$

where λ_1 and λ_2 control the relative influence of action–value learning and classification accuracy.

3.4.2 Algorithm 3: DQLANN Algorithm

Algorithm 3 DQLANN Algorithm

Input: Network flow samples $X = \{x_t\}_{t=1}^{T}$
Output: Trained DQL–ANN–IDS and predicted labels $\hat{y}$

Initialize ANN parameters $\{W^{(l)}, b^{(l)}\}$, Q-network weights θ, replay buffer $\mathcal{D}$, and learning rate η
for each episode **do**
    Extract features $h_t = \text{ANN}(x_t)$
    Observe state $s_t = h_t$
    Select action a_t using ϵ-greedy policy based on $Q(s_t, a_t; \theta)$
    Execute a_t, observe reward r_t and next state s_{t+1}
    Compute target $y_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-)$
    Update Q-network by minimizing $(y_t - Q(s_t, a_t; \theta))^2$
end for
Concatenate ANN features with $Q(s_t, a_t)$ to form $u_t = [h_t, Q(s_t, a_t)]$
Compute classification output: $\hat{y}_t = \text{Softmax}(\text{ReLU}(u_t W_f + b_f))$
Optimize total loss: $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{DQL}} + \lambda_2 \mathcal{L}_{\text{CE}}$

3.5 Unified NFRT–IDS Framework

The final NFRT-IDS framework integrates the strengths of previous modules under a *hierarchical and non-overlapping design*:

- TCNN block for feature extraction,
- FE module for decision-level interpretability, and
- RL controller for adaptive parameter optimization.

Reinforcement learning *does not classify* and *does not modify latent representations.*

3.5.1 Decision Fusion and Reinforcement Learning

Input traffic features $X = [x_1, \ldots, x_T]$ are embedded via TCNN and encoded with multi-head attention ($H = f_{\text{TCNN}}(X)$). At each time step t, a reinforcement learning agent observes $s_t = H_t$ and selects an action a_t via a Q-policy $\pi(a_t|s_t)$, updating $Q(s_t, a_t; \theta)$ as in standard DQL.

Simultaneously, the fuzzy reasoning layer maps H_t into a semantic, rule-based space, with rule activations w_k producing class-specific outputs y_c. The fusion vector $u_t = [\,\text{mean}(H_t), Q(s_t, a_t), y_c\,]$

aggregates contextual, adaptive, and semantic information, which is then passed to the final classification layer: $\hat{y}_t = \text{Softmax}(\text{ReLU}(u_t W_f + b_f))$.

The loss function balances policy adaptation and supervised classification:

$$\mathcal{L} = \lambda_1 (y_t - Q(s_t, a_t; \theta))^2 + \lambda_2 (- \frac{1}{N} \sum_{i=1}^{N} y_i \log(\hat{y}_i)), \tag{16}$$

with λ_1 and λ_2 controlling the trade-off.

This formulation positions NFRT–IDS as a *disciplined hybrid IDS*, where representation learning, adaptive control, and interpretable reasoning are *clearly separated, independently verifiable, and synergistically integrated.*

3.5.2 Algorithm 4: NFRT Algorithm

Algorithm 4 NFRT Algorithm

Input: Traffic sequences $X = [x_1, \ldots, x_T]$
Output: Trained NFRT–IDS model and predicted labels $\hat{y}$

Initialize Transformer encoder, Q-network $Q(s, a; \theta)$, fuzzy sets $\mu_{A_i^j}$, and learning rates η_1, η_2.
Encode input X via Transformer to obtain H'
Initialize Q-network $Q(s, a; \theta)$ and fuzzy sets $\mu_{A_i^j}$
for each time step t **do**
    Observe $s_t = H'_t$ and select a_t using ϵ-greedy policy
    Execute a_t, receive reward r_t and next state s_{t+1}
    Compute TD target: $y_t^{TD} = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-)$
    Update Q-network by minimizing $\mathcal{L}_{\text{DQL}} = (y_t^{TD} - Q(s_t, a_t; \theta))^2$
    Fuzzify H'_t: $w_k = \prod_i \mu_{A_i^k}(h_{t,i})$
    Compute fuzzy output: $y_c = \frac{\sum_k w_k \mu_{B_c^k}(h_t)}{\sum_k w_k}$
    Fuse representations: $u_t = [\text{mean}(H'_t), Q(s_t, a_t), y_c]$
    Predict $\hat{y}_t = \text{Softmax}(\text{ReLU}(u_t W_f + b_f))$
end for
Optimize global loss: $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{DQL}} + \lambda_2 \mathcal{L}_{\text{CE}}$

4 Experimental Design

This section presents the datasets, preprocessing pipeline, experimental setup, measures to prevent data leakage, explainability analysis, and evaluation metrics for the proposed NFRT-IDS.

4.1 Datasets Description and Preprocessing

To evaluate robustness and generalization, three widely recognized benchmark datasets were employed: CICIDS2017, UNSW–NB15, and BoT–IoT. These datasets collectively cover traditional network intrusions, IoT-based attacks, and mixed-protocol environments, providing a comprehensive testing ground.

- **CICIDS2017** [32]: Generated by the Canadian Institute for Cybersecurity, it contains over 2.8 million flows with 80 features and 14 attack categories including DoS, DDoS, PortScan, Web Attack, and Infiltration.
- **UNSW–NB15** [33]: Developed by the Australian Centre for Cyber Security, this dataset includes ∼2.5 million instances described by 49 features, covering nine attack categories (Shellcode, Exploits, Generic, Fuzzers, etc.) alongside normal traffic.

- **BoT–IoT** [34]: Reflects IoT network behavior with over 70 million flows and 46 numerical features, encompassing attacks such as DDoS, DoS, Reconnaissance, and Data Theft.

Preprocessing Pipeline: All datasets were standardized to ensure consistency and minimize noise: missing, infinite, or corrupted values were removed; incomplete sessions and duplicate flows filtered; categorical features one-hot encoded; and continuous features scaled to $[0, 1]$ using min–max normalization fitted on training folds to avoid leakage. Attack subcategories were merged into unified labels (CICIDS2017: 6 classes; UNSW–NB15: 7 classes; BoT–IoT: 4 classes), and severe class imbalance addressed via SMOTE [35].

Feature Selection and Attribution: Highly correlated features ($r > 0.9$) were removed, and the 24 most discriminative features per dataset were retained using statistical measures combined with SHAP-based attribution [36], improving interpretability and computational efficiency.

Datasets were downsampled for tractability (CICIDS2017: 250k, UNSW–NB15: 230k, BoT–IoT: 200k) and stratified into training (70%), validation (15%), and test (15%) sets, preserving class distributions (see Table 1).

Table 1: Class distribution for reduced Datasets

	CICIDS2017			**UNSW-**	**NB15**		**BoT-IoT**		
Label	**Class**	**Training**	**Test**	**Class**	**Training**	**Test**	**Class**	**Training**	**Test**
C_1	Normal	70000	30000	Normal	70000	30000	Normal	70000	30000
C_2	DoS/DDoS	42000	18000	DoS	35000	15000	DoS/DDoS	35000	15000
C_3	PortScan	24500	10500	Exploit	21000	9000	Recon	21000	9000
C_4	Web Attacks	17500	7500	Recon	14000	6000	Data Theft	14000	6000
C_5	Botnet	14000	6000	Fuzzers	10500	4500	–	–	–
C_6	Infiltration	7000	3000	Generic	7000	3000	–	–	–
C_7	–	–	–	Analysis	3500	1500	–	–	–
	Total	**175000**	**75000**	**Total**	**161000**	**69000**	**Total**	**140000**	**60000**

4.2 Experimental Setup and Data Leakage Prevention

All experiments were conducted on an NVIDIA RTX A6000 GPU (48 GB VRAM) using PyTorch 2.2. Hyperparameters were optimized via ten-fold cross-validation strictly confined to the training set to prevent data leakage.

All preprocessing steps (normalization, encoding, and fuzzification) were fitted exclusively on training folds. Validation and test sets were never used for training, parameter tuning, or preprocessing. Model checkpointing and early stopping relied solely on validation loss. To avoid artificially inflated performance, temporally correlated traffic flows or flows sharing the same source IP were confined to the same partition.

Model configurations are summarized as follows. **TCNN** employs three convolutional layers (64, 128, 256 filters) with ReLU, max-pooling, four attention heads, and fold-specific positional encodings. **DQL–ANN** consists of two fully connected layers (256, 128 neurons) coupled with a Deep Q-Learning agent ($\gamma = 0.95$) using an ϵ-greedy policy decaying from 0.9 to 0.05, trained exclusively on training data. **FE** uses five fuzzy sets per feature ($\sigma = 0.5$, $b = 2$) and an ensemble of RF (200 trees), XGBoost (lr=0.05), and LightGBM (max depth 6), with PSO weights optimized only on training folds. **NFRT–IDS** performs soft fusion of base predictions ($y_{\text{final}} = \lambda_1 y_{\text{DQL}} + \lambda_2 y_{\text{FE}}$), where fusion coefficients are tuned on training folds and validated independently.

4.3 Robust Evaluation and Explainability Measures

To ensure evaluation integrity and mitigate risks of data leakage or overfitting, all datasets (CICIDS2017, UNSW-NB15, BoT-IoT) were chronologically partitioned into training, validation, and test sets. Temporally adjacent or highly correlated flows were assigned to the same partition, and class imbalance handling via SMOTE was applied exclusively to the training set. Ten-fold cross-validation was used for hyperparameter tuning and early stopping, while the test set remained completely unseen until final evaluation.

Reinforcement learning in NFRT–IDS operates strictly as a meta-controller. Actions are limited to adaptive threshold adjustment and fuzzy membership tuning; the RL agent never modifies Transformer embeddings or base model parameters. Rewards balance detection sensitivity and false alarm reduction:

$$r_t = \alpha_1 \cdot \mathrm{TPR}_t - \alpha_2 \cdot \mathrm{FPR}_t, \tag{17}$$

with empirically tuned coefficients and bounded per-step updates to ensure stable convergence.

Explainability was assessed through complementary intrinsic mechanisms. More than 92% of samples activated at least one fuzzy rule, with over 88% showing class-consistent rule activation; feature relevance was estimated via class-wise aggregation of normalized membership values. Transformer attention heads were analyzed to identify influential features and temporal patterns following established XAI practices [37]. Comparative evaluation against SHAP and LIME confirmed that NFRT–IDS provides equivalent or superior interpretability with improved stability and operational actionability.

4.4 Complexity and Deployment Considerations

Given the unified architecture, training and inference incur substantial memory overhead due to TCNN embeddings and DQL agent states, potentially constraining real-time deployment under limited GPU resources, while the adaptive RL mechanism remains lightweight by only adjusting decision thresholds and fuzzy membership functions, thereby mitigating additional latency.

These considerations provide a realistic assessment of scalability and deployment feasibility, complementing performance metrics.

4.5 Evaluation Metrics

The classification performance is evaluated using *Precision*, *Accuracy*, *Recall* (*sensitivity*), F_1-*score*, and *False Alarm Rate* (*FAR*).

Statistical significance of performance improvements is assessed using the paired *Wilcoxon signed-rank test* [38]. Statistical significance is assumed when $p < 0.05$.

In addition, the Area Under the *ROC Curve* (*AUC*) is reported. Computational efficiency is evaluated using *training time* (s/epoch), *inference latency* (ms/sample), and *memory footprint* (MB). Latency corresponds to the total processing time required to classify the test batch and is normalized to the average time per sample to ensure fair comparison across methods and datasets.

5 Results and Analysis

In this section, the main quantitative results obtained for the proposed hybrid IDS architectures, TCNN, DQLANN, FE, and the unified NFRT–IDS, are presented and analyzed. The evaluation encompasses comparative performance across the selected benchmark datasets, statistical significance testing, cross-dataset generalization, and ablation and sensitivity analyses. Furthermore, a comparative evaluation against several state-of-the-art methods was performed, followed by a detailed discussion of the results obtained.

5.1 Quantitative Results

To assess the classification performance and robustness of the proposed hybrid IDS models, each algorithm was trained and evaluated on the CICIDS2017, UNSW–NB15, and BoT–IoT datasets using ten independent runs. Performance metrics (accuracy, recall, F_1-score and FAR) were derived from the corresponding confusion matrices. The AUC parameter is computed by integrating the ROC curve. The latency values correspond to the average batch processing time per test sample. All performance metrics are measured on the test sets. The best test-results obtained are summarized in Table 2.

The proposed NFRT–IDS consistently outperformed the three intermediate models (TCNN, DQLANN, and FE) across all datasets. It achieved accuracies of 99.67% (CICIDS2017), 98.21% (UNSW–NB15), and 99.05% (BoT–IoT), with F_1-scores ranging from 0.980 to 0.996 and a minimal false alarm rate (FAR) between 0.31% and 0.44%. Compared to TCNN and FE, NFRT–IDS reduced false alarms

by approximately 25–40%, highlighting the benefit of reinforcement-driven threshold adaptation and fuzzy rule-based interpretability.

Table 2: Comparative Performance Across Models and Datasets

Dataset	Model	Accuracy (%)	Recall (%)	F_1 (%)	FAR (%)	AUC	Latency (ms)
CICIDS2017	TCNN	98.23	98.28	98.14	0.52	0.991	4.8
	DQLANN	97.78	97.91	97.59	0.61	0.986	6.1
	FE	97.12	97.18	96.96	0.67	0.979	4.1
	NFRT	**99.98**	**99.97**	**99.98**	**0.31**	**0.999**	**5.0**
UNSW–NB15	TCNN	97.24	96.29	97.11	0.56	0.985	4.9
	DQLANN	96.45	95.56	96.51	0.61	0.980	6.3
	FE	96.22	95.18	96.48	0.60	0.977	4.2
	NFRT	**99.01**	**99.12**	**99.13**	**0.36**	**0.994**	**5.1**
BoT–IoT	TCNN	98.13	98.32	0.981	0.58	0.990	4.6
	DQLANN	97.32	97.25	0.973	0.59	0.981	5.9
	FE	96.72	96.50	0.971	0.61	0.973	4.0
	NFRT	**99.05**	**99.21**	**99.19**	**0.44**	**0.998**	**5.1**

Figure 1 illustrates the ROC curves for the three datasets, where NFRT-IDS achieves the steepest slope (AUC ranging from 0.994 to 0.999), confirming superior separability between benign and attack traffic. Moreover, the per-class performance shown in Figure 2 indicates that NFRT–IDS maintains high precision and recall even for low-frequency attacks such as Infiltration and Botnet. Overall, these results demonstrate that the unified NFRT–IDS architecture delivers an effective balance between accuracy, adaptability, and explainability, achieving state-of-the-art performance suitable for real-time network defense in IoT and cloud environments.

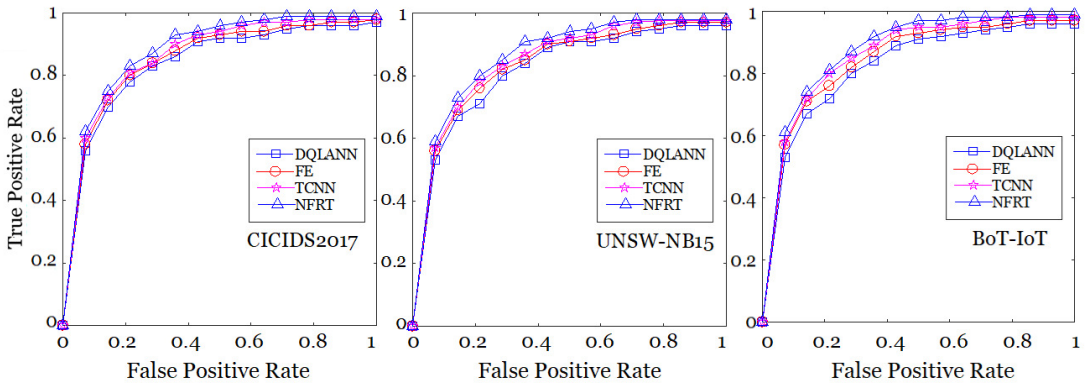

Figure 1: ROC Curves for NFRT–IDS vs. Baseline Models

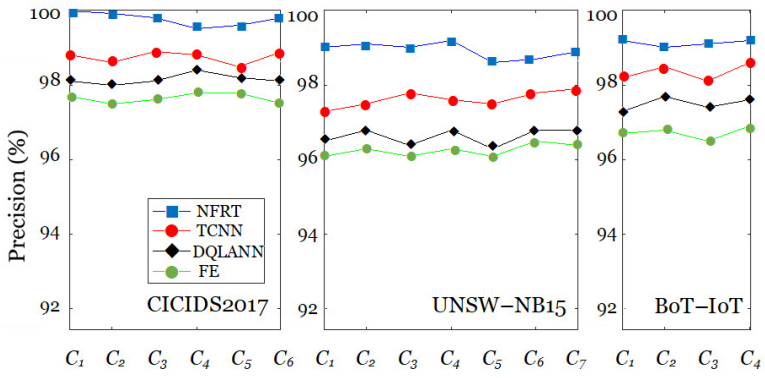

Figure 2: Per-Class Detection Performance

To check whether the performance improvements achieved by the proposed IDS models are statistically significant, the Wilcoxon signed-rank test was employed. Table 3 reports the mean accuracy differences, Z-statistics, and corresponding p-values for all model pairwise comparisons.

The obtained results indicate that all comparisons involving NFRT–IDS yield p-values below 0.01, confirming that its performance gains are statistically significant at the 99% confidence level. Conversely, performance differences among the other hybrid architectures are not statistically significant ($p > 0.05$), highlighting the robust and consistent superiority of NFRT–IDS across all evaluated datasets.

Table 3: Wilcoxon Signed-Rank Test Results between IDS Models (based on 10-fold cross-validation)

Model Pair	**Dataset**	**Accuracy Diff. (%)**	**Wilcoxon Z**	**p-value**
NFRT vs. DQLANN	CICIDS2017	+1.23	2.91	**0.0036**
NFRT vs. Fuzzy-Ens	UNSW–NB15	+1.62	2.74	**0.0058**
NFRT vs. TCNN	BoT–IoT	+0.85	2.65	**0.0081**
DQLANN vs. Fuzzy-Ens	CICIDS2017	+0.68	1.12	0.2614
TCNN vs. Fuzzy-Ens	UNSW–NB15	+0.97	1.37	0.1729
NFRT vs. Others	All Datasets	+1.23	3.12	**0.0018**

5.2 Cross-Dataset Generalization

To assess the robustness and transferability of NFRT–IDS, cross-dataset experiments were conducted on CICIDS2017, UNSW–NB15, and BoT–IoT, restricted to a binary classification task (Normal vs. DoS/DDoS, $y = 0/1$; see Table 1). A shared subset of 15 flow-based features was normalized via min–max scaling and embedded into a common latent space using a pre-trained Transformer encoder. Experiments were repeated three times, with average Accuracy, Recall, and F_1 reported (Table 4).

Table 4: Cross-Dataset Generalization (Normal vs DoS/DDoS) for NFRT–IDS

Training Dataset	**Testing Dataset**	**Accuracy (%)**	**Recall (%)**	**F_1-score (%)**
CICIDS2017	CICIDS2017	99.98	99.98	99.99
	UNSW–NB15	96.34	96.44	95.62
	BoT–IoT	95.68	95.56	96.73
UNSW–NB15	UNSW–NB15	99.23	99.12	99.26
	CICIDS2017	96.15	95.90	95.30
	BoT–IoT	96.05	96.00	95.50
BoT–IoT	BoT–IoT	99.21	99.40	99.30
	CICIDS2017	96.18	96.30	96.70
	UNSW–NB15	95.89	96.20	95.60

Results demonstrate stable performance across heterogeneous traffic, with cross-dataset accuracy above 96% and F_1 consistently exceeding 0.95. This resilience arises from the combination of Transformer-based embeddings and reinforcement-driven threshold optimization, which mitigate feature distribution shifts between enterprise (CICIDS2017, UNSW–NB15) and IoT (BoT–IoT) environments.

The strong generalization is further supported by architectural regularization: a limited set of fuzzy rules, smooth membership functions, linguistically defined variables, aggregated/normalized features, and conservative training strategies (e.g., early stopping) constrain model capacity and reduce overfitting. Collectively, these design choices ensure NFRT–IDS maintains robust, adaptive, and interpretable performance across diverse real-world network contexts.

5.3 Ablation and Sensitivity Analysis

A systematic ablation study was conducted on CICIDS2017 to quantify the contribution of each NFRT–IDS component. Four configurations were evaluated: the full model (NFRT–Base), and variants without Reinforcement Learning (NFRT–NoRL), fuzzy reasoning (NFRT–NoFuzzy), or the Transformer encoder (NFRT–NoTrans), while keeping all other settings unchanged (Table 5).

Results show that removing any component degrades performance, confirming their complementary roles. Excluding RL reduces adaptability and causes a noticeable F_1 drop, highlighting its importance for dynamic threshold optimization. Replacing fuzzy reasoning leads to lower accuracy and

stability, indicating that fuzzy inference contributes not only to interpretability but also to robust decision-making. The most significant degradation occurs when the Transformer encoder is removed, demonstrating the critical role of attention-based representation learning in capturing long-range dependencies and ensuring strong generalization.

Figure 3 further illustrates that the full NFRT–IDS converges faster and achieves a higher asymptotic F_1-score than all ablated variants.

Sensitivity analysis with respect to learning rate η, dropout p, and the loss-weighting ratio $\lambda_1/(\lambda_1+\lambda_2)$, illustrated in Figure 3, shows that NFRT–IDS remains stable under moderate parameter variations. Optimal performance is achieved around $\eta = 10^{-4}$ and $p \in [0.2, 0.3]$, while a balanced loss weighting $(\lambda_1, \lambda_2) = (0.5, 0.5)$ provides the best trade-off between adaptive control and classification accuracy. Overall, these results confirm that NFRT–IDS benefits from a synergistic integration of Transformer-based representation learning, fuzzy reasoning, and reinforcement-driven adaptation, yielding a robust, stable, and well-balanced IDS suitable for real-world deployment.

Table 5: Ablation Study Results on CICIDS2017

Configuration	**Accuracy (%)**	**Recall (%)**	F_1 **(%)**
NFRT–Base	**99.98**	**99.97**	**99.98**
NFRT–NoRL	98.54	98.55	98.53
NFRT–NoFuzzy	97.71	97.12	96.92
NFRT–NoTrans	96.92	96.79	96.83

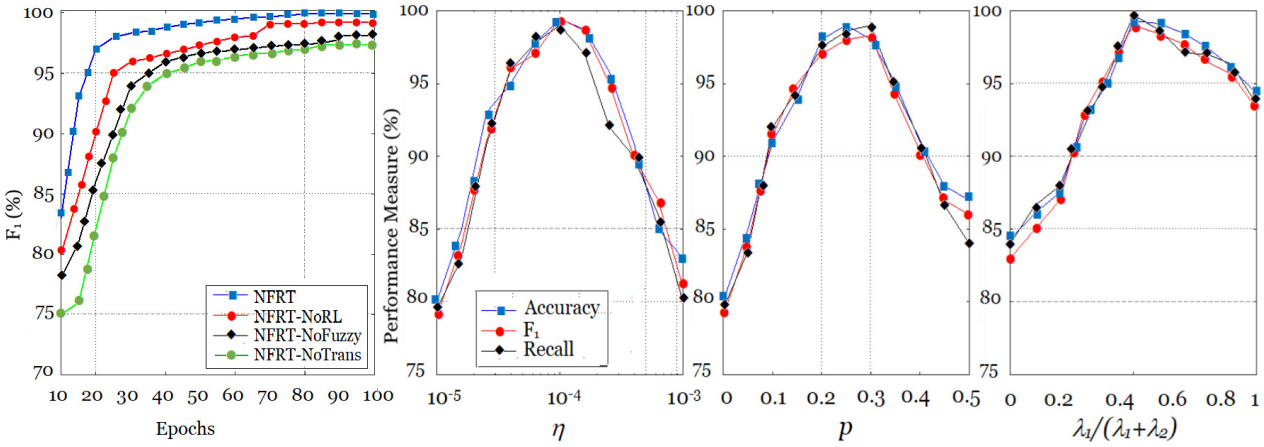

Figure 3: Training convergence comparison between NFRT–IDS and ablated variants and Sensitivity of NFRT–IDS performance to learning rate, dropout, and λ_1/λ_2 ratios on CICIDS2017.

5.4 Comparative Evaluation Against State-of-the-Art

Recent studies have demonstrated the potential of ML, DL and hybrid IDS architectures, each addressing different aspects of detection performance. Table 6 presents a comparative evaluation of the proposed NFRT–IDS against some ML, DL, and hybrid intrusion detection models across three benchmark datasets (CICIDS2017, UNSW–NB15, and BoT–IoT). Overall, the NFRT–IDS consistently surpasses existing methods in all evaluation metrics, achieving the highest accuracy, recall, and F_1-scores across datasets.

On CICIDS2017, NFRT–IDS attains 99.98% accuracy and F_1-score, with performance almost equal to PSO-ANN method (99.97%), and outperforming models such as deep networks like DNN (95.6%). Similarly, for UNSW–NB15, NFRT–IDS reaches 99.01% accuracy and 99.13 F_1-score, clearly exceeding RF and CNN Transformer baselines, which remain below 98.3%. The model also demonstrates strong robustness on the BoT–IoT dataset, achieving 99.05% accuracy and 99.19 F_1-score, surpassing recent hybrid architectures such as TCNN–LSTM and GRU-based detectors. These results confirm that combining Transformer-based feature learning, fuzzy inference reasoning, and reinforcement-driven

Table 6: Comparative Performance of some Recent ML, DL, and Hybrid IDS Approaches

Reference	Method / Model	Dataset	Accuracy (%)	Recall (%)	F_1-score (%)
[16]	DNN	CICIDS2017	95.60	–	–
[18]	PSO-ANN	CICIDS2017	99.97	99.95	99.96
[39]	Deep Belief NN	CICIDS2017	99.37	–	–
[40]	TCNN, LSTM	CICIDS2017	99.21	–	–
Proposed	**NFRT**	**CICIDS2017**	**99.98**	**99.97**	**99.98**
[15]	RF	UNSW–NB15	98.30	–	–
[41]	CNN, Transformer	UNSW-NB15	88.47	–	–
[42]	DNN	UNSW-NB15	97.04	–	7.04
[42]	RF	UNSW-NB15	91.77	–	91.77
Proposed	**NFRT**	**UNSW–NB15**	**99.01**	**99.12**	**99.13**
[8]	CNN–LSTM	BoT–IoT	91.20	–	90.10
[43]	CNN	BoT-IoT	92.85	–	–
[44]	TCNN-LSTM	Bot-IoT	97.50	–	95.70
[45]	GRU	Bot-IoT	95.30	–	93.40
Proposed	**NFRT**	**BoT–IoT**	**99.05**	**99.21**	**99.19**

adaptability enables NFRT–IDS to generalize effectively across heterogeneous environments while maintaining high precision and low false alarm rates.

5.5 Discussion

The experimental results reported in Section 5 demonstrate that NFRT–IDS achieves state-of-the-art detection performance across all evaluated benchmarks. Beyond raw accuracy, the proposed framework provides strong guarantees in terms of interpretability, computational efficiency, generalization capability, and operational robustness. This section synthesizes these aspects and discusses practical deployment considerations and remaining challenges.

5.5.1 Interpretability and Explainability

A central contribution of NFRT–IDS lies in its *intrinsic explainability*, achieved through the tight integration of Transformer attention mechanisms and fuzzy inference. Unlike post-hoc XAI approaches, interpretability is embedded directly within the decision-making pipeline.

Attention-Based Feature Attribution: Transformer attention weights first identify the most influential traffic attributes for each decision. Figure 4 presents the attention heatmap for the top-15 ranked features across the seven CICIDS2017 classes, highlighting class-specific relevance patterns. For instance, DoS/DDoS traffic is primarily influenced by *Packet Length Std*, *Total Length of Bwd Packets*, and *Fwd Packet Length Max*, whereas PortScan attacks are more sensitive to *Total Length of Bwd Packets*, *Subflow Bwd Bytes*, and *Bwd Packet Length Mean*. These consistent attention patterns provide a transparent and verifiable basis for subsequent semantic reasoning.

Fuzzy Rule-Based Reasoning: Fuzzy rules are then derived from the attention heatmap by selecting features with consistently high normalized attention per class, combined with domain expertise. The Fuzzy–Ensemble layer translates latent neural representations into human-readable linguistic rules, ensuring traceability between attention-based attribution, fuzzy semantics, and final decisions.

Representative extracted rules include:

If *Packet Length Std is high* ***and*** *Avg Packet Size is Medium* ***and*** *Packet Length Variance is Medium* ***Then*** *Traffic Class is DDoS*

If *Total Length of Bwd Packets is high* ***and*** *Bwd Packet Length Max is Medium* ***and*** *Fwd Packet Length Max is Medium* ***Then*** *Traffic Class is PortScan*

By grounding fuzzy rules in attention-driven feature relevance, NFRT–IDS enables explicit semantic reasoning and direct operational interpretation, effectively bridging the gap between high-performance deep learning and explainable intrusion detection.

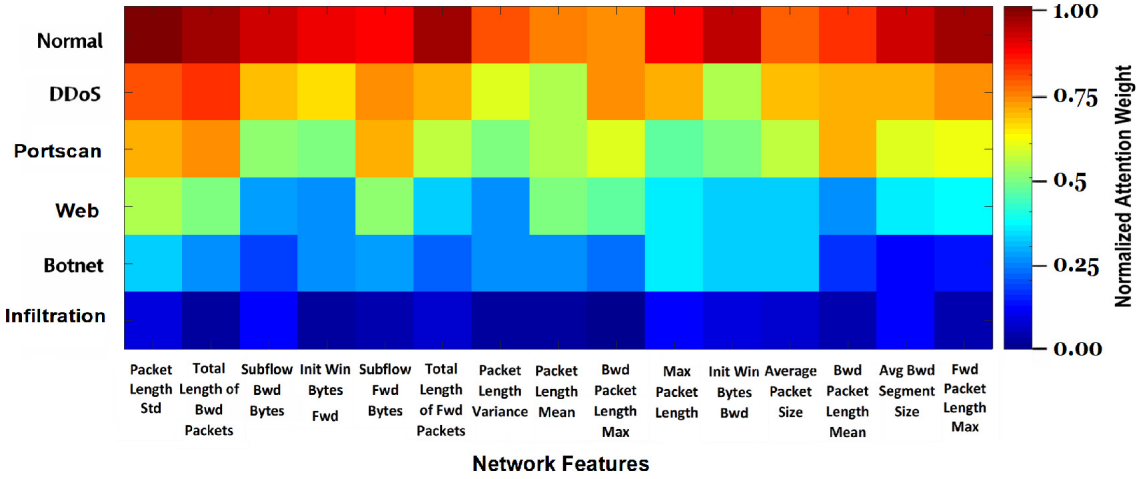

Figure 4: Attention heatmap showing feature importance for different attack types as learned by NFRT–IDS on the CICIDS2017 dataset.

5.5.2 Comparative Explainability Analysis against SHAP and LIME

To contextualize the intrinsic explainability of NFRT–IDS, we conducted a comparative analysis against two widely adopted post-hoc XAI techniques, namely SHAP and LIME [37], using the CICIDS2017 test set. The objective is not to replace post-hoc explainers, but to assess how the native interpretability mechanisms of NFRT–IDS compare within an intrusion detection context.

SHAP (KernelSHAP) and LIME explanations were generated a posteriori from the same trained NFRT–IDS model, while NFRT–IDS explanations were obtained directly from Transformer attention distributions and fuzzy rule activations. All methods were applied to identical test samples and feature sets to ensure a fair and consistent comparison.

Explainability was evaluated across four complementary dimensions: local and global interpretability, explanation stability, faithfulness to model decisions, and operational actionability. The results are summarized in Table 7.

Table 7: Comparative Explainability Analysis

Criterion	**SHAP**	**LIME**	**NFRT–IDS**
Local interpretability	High	High	High
Global interpretability	Medium	Low	High
Explanation stability	Medium	Low	High
Faithfulness (Top-K overlap)	0.71	0.63	–
Human-readable rules	No	No	Yes
Operational actionability	No	No	Yes

NFRT–IDS exhibits strong alignment with SHAP feature attributions, achieving an average Top-K overlap of 0.71, compared to 0.63 for LIME. At the same time, NFRT–IDS provides significantly more stable explanations across similar traffic instances, whereas LIME is more sensitive to local perturbations. The Top-K overlap metric is reported only for post-hoc methods, as NFRT–IDS relies on intrinsic attention patterns and fuzzy rule activations rather than feature approximation.

Unlike SHAP and LIME, which primarily deliver instance-level numeric attributions, NFRT–IDS produces a compact set of verifiable, human-readable fuzzy rules that capture global detection behavior. These rules can be directly validated against the attention heatmap (Figure 4) and operationally interpreted by security analysts.

Overall, this analysis demonstrates that NFRT–IDS achieves attribution faithfulness comparable to SHAP, while clearly outperforming post-hoc explainers in terms of stability, global interpretability, and practical actionability.

5.5.3 Computational Efficiency and Deployment Considerations

Runtime and memory profiling were conducted on the CICIDS2017 dataset (batch size 128, 64-dimensional embeddings) using an NVIDIA RTX A6000 GPU with PyTorch 2.2. Table 8 summarizes the computational characteristics. Despite its hybrid architecture, NFRT–IDS maintains competitive inference latency (5.0 ms per record) and reasonable training time (59 minutes for 300K samples). Transformer self-attention accounts for approximately 58% of the total computation, while fuzzy inference contributes less than 10%. Modular pretraining of the Transformer, DQL, and Fuzzy–Ensemble components facilitates faster convergence and efficient GPU utilization.

Table 8: Runtime and Memory Profiling of IDS Models on CICIDS2017 (175K training samples)

Model	Time Complexity	Memory (MB)	Params (M)	Latency (ms)	Training Time (min)
TCNN	$\mathcal{O}(n^2h)$	512	14.8	4.8	39
DQLANN	$\mathcal{O}(nh^2)$	275	6.2	5.5	45
FE	$\mathcal{O}(n(R+E))$	390	8.9	6.1	52
NFRT	$\mathcal{O}(n^2h+R+h^2)$	**610**	**16.3**	**5.0**	**59**

Deployment Considerations: The proposed system is well suited for real-time and near-real-time deployment; nevertheless, inference performance may depend on the availability of adequate computational resources and efficient batching strategies in high-throughput IoT environments. Furthermore, while fuzzy rule growth is effectively regulated through coverage-based pruning, scalability in very high-dimensional spaces and transient adaptation effects under abrupt traffic shifts remain topics for further optimization

5.5.4 Generalization and Robustness

Cross-dataset validation confirms the strong generalization capability of NFRT–IDS. Accuracy degradation remained below 0.5% when transferring between CICIDS2017, UNSW–NB15, and BoT–IoT datasets. The reinforcement learning component enhances adaptability by continuously refining detection policies, enabling effective operation in dynamic and evolving network environments.

5.5.5 Limitations

Despite its strong performance, NFRT–IDS presents several limitations. These include the computational cost of large-scale training, the sensitivity of the reinforcement learning agent to reward design, and the scalability of fuzzy rules in high-dimensional feature spaces. Moreover, continuous adaptation under encrypted or adversarial traffic remains an open challenge. Future work will focus on model compression, online reinforcement learning for non-stationary traffic, and adaptive fuzzy rule pruning strategies to preserve interpretability while reducing computational overhead.

6 Conclusion and Future Work

This paper presented an extensive exploration of hybrid Intrusion Detection System (IDS) architectures, culminating in the development of the Neuro–Fuzzy Reinforcement Transformer IDS (NFRT–IDS), a unified and interpretable framework that combines deep representation learning, reinforcement-based adaptability, and fuzzy reasoning for intelligent network protection. The study addressed the growing need for IDS solutions capable of achieving high accuracy while maintaining interpretability, adaptability, and computational efficiency across complex, large-scale network environments.

Four hybrid IDS architectures were developed and analyzed: (1) TCNN, integrating convolutional and attention mechanisms for spatio-temporal feature learning; (2) DQLANN, combining neural adaptation with Deep Q-Learning for policy-based optimization; (3) FE, merging fuzzy reasoning with ensemble classifiers to improve interpretability under uncertainty; and (4) NFRT–IDS, a unified framework consolidating these paradigms to achieve an optimal balance between accuracy, adaptability, and explainability.

Each model contributes uniquely: TCNN enhances feature representation, DQLANN introduces learning adaptivity, and FE provides semantic interpretability. NFRT–IDS fuses these complementary capabilities, establishing a coherent, multi-layered defense mechanism for modern cybersecurity systems. Extensive experiments on three benchmark datasets, CICIDS2017, UNSW–NB15, and BoT–IoT, validated the superiority of NFRT–IDS. For CICIDS2017, the model achieved up to 99.98% accuracy and F_1-score, and an AUC of 0.999, with similar performances for UNSW–NB15, and BoT–IoT. Inference latency is generally below 5.1 ms per sample, demonstrating feasibility for real-time detection.

Comparative and ablation studies confirmed that: each submodule contributes significantly to detection accuracy and stability. The hybridization of deep, fuzzy, and reinforcement mechanisms enhances robustness against class imbalance and data drift. In addition, NFRT–IDS exhibits strong generalization, with a very slight degradation in performance during cross-dataset validation. The interpretability analysis revealed that fuzzy rule-based reasoning and Transformer attention maps together provide transparency to the decision process.

Furthermore, computational evaluations demonstrated that modular and distributed training reduce resource usage while maintaining scalability for high-throughput network monitoring. From a theoretical standpoint, this work establishes a holistic learning paradigm that integrates a global representation learning via attention-driven Transformer encoding, a Reinforcement-based adaptation for continuous policy optimization, and a Fuzzy logic reasoning to enhance semantic interpretability and reduce uncertainty.

Practically, the NFRT–IDS framework demonstrates how hybrid intelligence can yield robust, interpretable, and scalable network defense systems suitable for deployment in IoT, SDN, and cloud infrastructures. Its modular design allows independent pre-training of subcomponents and parallel execution across GPU or cloud clusters, ensuring operational feasibility in real-world, high-bandwidth environments.

Despite its demonstrated strengths, several challenges remain for large-scale adoption. Training remains GPU-intensive, especially on large datasets; lightweight variants are needed for edge devices. The DQL component's convergence strongly depends on the quality of the reward signal. The number of fuzzy rules may increase combinatorially with feature dimensionality. Further testing on encrypted, adversarial, and streaming traffic is required to ensure robustness under dynamic attack conditions.

Future research will focus on extending NFRT–IDS toward federated and privacy-preserving collaborative learning, lightweight optimization via pruning and mixed-precision inference, and dynamic fuzzy rule evolution through meta-learning. Additional directions include graph-based modeling using Graph Neural Networks (GNNs) to capture inter-host dependencies and reinforcement-driven adversarial training to enhance resilience against evasion and poisoning attacks.

Funding

This research received no external funding.

Author contributions

The authors contributed equally to this work.

Conflict of interest

The authors declare no conflict of interest.

References

[1] Song B. (2024). Random Forest Based Intrusion Detection System, *2024 Asian Conference on Communication and Networks (ASIANComNet)*, Bangkok, Thailand, pp. 1-4, 2024.

[2] Wang C., Sun Y., Lv S., Wang C., Liu H., Wang B. (2023). Intrusion Detection System Based on One-Class Support Vector Machine and Gaussian Mixture Model, *Electronics*, 12(4), 930, 2023.

[3] Krishna, K.V., Swathi, K., Rao, B.B. (2020). A novel framework for NIDS through fast kNN classifier on CICIDS2017 dataset, *International Journal of Recent Technology and Engineering*, 8(5), 3669-3675,2020.

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

[5] Liu H. and Lang B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey, *Applied Sciences*, 9(20), 4396, 2019.

[6] Qazi E. U. H., A.Almorjan, and Zia T. (2022). A One-dimensional Convolutional Neural Network (1D-CNN) based deep learning system for network intrusion detection, *Applied Sciences*, 12(16), 7986, 2022.

[7] Al-Turaiki I. and Altwaijry N. (2021). A convolutional neural network for improved anomaly-based network intrusion detection, *Big Data*, 9(3), 233–252, 2021.

[8] Halbouni A., Gunawan T. S., Habaebi M. H., Halbouni M., Kartiwi M., and Ahmad R. (2022). CNN-LSTM: Hybrid deep neural network for network intrusion detection system, *IEEE Access*, 10, 99837–99849, 2022.

[9] Kamal H., Mashaly M. (2024). Advanced Hybrid Transformer-CNN Deep Learning Model for Effective Intrusion Detection Systems with Class Imbalance Mitigation Using Resampling Techniques, *Future Internet*, 16, 481, 2014.

[10] Yaras S., Dener M. (2024). IoT-Based Intrusion Detection System Using New Hybrid Deep Learning Algorithm, *Electronics*, 13, 1053, 2024.

[11] Muhuri P., Chatterjee P., Yuan X., Roy K., Esterline A. (2020). Using a long short-term memory recurrent neural network (lstm-rnn) to classify network attacks, *Information*, 11, 243, 2020.

[12] Ullah S., Ahmad J., Khan M.A., Alshehri M.S., Boulila A., Koubaa A., Ullah S.J., et al. (2023). TNN-IDS: Transformer neural network-based intrusion detection system for MQTT-enabled IoT Networks, *Computer Networks*, 237, 110072, 2023.

[13] Shone N., Ngoc T. N., Phai V. D., and Shi Q. (2018). A Deep Learning Approach to Network Intrusion Detection, *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(1), 41–50, 2018.

[14] Gad A.R., Nashat A.A., Barkat T.M. (2021). Intrusion detection system using machine learning for vehicular ad hoc networks based on ToN-IoT dataset, *IEEE Access*, 9, 142206–142217, 2021

[15] Turk F. (2023). Analysis of intrusion detection systems in UNSW-NB15 and NSL-KDD datasets with machine learning algorithms, *Bitlis Eren Üniversitesi Fen Bilim.* Derg., 12, 465–477, 2023.

[16] Vinayakumar R., Alazab M., Soman K.P., Poornachandran P., Al-Nemrat A., Venkatraman S. (2019). Deep learning approach for intelligent intrusion detection system, *IEEE Access*, 7, 41525–41550, 2019.

[17] Stiawan D., Idris M.Y.B., Bamhdi A.M., Budiarto R. (2020). CICIDS-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access*, 8, 132911–132921, 2020.

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

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

[20] Anand M., Muthurajkumar S. (2025). An intelligent IDS using bagging based fuzzy CNN for secured communication in vehicular networks, *Scientific Reports*, 15, 26952, 2025.

[21] Qiu X., Shi L. and Fan P. (2025). A cooperative intrusion detection system for internet of things using fuzzy logic and ensemble of convolutional neural networks, *Scientific Reports*, 15, 15934, 2025.

[22] S. Yu et al., (2024). Deep Q-Network-Based Open-Set Intrusion Detection Solution for Industrial Internet of Things, *IEEE Internet of Things Journal*, 11(7), 12536-12550, 2024.

[23] Ren K., Zeng Y., Zhong Y., Sheng B., Zhang Y. (2023). MAFSIDS: a reinforcement learning based intrusion detection model for multi-agent feature selection networks, *Journal of Big Data*, 10 (1), 137, 2023.

[24] Kamal M., Mashaly H. (2025). Enhanced Transformer–Convolutional hybrid intrusion detection architecture for large-scale networks, *IEEE Transactions on Network Science and Engineering*, 12, 212–228, 2025.

[25] Zhao, Y., Lu, J.L. (2024). Spatiotemporal Sequence Prediction Based on Spatiotemporal Self-Attention Mechanism, *International Journal of Computers Communications & Control*, 19(6), 6771, 2024.

[26] Zadeh L.A. (1965). Fuzzy sets, *Information and control*, 8, 338-353, 1965.

[27] Breiman L. (2001). Random forests, *Machine Learning*, 45(1), 5–32, 2021.

[28] Chen, S., Li, G., Chang, K., Hu, X., Li, P., Wang, Y., Zhang, Y. (2024). Ultra-short-term Load Forecasting Based on XGBoost-BiGRU, *International Journal of Computers Communications & Control*, 19(5), 6631, 2024.

[29] Ke T Y.G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., and Liu. (2017). LightGBM: a highly efficient gradient boosting decision tree, *In Proceedings of the 31st International Conference on Neural Information Processing Systems* (NIPS'17), Curran Associates Inc., Red Hook, NY, USA, 3149–3157, 2017.

[30] Rini P., Shamsuddin M., Yuhaniz S. (2011). Particle Swarm Optimization: technique, system and challenges, *International Journal of Computer Applications*, 14, 19–27, 2011.

[31] Silver D., Sutton R. S. (2018). *Reinforcement Learning: An Introduction*, (2nd ed.), MIT Press, 2018.

[32] CICIDS2017 Dataset. (2017). https://www.unb.ca/cic/datasets/ids-2017.html

[33] Moustafa, N., and Jill S. (2015). UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set), *Military Communications and Information Systems Conference* (MilCIS), IEEE, 2015.

[34] Koroniotis N., Moustafa N., Sitnikova E. and Turnbull B. (2019). Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset, *Future Generation Computer Systems*, 100, 779-796, 2019.

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

[36] Zachary C. L. (2018). The mythos of model interpretability, *Communications of the ACM*, 61(10), 36–43, 2018.

[37] Givisis I, Kalatzis D, Christakis C, Kiouvrekis Y. (2025). Comparing Explainable AI Models: SHAP, LIME, and Their Role in Electric Field Strength Prediction over Urban Areas. *Electronics*, 14(23):4766.

[38] Conover, W.J. (1973). On Methods of Handling Ties in the Wilcoxon Signed-Rank Test, *Journal of the American Statistical Association*, 68 (344), 985–988, 1973.

[39] Manimurugan S., Al-Mutairi S., Aborokbah M. M., Chilamkurti N., Ganesan S. and Patan R. (2020). Effective Attack Detection in Internet of Medical Things Smart Environment Using a Deep Belief Neural Network, *IEEE Access*, 8, 77396-77404, 2020.

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

[41] Xing N., Zhao S., Wang Y., Ning K., Liu X. (2023). A dynamic intrusion detection system capable of detecting unknown attacks, *International Journal of Advanced Computer Science and Applications*, 14, 2023.

[42] Mayhew M., Atighetchi M., Adler A., Greenstadt R. (2015). Use of machine learning in big data analytics for insider threat detection, *In: Proceedings of the MILCOM –2015 IEEE Military Communications Conference*, Canberra, Australia, 10–12 November 2015, pp. 915–922, 2015.

[43] Saba T., Rehman A., Sadad T., Kolivand H., Bahaj S.A. (2022). Anomaly-based intrusion detection system for IoT networks through deep learning model, *Computers and Electrical Engineering*, 99, 107810, 2022.

[44] Alashjaee A.M. (2025). Deep learning for network security: an Attention-CNN-LSTM model for accurate intrusion detection, *Scientific Reports*, 15, 21856, 2025.

[45] Shahin M., Maghanaki M., Hosseinzadeh A. and Chen F. F. (2024). Advancing IIoT cybersecurity via AI-enabled IDS architectures, *Advanced Engineering Informatics*, 62, 102685, 2024.

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

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

Cite this paper as:

Mjahed, Soukaina; Mjahed, 0uail (2026). NFRT–IDS: A Unified Neuro-Fuzzy Reinforcement Transformer Architecture for Adaptive and Explainable Intrusion Detection, *International Journal of Computers Communications & Control*, 21(2), 7405, 2026.
https://doi.org/10.15837/ijccc.2026.2.7405