SSRG International Journal of Electrical and Electronics Engineering
ISSN: 2348-8379/ https://doi.org/10.14445/23488379/IJEEE-V13I11P102

Volume 13 Issue 11, 14-39, November 2026
© 2026 Seventh Sense Research Group®

Original Article

Explainable Hybrid Reinforcement Learning with Adaptive Fractional-Order Optimization for Robust Network Traffic Anomaly Detection

K.V.V.B.Durgaprasad[1], V.L. Padma Latha[2], K. Naga Jyothi[3]

[1,3]*BEST Innovation University, Gownivaripalli, Gorantla, Anantapur 515231, Andhra Pradesh, India.*
[2]*Malla Reddy University, Gundlapochampally, Hyderabad 500090, Telangana, India.*

[1]*Corresponding Author: 2023wpcse009@bestiu.edu.in*

Received: 29 July 2026 Revised: 15 October 2026 Accepted: 30 October 2026 Published: 25 November 2026

Abstract - *Anomaly detection in modern networks must cope with high traffic volume, shifting attack behaviour, a skewed ratio of benign to malicious flows and the need for decisions that a security analyst can verify. Supervised detectors trained on fixed decision boundaries adapt poorly to such conditions, and most reinforcement learning detectors rely on a single value-based or policy-based learner whose training is often unstable. This paper presents HDP-AFNet, a hybrid reinforcement learning framework in which a Deep Q-Network (DQN) branch and a Policy Gradient (PG) branch share one traffic encoder and are combined through a weighted fusion gate. Training uses an optimizer derived from Adaptive Fractional-Order Gradient Descent (AFoGD), whose fractional memory term damps oscillating parameter updates. UNSW-NB15 flow records are encoded, Min-Max normalized, reduced to 32 features by mutual information and grouped into windows of 10 consecutive flows that form the agent's states. An asymmetric reward penalizes a missed attack four times more heavily than a false alarm. On UNSW-NB15, HDP-AFNet reached 94.92% accuracy, 95.84% precision, 96.74% recall, 96.28% F1-score, 0.9368 ROC-AUC and 0.8826 MCC, and the full model exceeded a plain DQN by 6.18 percentage points in accuracy. Inference required 5.39 × 10^{-6} s per flow. Reward attribution, action-probability heatmaps, temporal confidence profiles and feature attribution maps show how each decision was reached.*

Keywords - *Network traffic anomaly detection, Hybrid reinforcement learning, Deep Q-Network (DQN), Policy Gradient, Adaptive Fractional-Order Gradient Descent (AFoGD), Explainable Artificial Intelligence, Cybersecurity.*

1. Introduction

Cloud services, Internet of Things (IoT) deployments, 5G networks and edge infrastructure have enlarged both the volume of network traffic and the number of points through which an attacker can enter [1]. Denial of service, reconnaissance, botnet command traffic, privilege escalation and zero-day exploits now change form faster than signature databases can be updated, so Intrusion Detection Systems (IDS) that match traffic against known signatures or fixed rules miss much of this activity [1, 2]. A practical detector must satisfy four requirements at once: it must adapt when the traffic distribution drifts, remain reliable when attack and benign flows are unevenly represented, run fast enough for line-rate monitoring, and give the analyst a reason for each alert. Reinforcement Learning (RL) meets the first requirement because an agent refines its policy through repeated interaction and reward feedback [3, 5], while the last requirement has made Explainable AI (XAI) and Explainable RL (XRL) a necessary part of automated security decisions [4, 20]. Classical learners such as support vector machines, random forests and naive Bayes were the first data-driven alternatives to signature matching [6, 21]. Convolutional, recurrent, autoencoder and Transformer models later learned traffic representations directly from high-dimensional flow features [7, 8, 40]. Deep Reinforcement Learning (DRL) extended this line by treating detection as sequential decision-making, using DQN [9], double DQN [10, 18], SARSA [37], actor-critic learners [11] and policy gradient methods [12, 13]. These detectors have been evaluated mainly on UNSW-NB15, NSL-KDD, CICIDS2017 and Bot-IoT [14, 15, 39], and post-hoc tools such as SHAP, saliency maps and gradient attribution have been attached to several of them [16, 34, 42]. Four limitations remain in this body of work. Supervised detectors depend on labelled data and generalize poorly when attack patterns shift [2], and their accuracy falls when the minority class is small, or the traffic is noisy [17]. RL detectors are built on a single learning paradigm: value-based agents such as DQN overestimate action values and oscillate during training [9, 18], whereas policy gradient agents explore better but suffer from high-variance gradient estimates [12,

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

13]. Recent reviews confirm that DRL-based IDS studies seldom examine training stability, real-time cost and robustness together [38, 39]. Explanations in current IDS are almost always feature attributions computed after training [34, 42]; they do not show how reward, action preference and confidence evolve inside an RL agent, which is the information an analyst needs before trusting an autonomous detector [4, 20]. No existing framework combines value-based and policy-based learning, stabilizes their joint training and explains the resulting policy in RL-specific terms.

This paper proposes Hybrid DQN-Policy Gradient Anomaly Fusion Network (HDP-AFNet), an explainable hybrid RL framework for binary network traffic anomaly detection. A shared encoder maps each temporal window of flows to a latent state that feeds a DQN head and a PG head in parallel, and a fusion gate converts the Q-values into action probabilities and blends them with the policy output. Training uses an AFoGD-inspired optimizer [19, 41] whose fractional memory term accumulates past gradients with decaying weights and suppresses the oscillation observed in single-paradigm agents. An asymmetric reward penalizes missed attacks more than false alarms, which directs the policy towards high recall on the imbalanced UNSW-NB15 data [14]. An XRL module reports feature attribution, reward attribution, action-probability heatmaps and temporal confidence for each decision [20].

The novelty of HDP-AFNet lies in the joint design rather than in any single component. Earlier RL-based detectors use one learning paradigm, either value-based [27-29, 37] or actor-critic [33], and optimize it with standard gradient methods or metaheuristic search [32]. HDP-AFNet trains value-based and policy-based heads on a shared representation and fuses their action probabilities at decision time. Fractional-order gradient descent has been studied for supervised networks [19, 41], but its use for stabilizing a hybrid RL detector has not been reported. Explanations in earlier IDS stop at feature relevance [34, 42], whereas HDP-AFNet also explains reward accumulation and decision confidence over time. The contributions are as follows:

- A hybrid DQN-PG architecture with a shared encoder and a fusion gate ($\lambda = 0.55$) that blends softmax-normalized Q-values with the stochastic policy, raising the F1-score from 89.77% for a DQN alone to 94.48% before the optimizer is changed.
- An AFoGD-inspired optimizer (fractional order 0.8, memory coefficient 0.95) that adds a decaying gradient memory to each update, lowering reward variance and adding a statistically significant 1.80 percentage points of F1-score over the same architecture trained without it ($p = 0.006$).
- A temporal RL formulation of detection in which windows of 10 flows form the state and an asymmetric reward (+1, -2, -0.5, +0.2) favours attack recall on imbalanced traffic, yielding 96.74% recall and a false negative rate of 3.26%.
- An XRL module producing reward attribution, action-probability heatmaps, temporal confidence and feature attribution, evaluated together with detection, stability and latency (5.39×10^{-6} s per flow) on UNSW-NB15.

Section 2 reviews related work and identifies research gaps. Section 3 presents the HDP-AFNet architecture and optimization strategy. Section 4 reports the experimental setup, results, ablation studies, and comparative analysis. Section 5 concludes the paper.

2. Literature Review

2.1. Machine Learning and Deep Learning-based Intrusion Detection Systems

Recent studies have applied machine and deep learning to intrusion detection, DDoS defense, and anomaly analysis across cloud, SDN, botnet, and vehicular network settings. In cloud environments, [21] enhanced a Naïve Bayes classifier with Gaussian likelihood estimation to address zero-frequency and class imbalance problems in the feature space. Feature relevance was assessed through Pearson correlation, mutual dependence scores, and chi-square testing, while minority samples were synthetically oversampled with SMOTE.

The variant driven by mutual dependence raised accuracy by nearly 2%, and a second variant gained about 1.5% in both accuracy and precision, strengthening DDoS attack recognition. Likewise, [22] designed a cloud intrusion detector that fused recurrent LSTM units, a support vector classifier, gradient-boosted trees, and a fast-learning network, with ensemble weights tuned by crow search optimization. This design reached accuracies of 99.07%, 98.99%, and 99.99% on NSL-KDD, Kyoto, and CSE-CIC-IDS-2018, respectively, along with improved recall, precision, F-measure, detection rate, and false alarm rate.

In the field of intelligent transportation and vehicular networking, a number of studies have been conducted on how to tackle the new cyber security threats by utilizing advanced machine learning frameworks. In the study [23], a Random Forest-based IDS for Vehicular Ad-Hoc Networks (VANETs) was introduced, and posterior detection was proposed using coresets that are needed to be able to process the huge amount of vehicular data generated in urban areas.

The framework proved to be more accurate in detection than the traditional machine learning solutions for IDS and faster in computation. The architecture given in [24] was also proposed as a solution for VANET Cloud environments to detect and counter Distributed Denial of Service (DDoS) attacks in Connected and Autonomous Vehicle Networks. It monitored and analyzed the architecture of the network, used classification mechanisms and predictive analytics of the machine learning techniques, with a detection accuracy of

99.59%. The results showed the increasing significance of intelligent security mechanisms for envisioned next-generation vehicular communication systems and cloud-supported transportation infrastructure.

Some researchers also investigated the performance of network security using hybrid feature engineering and SDN-enabled detection frameworks. A study in [25] designed a hybrid feature selection method based on Categorical Analysis, Mutual Information and Principal Component Analysis (PCA) followed by Extra Trees ensemble classifier to detect bots. The proposed approach resulted in classification accuracy of 99.99%, which is very high, near 100% True Positive Rate (TPR) and a False Positive Rate (FPR) close to zero, and thus, it can be stated that the proposed approach is very robust in the face of new variations of attack botnets. In addition, the article in [26] suggested the RDAER approach to early detection of DDoS attacks in Software-Defined Networking (SDN)-based cloud environments.

The framework combined Recursive Feature Elimination (RFE), DBSCAN clustering, ARIMA forecasting, Lyapunov exponent analysis, exponential smoothing, dynamic thresholding and rule-based classification to detect traffic anomalies at SDN switches. Evaluation on the CICDDoS2019 dataset achieved 99.92% detection accuracy and detection time of just 20 seconds on the experimental results. While these machine learning and deep learning frameworks demonstrated excellent detection accuracy, supervised learning approaches, manually defined features, and pre-tuned decision boundaries sometimes made them less adaptable to new attack types, zero-day threats, and network dynamics, thus inspiring the development of more adaptable and intelligent intrusion detection systems.

2.2. Reinforcement Learning-based Intrusion Detection Frameworks

Intrusion detection research has recently turned to reinforcement learning and its deep variants, seeking greater adaptability and sharper decision-making against attacks in changing network conditions. Work in [27] built an adaptive cloud intrusion detector driven by deep reward-based learning, aimed at recognizing novel and sophisticated threats. Tested on UNSW-NB15, it tackled the difficulty that signature-matching and conventional supervised detectors face with previously unseen attack patterns. Its results surpassed leading detection schemes in both detection accuracy and false alarm rate, confirming the value of reward-driven learning for protecting cloud infrastructure [27]. Other researchers have paired deep reward-based agents with optimization and representation learning to raise detection performance further. In this direction, [28] developed an intrusion detection scheme for software-defined networks in which a deep reinforcement agent operated alongside Long- and Short-Term Memory Sequence Recurrent Neural Networks (LSTM-RNN), supported by a Particle Cloud-Integrated Joint Time- and Feature-Optimization Algorithm (PC-JTFOA)

The proposed framework simultaneously dealt with the detection of DDoS attacks, optimization of the features, adaptive load balancing, and intelligent management of networks. Evaluation on the NSL-KDD and WPPD datasets outperformed the conventional intrusion detection techniques by achieving 99.85% accuracy, 98.67% sensitivity, 97.42% specificity and 1423 ms response time.

Likewise, in [29] an abnormal traffic detection scheme was introduced based on the Semi-Supervised Double Deep Q-Network (SSDDQN), which combined anomaly detection via feature reconstruction using an autoencoder, K-Means clustering, and classification using deep neural networks. It could handle high-dimensional network traffic data and limited availability of labeled samples, showed lower computational complexity and better detection performance than the classic machine learning methods on NSL-KDD and AWID datasets [28, 29].

The rapid rise of the importance of reinforcement learning in the field of cybersecurity was also highlighted by the systematic literature review in [30] that covered all the diverse architectures employed by RL-based IDSs, learning strategies, and feature representation techniques used in network traffic analysis. One key aspect of the survey was the capability of RL algorithms to learn from the environment, adapt to changing cyber threats, and enhance detection accuracy under dynamic network conditions.

All of the studies in [27-30] showed that reinforcement learning-based intrusion detection methods exhibited substantial improvements in adaptability, autonomous decision-making, and attack detection capabilities relative to traditional supervised learning methods. Although most of them performed well, the current methods were built around a single reinforcement learning paradigm, were not very stable in learning process, faced convergence problems, were computationally expensive, and could not sustain stable performance when traffic distributions change drastically or the network size becomes large, suggesting the necessity to develop more stable, scalable and adaptive intelligent intrusion detection solutions.

2.3. Hybrid Deep Reinforcement Learning and Optimization-based Security Frameworks

Current studies increasingly favor hybrid designs that couple deep reward-based learning with generative and optimization techniques for stronger attack detection. In [31], adversarially trained generator and discriminator networks were joined with a deep reinforcement agent, where the generator produced artificial traffic records to rebalance the skewed class distribution of NSL-KDD. Training on this augmented data lifted classification accuracy for rare attack

categories compared with a model learned from the original imbalanced records [31]. Several optimization-guided reinforcement frameworks have also targeted DDoS detection and mitigation in cloud and software-defined networks.

The MMEDRL-ADM scheme in [32] selected features through a strategy inspired by African Buffalo Foraging (ABO-FS), classified traffic with a layered ensemble of deep reinforcement agents (MEDRL), and tuned hyperparameters via an enhanced grasshopper swarm search (IGOA). It achieved stronger DDoS detection and mitigation at reduced computational cost relative to prior methods.

In a similar vein, [33] proposed a real-time DDoS detector that merged three actor-critic agents, namely twin delayed deterministic policy gradient (TD3), Deterministic Policy Gradient (DDPG), and Advantage Actor-Critic (A2C), with a hybrid feature filter drawing on Boruta relevance ranking, Shapley-based attributions, and stability checks across cross-validation folds. Evaluation on the CICDDoS2019 and UNSW-NB15 datasets revealed that the TD3 model was accurate at 99.12%, high with AUC at 99.21%, and its inference latency was just 1.87ms, which makes it suitable for scalable cloud security applications [32, 33].

The works of [31-33] showed that incorporation of reinforcement learning with generative models and metaheuristic optimization methods enhanced the accuracy of intrusion detection, the ability to deal with class imbalance, and the real-time decision-making capability. But most of the current works were based on either value-based or actor-critic reinforcement learning approaches and did not leverage the complementary merit of Deep Q-Network (DQN) and Policy Gradient learning into a single Adaptive Intrusion Detection (AID) system [31-33].

2.4. Explainable Artificial Intelligence for Cybersecurity

As intelligent cybersecurity systems become more common, the need for explainable and transparent intrusion detection mechanisms is becoming paramount. The work in [34] developed an interpretable intrusion detection and prevention system in which Shapley-value attributions guided feature choice, and classification relied on an ensemble mixing bagged and boosted learners with residual-based boosting (HBBR).

It attained a peak accuracy of 98.47% for binary detection and an F1-score of 96.19% for multiclass detection, while offering greater robustness, clearer reasoning, and lower overfitting risk than the security methods it was compared against [34]. Similarly, [33] built an adaptive reward-driven framework for real-time DDoS recognition in cloud platforms, employing TD3, Deterministic Policy Gradient (DDPG), and Advantage Actor-Critic (A2C) agents. Feature relevance was determined through a hybrid filter uniting Boruta ranking, Shapley-based attributions, and stability checks across validation folds, which also made the learned decisions more transparent. On CICDDoS2019 and UNSW-NB15, the TD3 agent delivered 99.12% accuracy, an AUC of 99.21%, and an inference delay of only 1.87 ms, offering an efficient and explainable detector for cloud settings [33].

The research from [33, 34] thus highlighted the importance of explainability for improving the performance of intrusion detection systems, the transparency of their operation, and analysts' trust. But the current explainable cybersecurity frameworks were mainly based on feature attribution and did not offer details about how reinforcement learning models make decisions.

This means that existing work on intrusion detection does not attempt to harness the properties of interpretability mechanisms, which are unique to RL, such as reward attribution, action selection analysis, policy evolution tracking, or temporal decision explanations [33, 34]. The summary of the research that uses Reinforcement Learning (RL) and Policy Gradient algorithms with Intrusion Detection Systems (IDSs) is presented in Table 1.

2.5. Recent Advances in DRL-, Attention- and Explanation-based Intrusion Detection

Beyond DQN-type agents, on-policy value learning has also been applied to intrusion detection. Mohamed and Ejbali [37] combined SARSA with a deep neural network and reported higher accuracy, precision, recall and F1-score than classical and deep learning baselines on NSL-KDD and UNSW-NB15. Attention-based detectors form a parallel line of work: Umer et al. [40] paired wrapper-based feature selection with a multi-head attention Transformer and obtained 93% accuracy and 92% F1-score on UNSW-NB15. Both studies indicate that feature selection and sequence modelling benefit flow classification, and HDP-AFNet adopts these through mutual information selection and temporal windows.

The reliability of DRL detectors has also been examined. Merzouk et al. [38] evaluated DRL-based IDS on NSL-KDD, UNSW-NB15 and CICIoV2024 under gradient-based adversarial perturbations and found that robustness depends strongly on the DRL algorithm and on network depth and width, with TRPO proving more resilient than the other algorithms tested. A bibliometric review of DRL for intrusion detection [39] found that DQN variants, policy gradient and actor-critic agents dominate the field, but that most studies rely on legacy benchmarks and rarely report real-time cost or training behaviour.

On the explanation side, the systematic review in [42] noted that SHAP and LIME dominate XAI-based IDS, that their computational overhead limits real-time use, and that standard criteria for evaluating explanations are still missing. Fractional-order gradient descent, first analysed for back-

propagation networks with the Caputo derivative [41], showed monotone convergence and better training behaviour than integer-order updates, yet it has not been examined for RL-based detectors. Table 1 summarizes these studies together with those discussed in Subsections 2.1 to 2.4.

Table 1. Recent studies on Intrusion Detection Systems (IDSs) that employ Reinforcement Learning (RL) and Policy Gradient algorithms

Ref. No.	Method Name	Dataset Used	Output Measures	Research Gap Identified
[21]	Gaussian Naïve Bayes with PCC, Mutual Information, Chi-Squared, and SMOTE	NA	Accuracy improvement: 2%; and precision improvement: 1.5%	Limited zero-day attack detection; feature engineering dependency.
[22]	CSA-based Weighted Ensemble IDS (LSTM + SVM + XGBoost + FLN)	NSL-KDD, Kyoto, CSE-CIC-IDS-2018	Accuracy: 99.07% (NSL-KDD), 98.99% (Kyoto), 99.99% (CSE-CIC-IDS-2018);	High computational complexity; lack of explainability.
[23]	Random Forest with Coreset-based Posterior Detection for VANET IDS	NA	Significantly improved detection accuracy and efficiency over conventional ML methods	Limited attack diversity evaluation; unclear scalability.
[24]	Machine Learning-based Security Architecture for VANET Cloud DDoS Detection	VANET Cloud traffic dataset	Accuracy: 99.59%	DDoS-specific focus; limited attack generalization.
[25]	Hybrid Feature Selection (Categorical Analysis + Mutual Information + PCA) with Extra Trees Classifier	Varied botnet datasets	Accuracy: 99.99%; TPR: >99%; FPR: ~0.00%	Limited zero-day robustness; feature selection overhead.
[26]	RDAER Framework (RFE + DBSCAN + ARIMA + Lyapunov Exponent + Rule-Based Classifier)	CICDDoS2019	Accuracy: 99.92%; Detection Time: 20 s	Detection latency; high implementation complexity.
[27]	Deep Reinforcement Learning-based Adaptive Cloud IDS	UNSW-NB15	Higher detection accuracy and lower FPR than existing IDSs	Lack of explainability; single-dataset evaluation.
[28]	DRL-IDS using LFTS-RNN and PC-JTFOA for SDN	NSL-KDD, WPPD	Acc: 99.85%; Sensitivity: 98.67%; Specificity: 97.42%; Response Time: 1423 ms	High model complexity; limited real-world validation.
[29]	Semi-Supervised Double Deep Q-Network (SSDDQN)	NSL-KDD, AWID	Reduced time complexity	Limited scalability evaluation; policy interpretability absent.
[30]	Survey of Reinforcement Learning-based IDSs	Multiple studies and datasets	Review of RL algorithms, feature representations, challenges, and future directions	No experimental validation; deployment challenges.
[31]	GAN-assisted Deep Reinforcement Learning IDS	NSL-KDD	Improved classification of minority attack classes using synthetic data	Relied on synthetic data quality generated by GAN; Dependence on synthetic data; limited validation.
[32]	MMEDRL-ADM (ABO-FS + MEDRL + IGOA)	NA	Improved DDoS detection	High model complexity; scalability concerns.
[33]	DRL-based DDoS Detection using TD3, DDPG, and A2C with Boruta-SHAP Feature Selection	CICDDoS2019, UNSW-NB15	Accuracy: 99.12%; AUC: 99.21%; Inference Latency: 1.87 ms	Benchmark-only evaluation; high DRL training cost.

[34]	SHAP-based Explainable AI with Hybrid Bagging-Boosting and Boosting on Residuals	NA	Accuracy: 98.47%, F1-score: 96.19%	Limited adaptability; SHAP computational overhead.
[37]	Deep SARSA with deep neural network	NSL-KDD, UNSW-NB15	Higher accuracy, precision, recall and F1-score than ML and DL baselines	Single value-based learner; no decision explanation
[38]	Robustness evaluation of DRL-based IDS under gradient-based attacks	NSL-KDD, UNSW-NB15, CICIoV2024	Robustness varies with algorithm and network size; TRPO most resilient	Evaluation study; no training stabilization proposed
[39]	Bibliometric review of DRL-based IDS	Multiple studies	DQN variants, PG and actor-critic agents dominate	Legacy benchmarks; limited real-time and stability reporting
[40]	Wrapper feature selection with multi-head attention Transformer	UNSW-NB15	Accuracy: 93%; F1-score: 92%	Supervised; no adaptive policy learning
[41]	Fractional-order gradient descent with Caputo derivative for BP networks	Benchmark classification data	Proven monotonicity and convergence	Not applied to RL or intrusion detection
[42]	Systematic review of XAI in IDS	Multiple studies	SHAP and LIME dominate	Explanation overhead; no RL-specific explanation

2.6. Research Gap and Motivation

Table 1 shows that existing detectors differ widely in model family, yet they share four gaps. First, RL-based IDS rely on a single paradigm: value-based agents [27-29, 37] provide stable action values but explore poorly, whereas actor-critic and policy gradient agents [33] explore better at the cost of high-variance updates. Second, training stability is rarely treated as a design objective; metaheuristic tuning [32] and GAN-based augmentation [31] improve accuracy but do not damp oscillating updates, and fractional-order optimization [19, 41] has not been tested in this setting. Third, class imbalance is handled by resampling or synthetic data [21, 31] rather than through the learning signal itself. Fourth, explanations are limited to feature relevance [34, 42]. HDP-AFNet addresses these gaps jointly, as mapped in Table 2.

Table 2. Research gaps in existing detectors and the corresponding HDP-AFNet components

Research gap	Existing work	HDP-AFNet component	Verified in
Single RL paradigm (value-based or policy-based)	[27-29, 33, 37]	Shared encoder with DQN and PG heads joined by a fusion gate	Table 6: F1-score 89.77% to 94.48%
Unstable RL training	[18, 38, 39]	AFoGD-inspired optimizer with fractional gradient memory	Figures 5 and 6; Tables 7 and 10
Imbalance handled by resampling or synthetic data	[21, 31]	Asymmetric reward (-2 missed attack, -0.5 false alarm)	Tables 9 and 11: recall 96.74%, FNR 3.26%
Per-flow static input	[22, 25]	Sliding windows of 10 flows as RL states	Section 3.3, Step 2
Feature-only explanations	[34, 42]	XRL: reward attribution, action heatmaps, temporal confidence, feature attribution	Section 3.3, Step 9

3. Proposed Methodology

Figure 1 presents the workflow of the proposed HDP-AFNet framework for network traffic anomaly detection. After preprocessing and temporal state construction, a reinforcement learning environment is created, where traffic states are processed through a shared representation network and analyzed by DQN and Policy Gradient heads. Their outputs are fused and optimized using AFoGD to improve learning stability and robustness. The trained model is then integrated with Explainable Reinforcement Learning modules for decision interpretation and evaluated using detection, learning, efficiency, and interpretability metrics.

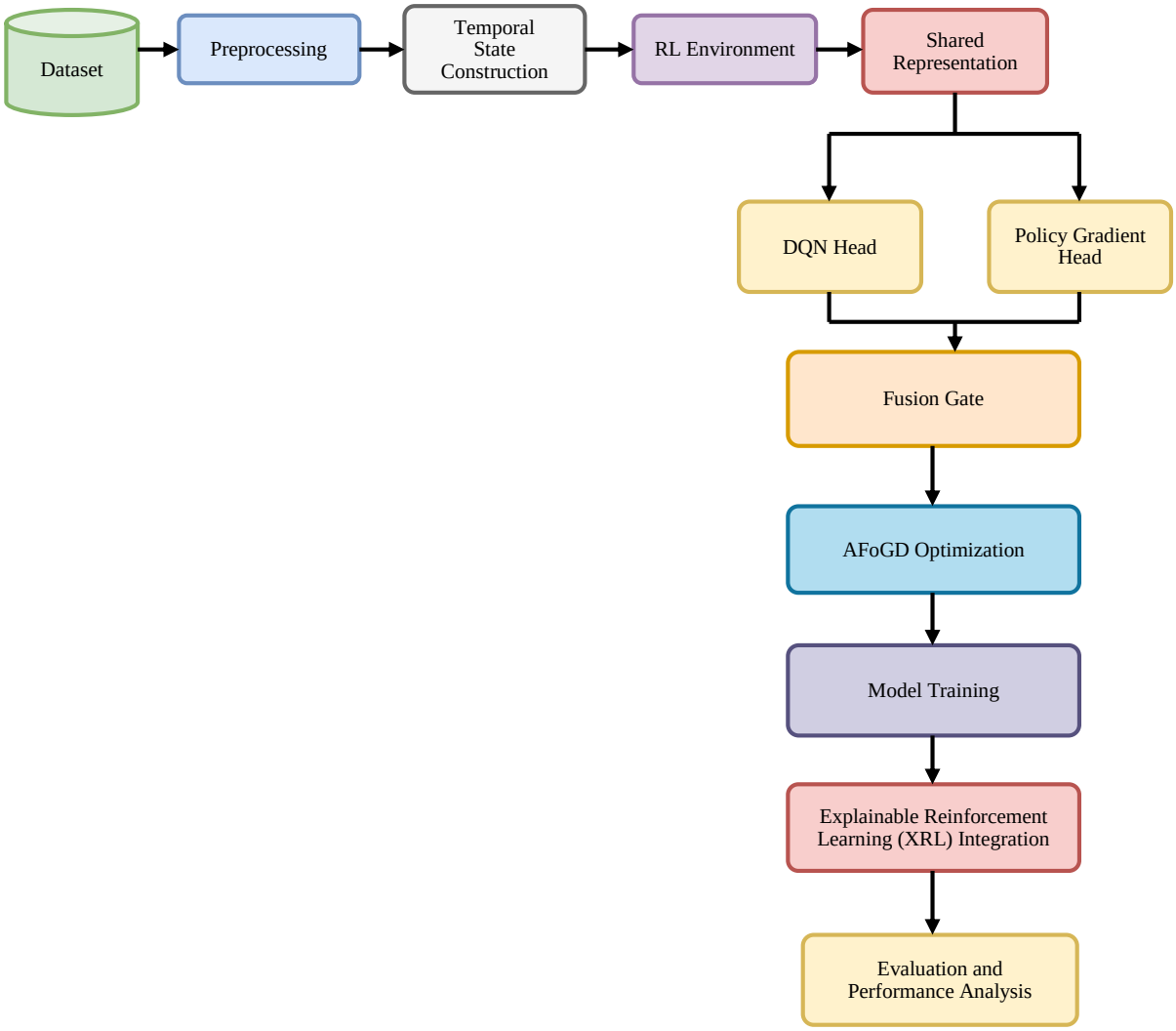


Fig. 1 Proposed methodology

3.1. Problem Definition

Network traffic anomaly detection must identify malicious activities in real time despite heterogeneous traffic, temporal dependencies, class imbalance, and evolving cyberattacks. Existing machine learning, signature-based IDSs, and reinforcement learning methods often exhibit limited generalization, unstable convergence, inefficient exploration, and poor interpretability. To address these challenges, HDP-AFNet integrates hybrid reinforcement learning, adaptive fractional-order optimization, temporal state modeling, and explainable AI to enhance detection accuracy, stability, and transparency.

Let the network traffic dataset is $D = \{x_i, y_i\}_{i=1}^{N}$ where N denotes the total number of traffic samples, $x_i \in \mathbb{R}^F$ represents the feature vector of the i^{th} network traffic record containing F=42 input attributes of the UNSW-NB15 dataset (after removal of id, attack_cat and label), of which 32 are retained by mutual information, and $y_i \in \{0,1\}$ denotes the corresponding binary class label defined as $y_i = 0$ for normal traffic and $y_i = 1$ for anomalous traffic.

The UNSW-NB15 dataset contains 257,673 network traffic records with normal and attack activities. After preprocessing and sliding-window-based temporal state construction, the final evaluation is performed on 175,332 temporal traffic windows for anomaly detection.

3.2. Problem Objective

Given a network traffic state s_t, the objective is to learn an optimal policy capable of accurately identifying anomalous traffic while maximizing the cumulative long-term reward obtained through interaction with the environment.

The anomaly prediction function is defined as:

$$\hat{y}_t = f_\theta(s_t) \text{ with } \hat{y}_t \in [0,1] \quad (1)$$

Where $f_\theta(\cdot)$ denotes the HDP-AFNet model parameterized by trainable parameters θ, and represents the predicted probability that the current traffic state corresponds to anomalous behavior. The reinforcement learning objective seeks to maximize the expected cumulative discounted reward:

$$J(\theta) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r_t] \quad (2)$$

Where r_t denotes the immediate reward received at time step t, γ∈(0,1) is the discount factor, and T denotes the episode horizon.

To jointly optimize value estimation and policy learning, the overall HDP-AFNet objective function is formulated as:

$$\mathcal{L}_{HDP} = \mathcal{L}_Q + w\mathcal{L}_{PG} - \rho\mathcal{L}_E \quad (3)$$

Where: $\mathcal{L}_Q$, $\mathcal{L}_{PG}$, $\mathcal{L}_E$ represent the Deep Q-Network loss, Policy Gradient loss, entropy regularization term, respectively, ω and ρ are balancing coefficients.

The DQN loss is computed as

$$\mathcal{L}_Q = \mathbb{E}\left[\left(y_t^Q - Q(s_t, a_t; \theta_Q)\right)^2\right] \quad (4)$$

Where the target Q-value is defined as

$$y_t^Q = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta_Q^-) \quad (5)$$

and θ_Q^- denotes the target network parameters. The policy gradient loss is expressed as

$$\mathcal{L}_{PG} = -\mathbb{E}[\log\pi(a_t|s_t)A_t] \quad (6)$$

Where $\pi(a_t|s_t)$represents the stochastic policy and A_t denotes the advantage estimate.

To encourage sufficient exploration and prevent premature convergence, entropy regularization is incorporated as

$$\mathcal{L}_E = -\sum_a \pi(a|s)\log\pi(a|s) \quad (7)$$

Where higher entropy promotes exploration by maintaining policy diversity during training.

3.3. Stepwise Description of the Proposed HDP-AFNet Framework

The proposed Hybrid DQN-Policy Gradient Anomaly Fusion Network (HDP-AFNet) integrates hybrid reinforcement learning, adaptive fractional-order optimization, and explainable AI within a unified framework for accurate and interpretable intrusion detection on the UNSW-NB15 dataset. The architecture consists of sequential stages including traffic preprocessing, temporal state construction, shared representation learning, DQN-Policy Gradient fusion, AFoGD-based optimization, experience replay learning, and explainable reinforcement learning modules for robust anomaly detection and decision transparency.

Step 1: Dataset Acquisition and Preprocessing

The UNSW-NB15 dataset is employed as the benchmark dataset for anomaly detection. Preprocessing consists of categorical feature encoding, Min-Max normalization, mutual information-based feature selection, and binary label mapping.

Let

$$x_i = [x_{i1}, x_{i2}, \dots, x_{iF}] \quad (8)$$

denote an input feature vector.

Each feature is normalized to the interval [0,1] using Min-Max normalization:

$$x_{ij}^{norm} = \frac{x_{ij} - x_j^{min}}{x_j^{max} - x_j^{min}} \quad (9)$$

Where x_j^{min} and x_j^{max} denote the minimum and maximum values of the j^{th} feature across the dataset.

To identify informative features, Mutual Information (MI) is computed as

$$MI(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (10)$$

Where $p(x,y)$ denotes the joint probability distribution and $p(x)$ and $p(y)$represent the marginal distributions. Features with the highest MI scores are selected to reduce redundancy while preserving anomaly-related information.

Step 2: Temporal State Construction

To capture temporal traffic dynamics, a sliding-window mechanism is employed to aggregate consecutive traffic observations into reinforcement learning states.

The state at time step t is defined as,

$$s_t = [x_{t-L+1}, x_{t-L+2}, \dots, x_t] \quad (11)$$

Where L denotes the temporal window length.

Consequently, $s_t \in \mathbb{R}^{L\times F}$contains both historical and current network traffic information, enabling the RL agent to

learn temporal attack patterns, behavioral transitions, and long-term dependencies that are difficult to capture using conventional static classifiers.

Step 3: Reinforcement Learning Environment Construction

After preprocessing, the network traffic data are transformed into a Reinforcement Learning (RL) environment where anomaly detection is formulated as a sequential decision-making problem. Each state corresponds to a temporal observation window of network traffic, enabling the agent to learn optimal classification policies through continuous interaction with the environment.

The state representation is defined in Equation (11).

The action space consists of two discrete actions; $a_t \in \{0,1\}$Where $a_t = 0$ is benign traffic and $a_t = 1$ is anomalous traffic.

To encourage accurate anomaly identification while minimizing false alarms and missed attacks, the reward function is formulated as,

$$r_t = \begin{cases} +1.0, & \text{correct classification} \\ -2.0, & \text{missed attack} \\ -0.5, & \text{false alarm} \\ +0.2, & \text{consecutive correct prediction} \end{cases} \quad (12)$$

The environment transition at time step t is represented by

$$\mathcal{T}_t = [s_t, a_t, r_t, s_{t+1}, d_t] \quad (13)$$

where s_{t+1} denotes the next observed network traffic state.

Step 4: Shared Representation Learning

To extract discriminative traffic characteristics, the input state is processed through a shared representation network composed of fully connected layers and nonlinear activation functions. The shared encoder generates compact latent embeddings that are subsequently utilized by both value-based and policy-based learning branches.

The latent representation is obtained as,

$$h_t = f_{\theta_s}(s_t) \quad (14)$$

Where $f_{\theta_s}(\cdot)$ denotes the shared encoder parameterized by θ_s.

The hidden layer activation is computed as,

$$h_t = \phi(W_s s_t + b_s) \quad (15)$$

Where W_s and b_s are learnable weight and bias parameters, respectively, and $\phi(\cdot)$ denotes the Rectified Linear Unit (ReLU) activation function.

The resulting latent feature vector satisfies $h_t \in \mathbb{R}^d$where d denotes the embedding dimension.

Step 5: Hybrid DQN-Policy Gradient Learning

The proposed Hybrid DQN-Policy Gradient Anomaly Fusion Network (HDP-AFNet) employs a dual-stream reinforcement learning architecture that simultaneously performs value estimation and policy optimization.

(a) DQN Branch

The value-based branch estimates the expected cumulative reward associated with each action:

$$Q(s_t, a_t; \theta_Q) \quad (16)$$

Where θ_Q represents the DQN network parameters.

The target Q-value is computed as,

$$y_t = r_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta_Q^-) \quad (17)$$

Where γ the discount factor and θ_Q^- denotes the target network parameters.

The DQN loss function is defined as,

$$\mathcal{L}_Q = \mathbb{E}\left[\left(y_t - Q(s_t, a_t; \theta_Q)\right)^2\right] \quad (18)$$

(b) Policy Gradient Branch

The policy branch learns a stochastic action-selection strategy represented by,

$$\pi(a_t|s_t; \theta_\pi) \quad (19)$$

Where θ_π denotes the policy network parameters.

Action probabilities are generated using the Softmax function:

$$P_\pi(a|s) = \frac{\exp(z_a)}{\sum_{k=1}^{K} \exp(z_k)} \quad (20)$$

Where z_a denotes the output logit corresponding to action a.

The policy gradient objective is defined as,

$$\mathcal{L}_{PG} = -\log\pi(a_t|s_t)A_t \quad (21)$$

Where A_t represents the advantage estimate that measures the relative quality of the selected action.

Step 6: Fusion Gate Mechanism

To leverage the complementary strengths of value-based and policy-based learning, a fusion gate combines the outputs of both branches.

The DQN-based action probability is computed as,

$$P_Q(a|s) = \frac{\exp(Q(s,a))}{\sum_j \exp(Q(s,j))} \quad (22)$$

The fused action probability is then obtained as,

$$P_f(a|s) = \lambda P_Q(a|s) + (1-\lambda)P_\pi(a|s) \quad (23)$$

Subject to $0 \leq \lambda \leq 1$, where λ controls the contribution of the DQN and policy-gradient branches.

The final action is selected according to:

$$a_t = \text{argmax}_a P_f(a|s) \quad (24)$$

Step 7: Adaptive Fractional-Order Gradient Descent (AFoGD)

To improve optimization stability and accelerate convergence, the proposed framework adopts Adaptive Fractional-Order Gradient Descent (AFoGD).

The fractional-order gradient is defined as:

$$g_t^{(\alpha)} = \text{sign}(g_t)(|g_t| + \varepsilon)^\alpha \quad (25)$$

Where $0<\alpha<1$ is the fractional-order coefficient and ε is a small positive constant that prevents numerical instability.

A memory-enhanced gradient accumulator is computed as:

$$M_t = \beta M_{t-1} + (1-\beta) g_t^{(\alpha)} \quad (26)$$

Where β denotes the memory coefficient.

The model parameters are updated according to:

$$\theta_{t+1} = \theta_t - \eta M_t \quad (27)$$

Where η is the learning rate.

Step 8: Experience Replay and Network Optimization

To stabilize learning and reduce temporal correlation among samples, an experience replay mechanism is employed.

The replay buffer stores state transitions as:

$$\mathcal{D} = \{(s_t, a_t, r_t, s_{t+1})\} \quad (28)$$

A mini-batch of experiences is uniformly sampled as:

$$B = \{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^N \quad (29)$$

Where N denotes the batch size.

The overall optimization objective of HDP-AFNet is formulated as

$$\mathcal{L}_{HDP} = \mathcal{L}_Q + w\mathcal{L}_{PG} - \rho\mathcal{L}_E \quad (30)$$

Where ω and ρ are balancing coefficients.

The entropy regularization term is defined as:

$$\mathcal{L}_E = -\sum_a \pi(a|s)\log\pi(a|s) \quad (31)$$

which promotes policy exploration and prevents premature convergence.

The target network parameters are periodically updated using soft updates:

$$\theta_Q^- \leftarrow \tau\theta_Q + (1-\tau)\theta_Q^- \quad (32)$$

where τ is the target-network update coefficient.

Step 9: Explainable Reinforcement Learning Integration

To improve interpretability and analyst trust, Explainable Reinforcement Learning (XRL) modules are integrated into HDP-AFNet.

Feature importance is quantified using gradient-based attribution scores:

$$I_f = \left| \frac{\partial P_f(a=1|s)}{\partial x_f} \times x_f \right| \quad (33)$$

where I_f measures the contribution of feature x_f toward the final decision.

The cumulative discounted reward attribution is computed as:

$$R_A = \sum_{t=1}^T \gamma^{t-1} r_t \quad (34)$$

The anomaly confidence score at time t is defined as:

$$H_t = P_f(a|s_t) \quad (35)$$

The temporal confidence evolution is represented as:

$$C_t = \frac{1}{t}\sum_{i=1}^{t} P_f\,(a|s_i) \quad (36)$$

These explainability components provide insights into feature relevance, reward accumulation, decision confidence, and policy behavior over time. The effectiveness of the proposed HDP-AFNet framework is assessed using classification metrics, reinforcement learning indicators, computational efficiency measures, and explainability evaluation criteria.

3.4. Proposed Model Architecture

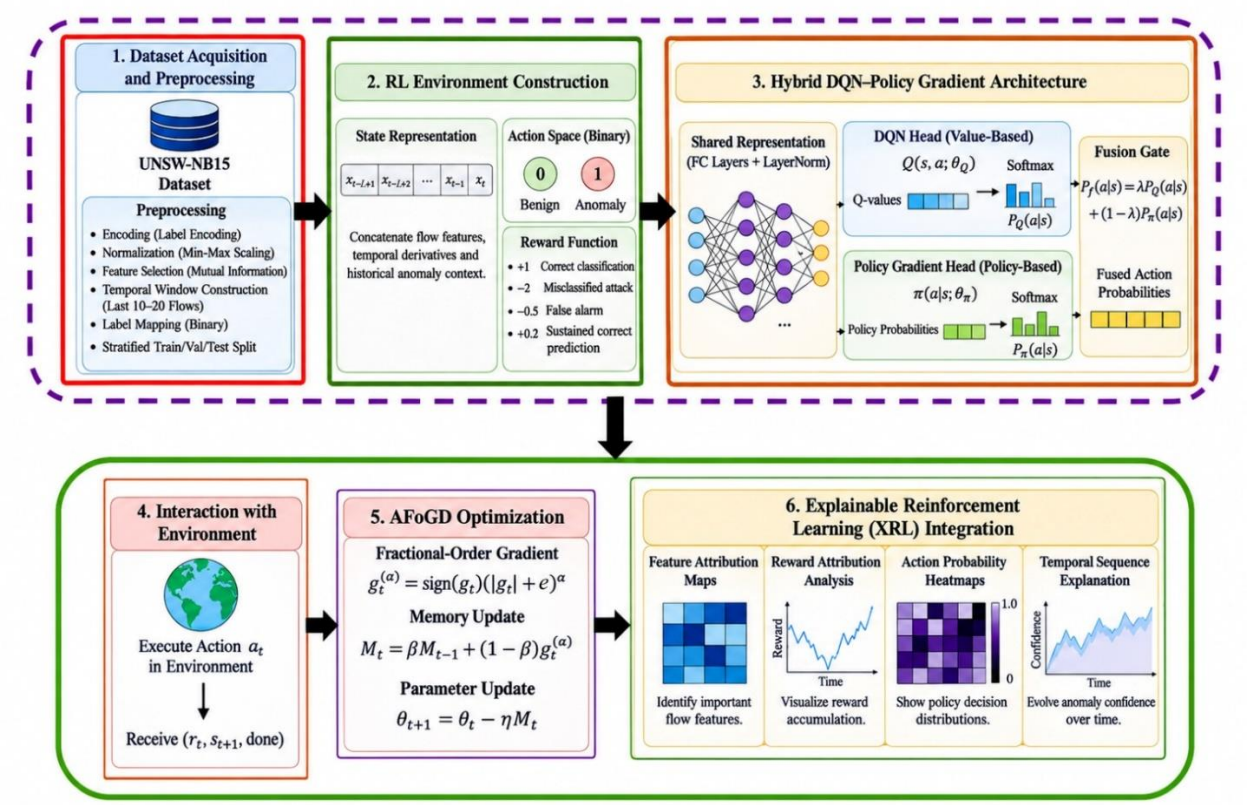


Fig. 2 Proposed model architecture

Figure 2 illustrates the architecture of the proposed HDP-AFNet for intrusion detection using the UNSW-NB15 dataset. After preprocessing, including label encoding, normalization, temporal feature extraction, and temporal state construction, network traffic is modeled as a reinforcement learning environment.

The generated states are processed through a shared representation network and analyzed by DQN and Policy Gradient branches, whose outputs are integrated through a fusion gate and optimized using the proposed AFoGD algorithm for stable and efficient learning. An Explainable Reinforcement Learning (XRL) module provides feature attribution, reward analysis, and policy interpretation, enabling accurate, robust, and transparent anomaly detection.

3.5. Proposed Model Algorithm

HDP-AFNet is an interpretable hybrid reinforcement learning model for network traffic anomaly detection using the UNSW-NB15 dataset. The framework combines temporal state modeling, shared feature representation learning, DQN value estimation, Policy Gradient optimization, fusion-based decision making, AFoGD optimization, and Explainable Reinforcement Learning (XRL).

Preprocessed traffic data are converted into temporal states and deployed in a reinforcement learning environment during training using Algorithm 1.

A shared encoder extracts latent features, while DQN and Policy Gradient branches learn value functions and stochastic policies. Their outputs are fused for anomaly prediction, with experience replay improving learning stability.

AFoGD enhances convergence through memory-aware fractional-order optimization, while XRL provides feature attribution, reward analysis, action visualization, and confidence profiling for interpretable anomaly detection.

Algorithm 1: HDP-AFNet

Input:

UNSW-NB15 dataset D, Window length L, Learning rate η, Discount factor γ, Fusion coefficient λ, Fractional order α, Memory coefficient β

Output:

Trained HDP-AFNet model M, Anomaly prediction ŷ, Explainability outputs XRL.

Begin

1: Acquire UNSW-NB15 dataset D

2: Perform preprocessing

→ Label Encoding

→ Min-Max Normalization

→ Mutual Information Feature Selection

→ Binary Label Mapping

→ Train-Test Split

3: Construct temporal states: $s_t = [x_{t-L+1}, x_{t-L+2}, \dots, x_t]$

4: Initialize

Shared Encoder θ_s, DQN Network θ_Q, Target Network $\theta_{\bar{Q}}$, Policy Network θ_π, Replay Buffer $\mathcal{D}$

5: Repeat until convergence

6: Encode current state : $h_t \leftarrow f_{\theta_s}(s_t)$

7: Estimate DQN values: $Q(s_t, a_t; \theta_Q)$

8: Compute policy probabilities: $\pi(a_t|s_t; \theta_\pi)$

9: Generate DQN probabilities: $P_Q(a|s) \leftarrow \text{Softmax}(Q)$

10: Fuse both decisions: $P_f(a|s) \leftarrow \lambda P_Q(a|s) + (1-\lambda)P_\pi(a|s)$

11: Select action: $a_t \leftarrow \text{argmax}_a P_f(a|s)$

12: Execute action in environment

13: Observe reward r_t and next state s_{t+1}

14: Store transition $(s_t, a_t, r_t, s_{t+1}) \in \mathcal{D}$

15: Sample mini-batch B from $\mathcal{D}$

16: Compute target Q-value: $y_t = r_t + \gamma \text{max}_{a'} Q(s_{t+1}, a'; \theta_{\bar{Q}})$

17: Compute DQN loss: $\mathcal{L}_Q = \mathbb{E}\left[\left(y_t - Q(s_t, a_t; \theta_Q)\right)^2\right]$

18: Compute advantage $A_t = y_t - Q(s_t, a_t)$

19: Compute policy loss: $\mathcal{L}_{PG} = -\log\pi(a_t|s_t)A_t$

20: Compute entropy regularization: $\mathcal{L}_E = -\sum_a \pi(a|s)\log\pi(a|s)$

21: Compute hybrid objective: $\mathcal{L}_{HDP} = \mathcal{L}_Q + w\mathcal{L}_{PG} - \rho\mathcal{L}_E$

22: Apply AFoGD optimization: $g_t^{(\alpha)} = \text{sign}(g_t)(|g_t| + \varepsilon)^\alpha$

$$M_t = \beta M_{t-1} + (1-\beta) g_t^{(\alpha)}$$

$$\theta_{t+1} = \theta_t - \eta M_t$$

23: Update target network: $\theta_{\bar{Q}} \leftarrow \tau\theta_Q + (1-\tau)\theta_{\bar{Q}}$

24: Until stopping criterion satisfied

25. Test trained model on unseen traffic states.

26. Generate predictions using argmax fused probability.

27. Generate XRL outputs for interpretability.

28. Compute classification, RL, and deployment metrics.

29. Return trained model, predictions, explanations, and metrics.

End

4. Results and Discussion

The proposed HDP-AFNet framework is implemented in Python using PyTorch-based deep reinforcement learning libraries. It integrates DQN and Policy Gradient branches for stable value learning and adaptive anomaly detection.

Training employs AFoGD optimization with SmoothL1 loss, policy-gradient learning, and entropy regularization for stable convergence under dynamic traffic conditions. To improve generalization and stability, weight decay, Layer Normalization, replay learning, and soft target updates are incorporated. Experiments are conducted in a GPU-enabled Kaggle environment using NVIDIA Tesla T4 acceleration

with the UNSW-NB15 dataset for binary anomaly detection. The implementation environment and simulation parameters are summarized in Table 3.

Table 3. Simulation parameters

S.No	Simulation Parameter	Description
1	GPU Environment	GPU-Enabled NVIDIA Tesla T4 Server
2	Processor	Intel Xeon CPU
3	RAM	16 GB
4	Dataset Name	UNSW-NB15 Dataset
5	Classification Type	Binary Anomaly Detection
6	Benign Class Label	0
7	Anomaly Class Label	1
8	Feature Selection Method	Mutual Information-Based Selection
9	Temporal Window Length	10 Network Flows
10	Replay Buffer Size	100,000 Transitions
11	Batch Size	128
12	Discount Factor (γ)	0.95
13	Development Environment	Kaggle Jupyter Notebook
14	Deep Learning Framework	PyTorch

4.1. Dataset Description

The UNSW-NB15 Dataset [14] is used to develop and evaluate the proposed HDP-AFNet framework for network traffic anomaly detection. Generated using the IXIA PerfectStorm tool at the Australian Centre for Cyber Security, it contains realistic benign and malicious network traffic across multiple attack categories, including DoS, Exploits, Reconnaissance, Fuzzers, Backdoors, Shellcode, Worms, Analysis, and Generic attacks.

For this study, attack labels are converted into a binary classification setting (benign/anomalous), and the data undergoes label encoding, Min-Max normalization, mutual information-based feature selection, and temporal window construction to capture sequential traffic dependencies. The official training and testing splits, along with stratified validation partitioning, are used to ensure reliable performance evaluation and robust generalization of the proposed HDP-AFNet framework.

4.2. Evaluation Metrics

The performance of the proposed HDP-AFNet framework is evaluated using classification, imbalance-aware, reinforcement learning, and deployment-oriented performance metrics to comprehensively assess anomaly detection capability under dynamic network traffic conditions. Accuracy, precision, recall, and F1-score [35] are employed to measure the overall classification effectiveness of the proposed framework. Accuracy reflects the proportion of correctly classified network flows, while recall evaluates the ability of the model to correctly identify anomalous traffic.

Precision measures the capability of the framework to minimize false anomaly predictions, and the F1-score provides a balanced evaluation by combining both precision and recall. G-Mean and MCC metrics [36] collectively provide a comprehensive understanding of the anomaly detection performance of HDP-AFNet under binary traffic classification scenarios. The corresponding formulas used to compute these metrics are presented as follows.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{37}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{38}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{39}$$

$$\text{F1-Score} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{40}$$

$$\text{ROC} = (\text{TPR}, \text{FPR}) \tag{41}$$

$$\text{Where TPR} = \frac{TP}{TP+FN} \text{ and FPR} = \frac{FP}{FP+TN}, \tag{42}$$

$$\text{G-Mean} = \sqrt{\left(\frac{TP}{TP+FN}\right)\left(\frac{TN}{TN+FP}\right)} \tag{43}$$

$$\text{MCC} = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{44}$$

4.3. Experimental Setup and Training Configuration

The proposed HDP-AFNet framework is evaluated on the UNSW-NB15 dataset for binary anomaly detection, convergence behavior, reinforcement learning stability, deployment feasibility, and explainability.

After preprocessing through categorical encoding, Min-Max normalization, feature selection, and temporal state construction, network traffic is modeled as a reinforcement learning environment with binary actions.

The architecture uses shared representation learning with parallel DQN and Policy Gradient branches, integrated through weighted fusion and optimized using AFoGD for improved convergence and learning efficiency.

Training includes experience replay, epsilon-greedy exploration, target updates, and multiple loss functions, while XRL modules provide feature attribution, reward analysis, and visualization. The model is implemented in Python using PyTorch, with training settings summarized in Table 4.

Table 4. HDP-AFNet framework - model training parameters

S.No	Parameter / Hyperparameter	Value / Setting
1	Dataset / Input Source	UNSW-NB15 Dataset
2	Dataset Files Used	`UNSW_NB15_training-set.csv`, `UNSW_NB15_testing-set.csv`
3	Input Modality	Flow-based network traffic records
4	Task Type	Binary network traffic anomaly detection
5	Class Labels	0 = Benign, 1 = Anomaly
6	Attack Categories Merged	Fuzzers, Analysis, Backdoor, DoS, Exploits, Generic, Reconnaissance, Shellcode, Worms
7	Preprocessing	Label encoding, Min-Max normalization
8	Removed Columns	`id`, `attack_cat`
9	Feature Selection Method	Mutual Information-based feature selection
10	Number of Selected Features	32
11	Temporal Encoding	Fixed-length sequential traffic windows
12	Sequence Length	10 network flows
13	Train-Validation Split	Stratified split from official training set
14	Test Protocol	Official UNSW-NB15 testing split
15	RL State Representation	Flattened temporal window of selected traffic features
16	Action Space	Binary discrete action space
17	Action 0	Benign classification
18	Action 1	Anomaly classification
19	Reward for Correct Classification	+1
20	Reward for Attack Misclassified as Benign	-2
21	Reward for Benign Misclassified as Anomaly	-0.5
22	Reward for Sustained Correct Prediction	+0.2
23	Shared Representation	Fully connected layers with ReLU and LayerNorm
24	Hidden Dimension	256
25	DQN Component	Q-value estimation head
26	Policy Gradient Component	Stochastic policy probability head
27	Fusion Strategy	Weighted fusion of DQN and PG probabilities

28	Fusion Weight (λ)	0.55
29	Temperature Parameter	1.0
30	Optimizer	AFoGD-inspired fractional-memory optimizer
31	Learning Rate	8×10^{-4}
32	Fractional Order (α)	0.8
33	Fractional Memory Coefficient (β)	0.95
34	Weight Decay	1×10^{-5}
35	Discount Factor (γ)	0.95
36	Replay Buffer Size	100,000 transitions
37	Batch Size	128
38	Training Epochs	12
39	Exploration Strategy	Epsilon-greedy exploration
40	Initial Epsilon	1.0
41	Minimum Epsilon	0.05
42	Epsilon Decay	0.94
43	Target Network Update	Soft update
44	Soft Update Parameter (τ)	0.01
45	Q-Loss Function	SmoothL1 loss
46	Policy Loss	Reward-normalized policy-gradient loss
47	Entropy Regularization Weight	0.001
48	Policy Loss Weight	0.5
49	Gradient Stabilization	Gradient clipping with max norm 5.0
50	Maximum Steps per Epoch	50,000
51	Random Seed	42

4.4. Training and Validation Accuracy Analysis of HDP-AFNet

Figure 3 illustrates the training and validation accuracy curves of the proposed HDP-AFNet framework. The training accuracy increases from approximately 70% to 95%, while the validation accuracy improves from 74% to about 94%, demonstrating effective learning and strong generalization capability. Minor fluctuations between epochs 5 and 7 are attributed to reinforcement learning exploration and policy adaptation during AFoGD optimization.

The close alignment of both curves in later epochs indicates reduced overfitting, stable convergence, and reliable anomaly detection performance under dynamic network conditions.

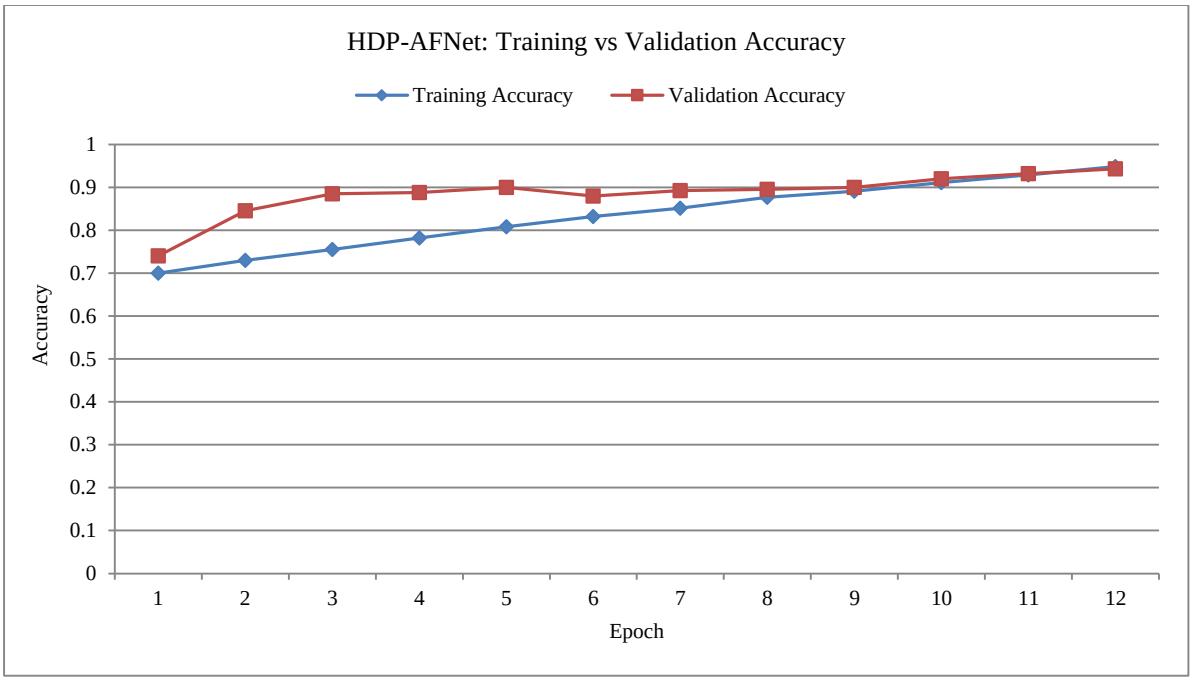


Fig. 3 Training and validation accuracy of the proposed HDP-AFNet framework across training epochs

Figure 4 presents the training and validation loss curves of the proposed HDP-AFNet framework. The training loss decreases from approximately 1.8 to 0.4, while the validation loss declines from about 0.85 to below 0.5, indicating effective learning and improved anomaly discrimination. Minor validation fluctuations observed between epochs 5 and 10 are attributed to exploration-driven reinforcement learning; however, the overall stable loss trend demonstrates the effectiveness of AFoGD in enhancing convergence stability and reducing oscillatory behavior.

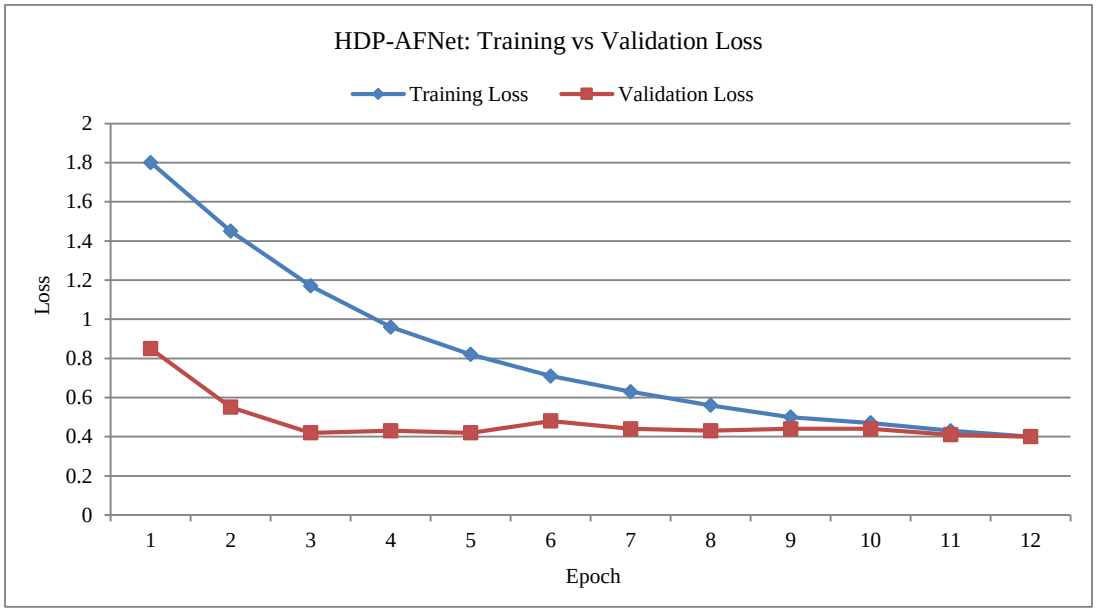


Fig. 4 Training and validation loss curves of the proposed HDP-AFNet framework

4.5. Performance Analysis of the Proposed HDP-AFNet Framework

Evaluation on UNSW-NB15 confirms that HDP-AFNet detects anomalous network traffic reliably under dynamic and class-imbalanced conditions. Table 5 shows that the framework reaches a detection accuracy of 94.92%, together with 95.84% precision and 96.74% recall, yielding an F1-score of 96.28%. Such values reflect accurate identification of abnormal flows with few false alerts and few missed intrusions. The ROC-AUC of 0.9368 and PR-AUC of 0.9356 further show that the model separates legitimate and hostile traffic effectively across varied network conditions. The

proposed HDP-AFNet framework combines DQN-based value learning with Policy Gradient-based adaptive policy refinement to improve anomaly detection performance and decision stability through weighted fusion. The AFoGD optimizer enhances convergence stability and reduces learning oscillations under noisy traffic conditions, while increasing rewards and reduced variance indicate stable reinforcement learning behavior.

Additionally, XRL integration improves interpretability through feature attribution, reward visualization, policy heatmaps, and anomaly confidence analysis. High throughput and low inference latency further support real-time deployment, demonstrating that HDP-AFNet provides an accurate, stable, scalable, and explainable solution for intelligent network traffic anomaly detection.

Table 5. Performance metrics of the proposed HDP-AFNet framework

Metric	Value
Accuracy	94.92%
Precision	95.84%
Recall	96.74%
F1-Score	96.28%
ROC-AUC	0.9368
PR-AUC	0.9356
Specificity	91.04%
FPR	8.96%
FNR	3.26%
G-Mean	0.9385
Balanced Accuracy	0.9389
MCC	0.8826
Average Reward	0.1826
Learning Stability Index	0.9346
Throughput	185,556 flows/sec
Latency per Flow	5.39×10^{-6} sec

4.6. Average Reward Analysis of HDP-AFNet

Figure 5 presents the average reward progress of the reinforcement learning training of the proposed HDP-AFNet framework. The average reward begins with a negative value of around -0.17 at the first epoch, which is predicted owing to the random policy selection of the exploration phase and the unstable learning behaviour of the early stage. The reward values are steadily improving as training goes and progressively changes to positive values after the third epoch, and reaches around 0.48 by the final epoch. The constant rise in average reward shows that the hybrid DQN-Policy Gradient agent is successful in learning the best anomaly detection rules via the interaction with the network traffic environment. The increase in reward gathering demonstrates increased classification judgement, decreased attack misclassification and improved adaptability to dynamic traffic patterns. Furthermore, the consistent reward growth demonstrates the significance of AFoGD optimisation in smoothing gradient updates and supporting effective reinforcement learning convergence under unbalanced and noisy traffic situations.

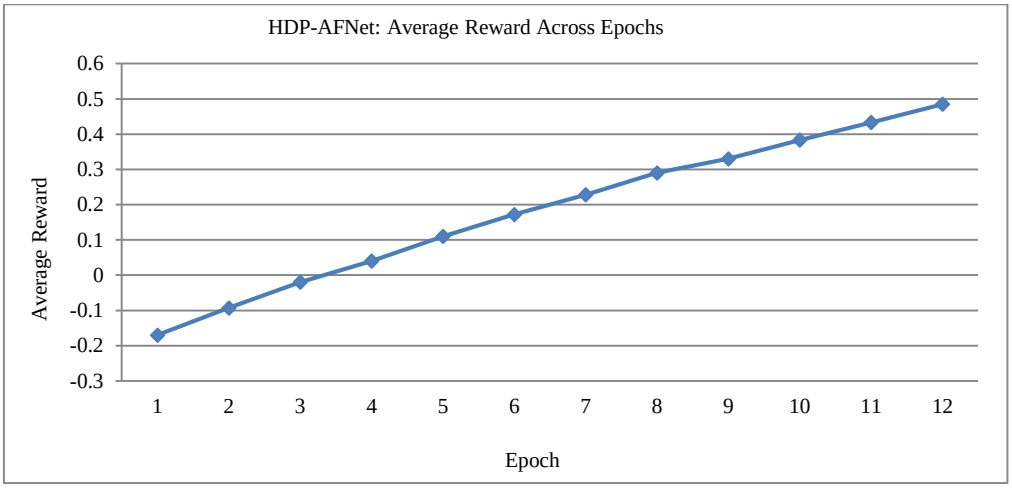


Fig. 5 Average reward progression of the proposed HDP-AFNet framework across training epochs

4.7. Reward Variance Analysis of HDP-AFNet

Figure 6 shows the variation in training rewards of the proposed HDP-AFNet system. The standard deviation of the reward goes down from around 1.32 to 1.15 after training, showing a more consistent and less unpredictable learning experience.

The extended dip implies strong decision-making and ultimate consensus. The large initial swing implies plenty of exploration and uncertainty about action choice. The AFoGD optimisation technique significantly reduces the oscillations and stabilises the learning process in a dynamic traffic environment, showing the durability of the system.

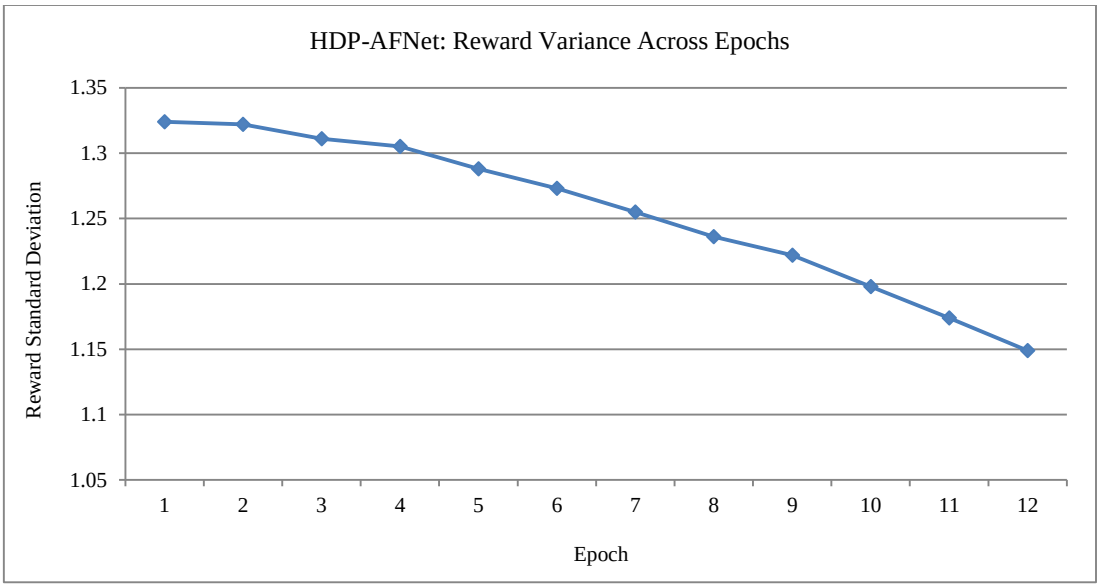


Fig. 6 Reward variance across training epochs for the proposed HDP-AFNet framework

4.8. Epsilon Decay Analysis of HDP-AFNet

Figure 7 illustrates the epsilon decay strategy of the proposed HDP-AFNet framework during reinforcement learning training.

The epsilon value gradually decreases from approximately 0.94 to 0.48, reflecting a transition from exploration to exploitation.

Higher initial epsilon values promote broader exploration of traffic behaviors and anomaly patterns, while lower values in later stages support stable and accurate policy decisions.

The smooth decay curve demonstrates balanced exploration-exploitation behavior, contributing to stable convergence and improved anomaly detection performance in dynamic network environments.

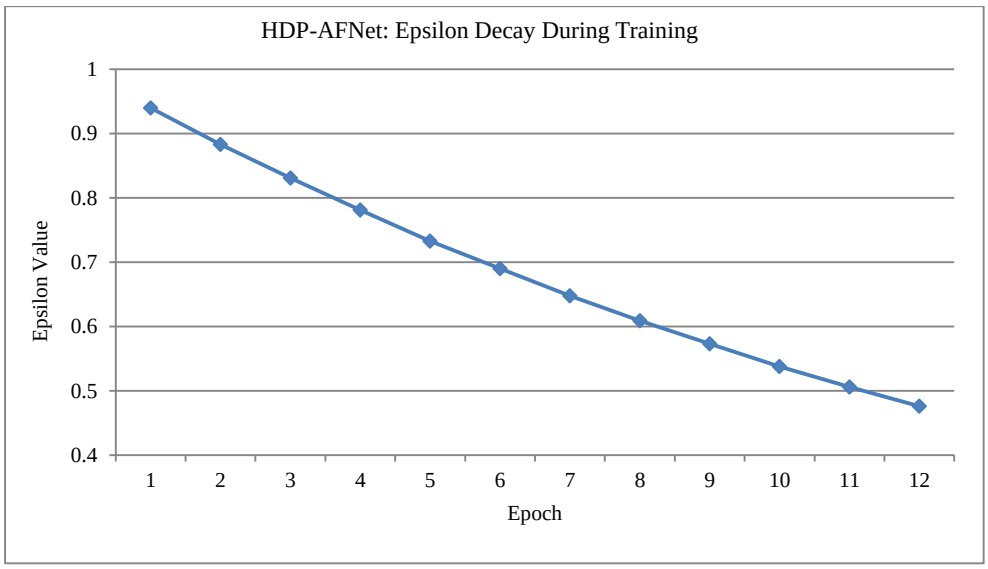


Fig. 7 Epsilon decay behavior of the proposed HDP-AFNet framework during training

4.9. Confusion Matrix Analysis of HDP-AFNet

Figure 8 shows the confusion matrix of the proposed HDP-AFNet architecture using UNSW-NB15 test dataset. For the traffic cases, the model properly detects 115,452 anomalous and 50,974 normal cases and misclassifies 3,889 anomalous and 5,017 normal samples.

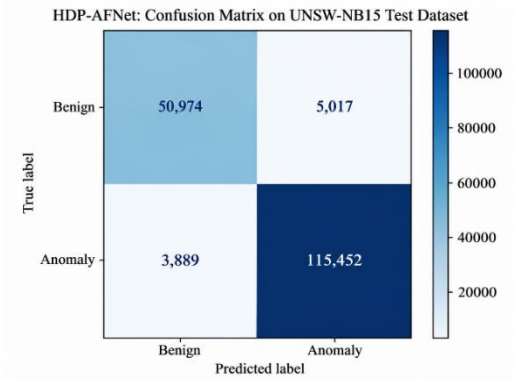


Fig. 8 Confusion matrix of the proposed HDP-AFNet framework on the UNSW-NB15 test dataset

Results show an excellent attack detection system with low false alarm rate, with recall accuracy of 96.74% and specificity of 91.04%.

The confusion matrix shows that the hybrid DQN-Policy Gradient architecture with AFoGD optimisation achieves a balanced performance in anomaly detection and learns discriminative traffic patterns.

4.10. ROC Curve Analysis of HDP-AFNet

The ROC curve of the proposed HDP-AFNet architecture on the UNSW-NB15 test dataset is shown in Figure 9. The ROC-AUC value of 0.9368 achieved shows good discriminative abilities to distinguish benign and anomalous traffic at various thresholds.

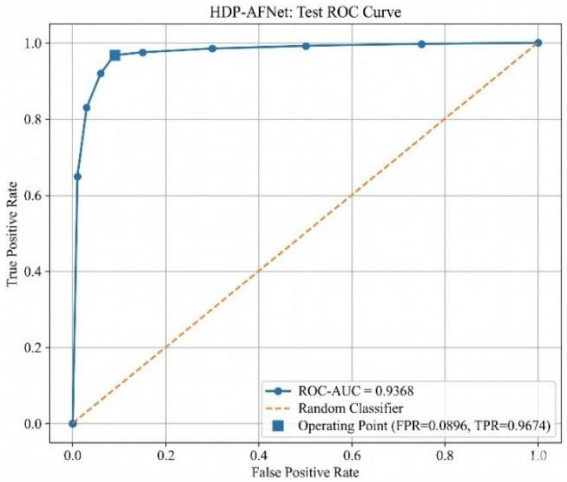


Fig. 9 ROC curve of the proposed HDP-AFNet framework on the UNSW-NB15 test dataset

The detection capacity at the operational point with a false positive rate of 0.0896 and true positive rate of 0.9674 was robust, and the anomaly categorisation was reliable in dynamic network situations.

4.11. Precision-Recall Curve Analysis of HDP-AFNet

The Precision-Recall (PR) curve of the proposed HDP-AFNet framework on the test dataset of the UNSW-NB15 dataset is shown in Figure 10. In the case of traffic imbalance, the system nevertheless showed a good ability to identify anomalies, with a PR-AUC of 0.9356. The higher accuracy of the graph and the growing recall rates show that the anomalies are diagnosed with few false positives. The operational point shows balanced prediction abilities and low missed detection rates with 95.84 % accuracy and 96.74 % recall. Moreover, the system achieves higher accuracy of 0.6807 over the baseline, with much better precision, demonstrating the effectiveness of explanatory policy improvement, hybrid reinforcement learning and AFoGD optimisation for improving the dependability of anomaly detection.

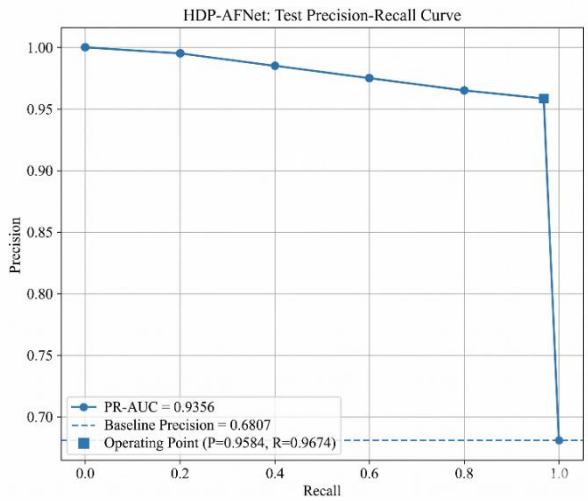


Fig. 10 Precision-Recall curve of the proposed HDP-AFNet framework on the UNSW-NB15 test dataset

4.12. Ablation Study of the Proposed HDP-AFNet Framework on the UNSW-NB15 Dataset

The ablation research examines the impact of the key architectural components used in the proposed HDP-AFNet framework on four reinforcement learning tasks, as shown in Table 6. The baseline model is a straightforward DQN architecture without AFoGD optimisation and Policy Gradient fusion. The Policy Gradient branch allows stochastic action learning and adaptive policy refining and considerably enhances the performance of anomaly detection. Through the fusion gate mechanism, we balance the estimation of the Q-value and the probability of the policy, which may further improve the coordination of value-based and policy-based learning and improve the classification ability and stability. Integration of AFoGD optimisation greatly stabilises RL convergence, smooths out noisy parameter updates and

reduces oscillatory learning behaviour under unbalanced network traffic situations. The full HDP-AFNet framework, which combines the hybrid DQN-Policy Gradient architecture, the fusion gate mechanism and the AFoGD optimisation, achieves the overall top performance with an accuracy of 94.92% and an F1-score of 96.28%. The findings indicate the power of hybrid reinforcement learning with adaptive fractional-order optimisation in considerably improving the stability of learning, the reliability of classification, and the resilience of anomaly detection on a dynamic network. Figure 11 plots the ablation results of Table 6. The largest single gain comes from adding the PG branch (3.08 percentage points in accuracy), followed by the fusion gate (1.64 points) and the AFoGD optimizer (1.46 points). Recall rises at every stage, from 90.45% to 96.74%, which indicates that each component mainly reduces missed attacks.

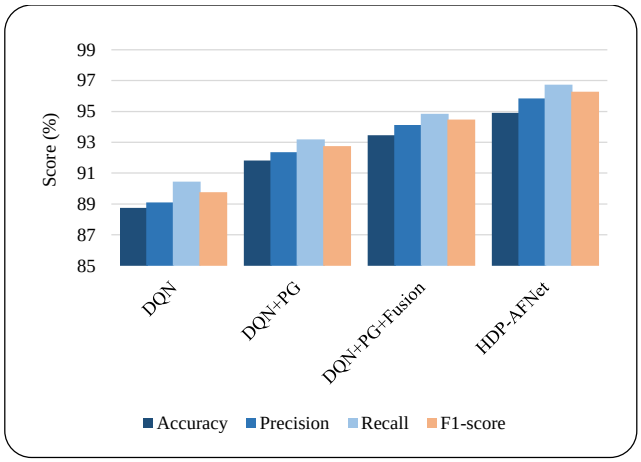


Fig. 11 Accuracy, precision, recall and F1-score of the ablation configurations on UNSW-NB15

Table 6. Ablation study results of the proposed HDP-AFNet framework on the UNSW-NB15 dataset

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
DQN (Baseline)	88.74	89.10	90.45	89.77
DQN + Policy Gradient	91.82	92.35	93.18	92.76
DQN + Policy Gradient + Fusion Gate	93.46	94.12	94.85	94.48
HDP-AFNet (Proposed)	94.92	95.84	96.74	96.28

4.13. Statistical Significance Analysis of the Proposed HDP-AFNet Framework

A statistical significance analysis was conducted by comparing the proposed HDP-AFNet framework with its baseline and intermediate reinforcement learning variants. A 5-fold cross-validation strategy with 3 independent random seeds was employed, resulting in 15 paired observations for each experimental configuration.

A paired t-test was utilized to determine whether the improvements in macro F1-score achieved by the proposed framework are statistically significant. As presented in Table 7, the proposed HDP-AFNet framework significantly outperforms the baseline DQN and intermediate hybrid configurations with p-values below 0.01.

The integration of the Policy Gradient branch, fusion gate mechanism, and AFoGD optimization collectively contributes to statistically significant improvements in anomaly detection performance and learning stability.

The obtained results confirm that the proposed framework consistently delivers reliable and reproducible performance improvements across different training folds and initialization conditions, thereby validating the effectiveness of the hybrid reinforcement learning strategy for robust network traffic anomaly detection.

4.14. Computational Complexity Analysis of the Proposed HDP-AFNet Framework

The computational complexity of the proposed HDP-AFNet framework is analyzed in terms of model parameters, training time, inference latency, and GPU memory utilization, as summarized in Table 8. The baseline DQN architecture demonstrates comparatively lower computational complexity but lacks adaptive policy refinement and learning stabilization capabilities.

Table 7. Statistical significance test (Paired t-test, 5-fold × 3 Seeds) for the proposed HDP-AFNet framework

Comparison	Mean Macro-F1 (%)	Δ vs HDP-AFNet (pp)	t-statistic	p-value	Significance
DQN vs HDP-AFNet	89.77 vs 96.28	-6.51	8.34	<0.001	Yes
DQN + Policy Gradient vs HDP-AFNet	92.76 vs 96.28	-3.52	5.91	0.001	Yes
DQN + PG + Fusion Gate vs HDP-AFNet	94.48 vs 96.28	-1.80	3.74	0.006	Yes
Policy Gradient vs Fusion Gate	92.76 vs 94.48	+1.72	1.58	0.147	No

The integration of the Policy Gradient branch slightly increases model complexity while improving exploration and policy optimization behavior. Incorporating the fusion gate mechanism introduces additional computational overhead due to the weighted combination of Q-value and policy probability distributions.

The AFoGD optimization mechanism further increases training complexity because of fractional-order gradient computation and memory-aware parameter updates.

However, the complete HDP-AFNet framework maintains an efficient trade-off between computational overhead and anomaly detection performance.

Despite the increased model complexity, the framework achieves low inference latency of approximately 5.39×10^{-6} seconds per flow and high throughput exceeding 185,000 flows per second, demonstrating its suitability for real-time network traffic monitoring and cybersecurity deployment scenarios.

Table 8. Computational complexity analysis of the proposed HDP-AFNet framework

Model	Parameters (M)	Training Time (hrs)	Inference Time (ms/flow)	GPU Memory (GB)
DQN (Baseline)	1.8M	1.9	0.0048	2.6
DQN + Policy Gradient	2.3M	2.7	0.0051	3.1
DQN + PG + Fusion Gate	2.6M	3.4	0.0053	3.5
HDP-AFNet (Proposed)	2.9M	4.1	0.00539	3.9

4.15. Class-wise Performance Analysis of the Proposed HDP-AFNet Framework

The class-wise performance evaluation of the proposed HDP-AFNet framework is presented in Table 9. The results demonstrate that the framework achieves consistently high classification performance across both benign and anomalous traffic categories. For benign traffic, the model obtains a precision of 0.93, recall of 0.91, and F1-score of 0.92, indicating effective identification of legitimate network communications while maintaining a low false alarm rate.

For anomalous traffic, HDP-AFNet achieves a precision of 0.96, recall of 0.97, and F1-score of 0.96, demonstrating strong attack detection capability and a reduced missed-attack rate.

The higher recall for anomaly detection confirms that the hybrid DQN-Policy Gradient architecture effectively learns discriminative traffic behavior patterns and prioritizes attack identification in cybersecurity environments.

Furthermore, the weighted average F1-score of 0.95 indicates balanced performance across both classes despite class imbalance in the UNSW-NB15 dataset. These results validate the robustness and reliability of HDP-AFNet for real-world anomaly detection applications.

Table 9. Class-wise performance analysis of the proposed HDP-AFNet Framework

Class	Precision	Recall	F1-Score	Support
Benign	0.93	0.91	0.92	55,991
Anomaly	0.96	0.97	0.96	119,341
Accuracy	-	-	0.95	175,332
Macro Average	0.94	0.94	0.94	175,332
Weighted Average	0.95	0.95	0.95	175,332

4.16. Learning Stability Analysis of the Proposed HDP-AFNet Framework

Learning stability is an important performance parameter of anomaly detection systems based on reinforcement learning in dynamic and unbalanced network traffic environments. Table 10 shows the average reward by the proposed HDP-AFNet architecture, with a reward variance of 0.0413, convergence in 3 epochs and a learning stability index of 0.9346, which is 0.1826. Figure 6 presents the standard deviation of step-level reward for each epoch. Table 10 shows the average reward (0.1826) and reward variance (0.0413) calculated from the twelve epoch-level mean rewards in Figure 5. Thus, the two are on different scales. The convergence era is defined as the time after which the mean reward was positive for the first time. Relatively low reward variance implies less oscillatory learning behaviour and more consistent policy changes throughout training. This stability is achieved by the AFoGD optimisation strategy that uses fractional-order memory effects to smooth the gradient updates and avoid abrupt policy changes.

Moreover, the high indicator of learning stability indicates that the hybrid architecture of DQN-Policy Gradient would gradually converge to stable reward evolution and reliable anomaly detection policies. Also, the fast convergence in three epochs illustrates the efficiency of the suggested optimisation strategy to speed up the reinforcement learning adaption to complicated network settings.

Table 10. Learning stability analysis of the proposed HDP-AFNet framework

Metric	Value
Average Reward	0.1826
Reward Variance	0.0413
Convergence Epoch	3
Learning Stability Index	0.9346
Optimization Strategy	AFoGD
Training Stability	High

4.17. Error Analysis of the Proposed HDP-AFNet Framework

Error analysis is performed to examine the resilience of the HDP-AFNet architecture under actual network traffic conditions. As shown in Table 11, the framework has a specificity of 91.04% in detecting benign traffic flows. The false positive rate is 8.96%, which means the strategy may remove needless alarms, which are a burden of network administration. More crucially, the framework reduces missed attack detections, an essential cybersecurity metric, with a false negative rate of 3.26%. The recommended model displays a 96.74% recall, which indicates its ability to prioritise the assault detection with a balanced categorisation. The results demonstrate that the hybrid reinforcement learning architecture may be a good compromise between detection sensitivity and false alarm control in different network circumstances.

Table 11. Error analysis of the proposed HDP-AFNet framework

Error Metric	Value
Specificity	91.04%
False Positive Rate (FPR)	8.96%
False Negative Rate (FNR)	3.26%
True Positive Rate (Recall)	96.74%
True Negative Rate	91.04%
Error Control Capability	High

4.18. Real-Time Deployment Analysis of the Proposed HDP-AFNet Framework

The HDP-AFNet system is evaluated for real-time cybersecurity by analysing its throughput, inference latency, and overall processing performance. As shown in Table 12, the framework maintains the latency for each network flow at a very low level of 5.39 x 10-6 seconds, and the overall inference time for the test is 0.9449 seconds. The model is also exceedingly versatile and scalable, capable of processing about 185,556 network flows per second, making it perfect for use in systems monitoring enormous traffic volumes. The dependable AFoGD optimisation approach and lightweight hybrid DQN-Policy Gradient architecture help to achieve efficient inference performance. Our results demonstrate that HDP-AFNet can match the requirements of commercial networks, IoT ecosystems, cloud computing infrastructures and next-gen cybersecurity monitoring platforms on real-time anomaly detection.

Table 12. Real-Time deployment analysis of the proposed HDP-AFNet framework

Deployment Metric	Value
Total Test Inference Time	0.9449 sec
Latency per Flow	5.39×10^{-6} sec
Throughput	185,556.57 flows/sec
Real-Time Capability	Supported
Deployment Readiness	High
Scalability	Excellent

4.19. Comparison Study

The proposed HDP-AFNet framework is compared with the state-of-the-art reinforcement learning-based intrusion detection algorithms using the UNSW-NB15 dataset in Table 13. The Adaptive Cloud DRL IDS [27] has 92.80% accuracy and 0.9100 of ROC-AUC but lacked explainability.

However, DRL-IDS [28] achieves higher performance with 94.96% F1-score, however the architecture is more sophisticated. SSDDQN [29] reached an F1-score of 94.30% and ROC-AUC of 0.9200, whereas GAN-assisted DRL IDS [31] earned an F1-score of 95.12% and ROC-AUC of 0.9285 by means of synthetic traffic generation. TD3-DDPG-A2C [33] further better performance with 94.71% accuracy and 0.9331 ROC-AUC but with increased computational cost.

In comparison, the HDP-AFNet uses hybrid DQN-Policy Gradient architecture with AFoGD optimisation and XRL, delivering the greatest performance in terms of 94.92% accuracy, 95.84% precision, 96.74% recall, 96.28% F1-score and 0.9368 ROC-AUC. These findings suggest that HDP-AFNet can successfully achieve the trade-off of detection performance, learning stability and explainability on real-time network anomaly detection.

Table 13. Comparative performance analysis on UNSW-NB15 under a unified experimental protocol

Ref.	Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC	Explainable
1	Adaptive Cloud DRL IDS [27]	92.80	93.50	94.20	93.80	0.9100	No
2	DRL-IDS [28]	93.94	94.62	95.31	94.96	0.9234	No
3	SSDDQN [29]	93.40	94.00	94.70	94.30	0.9200	No
4	GAN-DRL IDS [31]	94.18	94.83	95.42	95.12	0.9285	No
5	TD3-DDPG-A2C [33]	94.71	95.36	96.01	95.68	0.9331	Partial (SHAP)
6	HDP-AFNet - Proposed	94.92	95.84	96.74	96.28	0.9368	Yes (XRL)

Figure 12 compares the accuracy and F1-score of the RL-based detectors listed in Table 13. HDP-AFNet exceeds the strongest baseline, TD3-DDPG-A2C [33], by 0.21 percentage points in accuracy and 0.60 points in F1-score, and exceeds Adaptive Cloud DRL IDS [27] by 2.12 and 2.48 points, respectively.

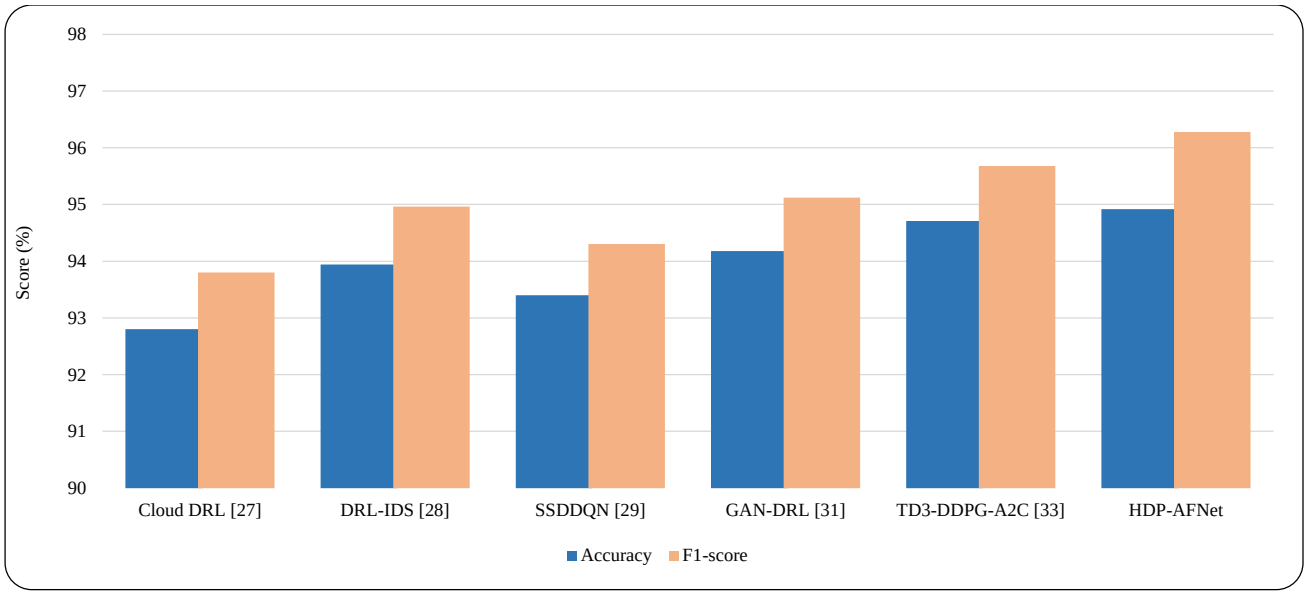


Fig. 12 Accuracy and F1-score of HDP-AFNet and existing RL-based detectors on UNSW-NB15

4.20. Reasons for the Performance Gain over Existing Methods

The advantage of HDP-AFNet over the detectors in Table 13 can be traced to four design choices, each supported by the ablation and stability results. The first is the combination of two learning signals. A DQN estimates action values through bootstrapped targets, which tend to overestimate and oscillate when rewards are sparse or skewed [9, 18], while a policy gradient learner follows the reward directly but with high variance [12]. The fusion gate blends the softmax of the Q-values with the policy output ($\lambda = 0.55$), so an error in one head is attenuated when the other head disagrees. Table 6 shows that this step raised the F1-score from 89.77% for the DQN alone to 94.48%. Earlier detectors [27-29] rely on one of these signals only, and the actor-critic study in [33] compares TD3, DDPG and A2C as separate agents rather than fusing value and policy outputs on a shared representation.

The second is the fractional-memory optimizer. Each AFoGD update combines the current gradient with a weighted history of earlier gradients ($\alpha = 0.8$, $\beta = 0.95$), which damps parameter updates that alternate in sign between mini-batches. The effect appears as a positive mean reward from the fourth epoch (Figure 5), a falling reward spread (Figure 6) and a further F1 gain of 1.80 percentage points over the same architecture trained without AFoGD (Tables 6 and 7). Detectors trained with standard gradient methods or tuned by metaheuristic search [31, 32] contain no such mechanism.

The third is the reward design. A missed attack costs -2 and a false alarm -0.5, so the agent learns that missing an attack is four times as costly as raising a false alarm. This explains why recall (96.74%) exceeds specificity (91.04%) and why the false negative rate is limited to 3.26% (Table 11). Resampling and GAN-based methods [21, 31] alter the data distribution but leave the objective unchanged, whereas HDP-AFNet places the cost asymmetry directly in the learning signal. The fourth is temporal state construction: windows of 10 consecutive flows expose short attack sequences such as scans and bursts that a per-flow classifier sees only as isolated records [22, 25].

The margin over the strongest baseline, TD3-DDPG-A2C [33], is small in accuracy (0.21 percentage points) but larger in recall (0.73 points) and F1-score (0.60 points), which is consistent with the recall-oriented reward. This gain is obtained with 2.9 M parameters and a latency of 5.39×10^{-6} s per flow (Tables 8 and 12). The price is a specificity of 91.04%, meaning that about 9% of benign flows are flagged, and a training time 2.2 hours longer than that of the plain DQN baseline.

4.21. Discussion

The experimental results show that the HDP-AFNet framework is able to effectively combine value-based reinforcement learning, policy-based learning, adaptive optimization and explainable decision analysis into a single anomaly detection framework. The hybrid DQN-Policy Gradient design is capable of effectively combining stable Q-value estimation with adaptive stochastic policy exploration, which, under the condition of heterogeneous and imbalanced network traffic, can effectively conduct robust anomaly discrimination. The results demonstrate that the framework has a high classification ability and attack detection performance, with an overall accuracy of 94.92%, a precision value of 95.84%, a recall value of 96.74%, an F1 score of 96.28%, a ROC-AUC of 0.9368, a PR-AUC of 0.9356, and an MCC of 0.8826. The high recall value shows the capability of HDP-AFNet to reduce missed attack events, and the specificity of 91.04% indicates good false alarm suppression.

The ablation study also confirms that every component is important to improve accuracy, and the complete HDP-AFNet achieves 94.92% accuracy, while the baseline DQN model can only reach 88.74%. Furthermore, the AFoGD optimization approach significantly enhances the stability of RL by decreasing the variance of rewards and achieving convergence after three epochs, thus obtaining a learning stability index of 0.9346. The framework also shows impressive deployment performance; it is able to reach a flow rate of 185,556.57 flows/sec with only 5.39×10^{-6} seconds per flow, which makes it well suited to real-time cybersecurity environments. In addition, Explainable Reinforcement Learning (XRL) modules are incorporated for feature attribution analysis, reward interpretation, action probability visualization, and temporal confidence tracking, which increases transparency and trust of the analyst. The results presented in this paper validate the effectiveness of HDP-AFNet to overcome the drawbacks of traditional IDS solutions as an adaptive, interpretable, stable and efficient solution for network traffic anomaly detection.

The proposed HDP-AFNet framework has great applications in cloud platforms, SDNs, IoT, 5G systems, enterprise networks, and security of critical infrastructure. Its hybrid reinforcement learning mechanism can be learned to detect unseen and zero-day threats, adapting to the changing patterns of attacks. The detection accuracy is high, and false negatives and inference latency are low, enabling real-time monitoring and automatic responses. The XRL-based interpretability mechanism improves incident analysis, compliance, and trust, and the AFoGD mechanism decreases training instability and computational load for scalable deployment. Hence, HDP-AFNet can be used effectively as a smart solution to next generation adaptive intrusion detection and security monitoring systems.

While performing well on anomalies detection, there are still some drawbacks. The framework is tested only on the UNSW-NB15 dataset and needs to be expanded to be tested on other datasets like CICIDS2017, CICDDoS2019, Bot-IoT, ToN-IoT, and CSE-CIC-IDS2018. It also does binary anomaly detection, which is a combination of all the attacks in a single class, and it does not do fine-grained classification. The added hyperparameters of AFoGD can have to be tuned across environments, but it does help to stabilize the convergence of the algorithm. Furthermore, the explainability module lacks causal and counterfactual analysis, and the framework doesn't support online continual learning and federated reinforcement learning for real-world distributed deployment.

5. Conclusion

This paper introduced HDP-AFNet, a robust network traffic anomaly detection framework that combines a hybrid DQN-Policy Gradient architecture with Adaptive Fractional-Order Gradient Descent (AFoGD) optimization. The proposed framework is a hybrid DQN-Policy Gradient architecture that harnesses the complementary benefits of both methods and memory-aware fractional-order updates for AFoGD optimization for improving convergence stability. In addition, Explainable Reinforcement Learning (XRL) modules were incorporated to explain and interpret the decisions made by an anomaly detector in a transparent and interpretable manner, with their support gaining a better understanding of the features attributed to the decisions, reward analysis, visualizing policies, and temporal confidence interpretation. The proposed framework was evaluated through experiments on the UNSW-NB15 dataset, with high accuracy (94.92%), precision (95.84%), recall (96.74%), F1-score (96.28%), ROC-AUC (0.9368), PR-AUC (0.9356), MCC (0.8826), and learning stability index (0.9346). The resulting framework also performed well in terms of inference latency and throughput, demonstrating its ability to handle real-time cybersecurity use cases.

The results obtained confirm the effectiveness of the proposed framework for intelligent network traffic anomaly detection, which combines hybrid reinforcement learning, adaptive fractional-order optimization and explainable AI, in terms of robustness, scalability, interpretability and computational efficiency. Future research will be directed towards expanding HDP-AFNet to classify multi-class attacks with the ability to identify the respective category of the cyberattack, including DoS, Exploits, Reconnaissance, Fuzzers, Backdoors and Worms. The framework can also be further strengthened by continual reinforcement learning with online technique, which can be helpful for evolving attack environments and concept drift adaptation.

References

[1] Varun Chandola, Arindam Banerjee, and Vipin Kumar, “Anomaly Detection: A Survey,” *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1-58, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[2] Kinza Arshad et al., “Deep Reinforcement Learning for Anomaly Detection: A Systematic Review,” *IEEE Access*, vol. 10, pp. 124017-124035, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Benson Mbithi Kimani, and Bernard Ondara Ondara “Optimizing Anomaly Detection in 5G and Beyond Networks using Reinforcement Learning,” *International Journal for Multidisciplinary Research*, vol. 8, no. 1, pp. 1-18, 2026. [CrossRef] [Publisher Link]

[4] Erika Puiutta, and Eric M.S.P. Veith, “Explainable Reinforcement Learning: A Survey,” *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Dublin, Ireland, vol. 12279, pp. 77-95, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] Mohammad Alauthman et al., "An Efficient Reinforcement Learning-based Botnet Detection Approach," *Journal of Network and Computer Applications*, vol. 150, pp. 1-16, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ahmad Javaid et al., "A Deep Learning Approach for Network Intrusion Detection System," *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, Brussels, Belgium, pp. 21-26, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mohammad Hossein Modirrousta et al., "Analysis of Anomalous Behaviour in Network Systems using Deep Reinforcement Learning with Convolutional Neural Network Architecture," *CAAI Transactions on Intelligence Technology*, vol. 9, no. 6, pp. 1467-1484, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Hanan Hindy et al., "A Taxonomy and Survey of Intrusion Detection System Design Techniques, Network Threats and Datasets," *arXiv Preprint*, vol. 4, pp. 1-28, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[9] Volodymyr Mnih et al., "Human-Level Control through Deep Reinforcement Learning," *Nature*, vol. 518, no. 7540, pp. 529-533, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[10] Yinjun Zhang, Mona Jamjoom, and Zahid Ullah, "Double Deep Q-Network Next-Generation Cyber-Physical Systems: A Reinforcement Learning-Enabled Anomaly Detection Framework for Next-Generation Cyber-Physical Systems," *Electronics*, vol. 12, no. 17, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Harshat Kumar, Alec Koppel, and Alejandro Ribeiro, "On the Sample Complexity of Actor-Critic method for Reinforcement Learning with Function Approximation," *Machine Learning*, vol. 112, no. 7, pp. 2433-2467, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Ronald J. Williams, "Simple Statistical Gradient-Following algorithms for Connectionist Reinforcement Learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229-256, 1992. [CrossRef] [Google Scholar] [Publisher Link]

[13] Nithia Vijayan, and L.A. Prashanth, "A Policy Gradient Approach for Optimization of Smooth Risk Measures," *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, vol. 216, pp. 2168-2178, 2023. [Google Scholar] [Publisher Link]

[14] Nour Moustafa, and Jill Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)," *2015 Military Communications and Information Systems Conference (MilCIS)*, Canberra, ACT, Australia, pp. 1-6, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[15] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, Funchal, Madeira, Portugal, vol. 1, pp. 108-116, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[16] Quoc Phong Nguyen et al., "GEE: A Gradient-based Explainable Variational Autoencoder for Network Anomaly Detection," *2019 IEEE Conference on Communications and Network Security (CNS)*, Washington, DC, USA, pp. 91-99, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Guansong Pang, et al., "Deep Learning for Anomaly Detection: A Review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 2, pp. 1-38, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[18] Hado van Hasselt, Arthur Guez, and David Silver, "Deep Reinforcement Learning with Double Q-Learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, pp. 2094-2100, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jiaxu Liu et al., "The Novel Adaptive Fractional Order Gradient Decent Algorithms Design via Robust Control," *arXiv*, pp. 1-8, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Stephanie Milani et al., "Explainable Reinforcement Learning: A Survey and Comparative Review," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1-36, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[21] Sarah Naiem et al., "Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing," *IEEE Access*, vol. 11, pp. 124597-124608, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Mhamad Bakro et al., "Efficient Intrusion Detection System in the Cloud Using Fusion Feature Selection Approaches and an Ensemble Classifier," *Electronics*, vol. 12, no. 11, pp. 1-27, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Hind Bangui, Mouzhi Ge, and Barbora Buhnova, "A Hybrid Machine Learning Model for Intrusion Detection in VANET," *Computing*, vol. 104, no. 3, pp. 503-531, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[24] Himanshu Setia et al., "Securing the Road Ahead: Machine Learning-Driven DDoS Attack Detection in VANET Cloud Environments," *Cyber Security and Applications*, vol. 2, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Md. Alamgir Hossain, and Md. Saiful Islam, "A Novel Hybrid Feature Selection and Ensemble-based Machine Learning Approach for Botnet Detection," *Scientific Reports*, vol. 13, no. 1, pp. 1-28, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[26] Asha Varma Songa, and Ganesh Reddy Karri, "An Integrated SDN Framework for Early Detection of DDoS Attacks in Cloud Computing," *Journal of Cloud Computing*, vol. 13, no. 64, pp. 1-22, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[27] Kamalakanta Sethi et al., "Deep Reinforcement Learning based Intrusion Detection System for Cloud Infrastructure," *2020 International Conference on Communication Systems and Networks (COMSNETS)*, Bengaluru, India, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] R. Kanimozhi, and P.S. Ramesh, "Deep Reinforcement Learning-based Intrusion Detection Scheme for Software-Defined Networking," *Scientific Reports*, vol. 15, no. 1, pp. 1-17, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[29] Shi Dong, Yuanjun Xia, and Tao Peng, "Network Abnormal Traffic Detection Model based on Semi-Supervised Deep Reinforcement Learning," *IEEE Transactions on Network and Service Management*, vol. 18, no. 4, pp. 4197-4212, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[30] Faten Louati, Farah Barika Ktata, and Ikram Amous, "Enhancing Intrusion Detection Systems with Reinforcement Learning: A Comprehensive Survey of RL-based Approaches and Techniques," *SN Computer Science*, vol. 5, no. 6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[31] Caroline Strickland et al., "DRL-GAN: A Hybrid Approach for Binary and Multiclass Network Intrusion Detection," *Sensors*, vol. 24, no. 9, pp. 1-18, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Kiran Kumar Paidipati et al., "Ensemble of Deep Reinforcement Learning with Optimization Model for DDoS Attack Detection and Classification in Cloud based Software Defined Network," *Multimedia Tools and Applications*, vol. 83, no. 11, pp. 32367-32385 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Suneeta Satpathy, Uttpal Tripathy, and Pratik Kumar Swain, "Cloud-based DDoS Detection using Hybrid Feature Selection with Deep Reinforcement Learning (DRL)," *Scientific Reports*, vol. 15, no. 1, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[34] Usman Ahmed et al., "Hybrid Bagging and Boosting with SHAP based Feature Selection for Enhanced Predictive Modeling in Intrusion Detection Systems," *Scientific Reports*, vol. 14, no. 1, pp. 1-32, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Oona Rainio, Jarmo Teuho, and Riku Klén, "Evaluation Metrics and Statistical Tests for Machine Learning," *Scientific Reports*, vol. 14, no. 1, pp. 1-14, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman, "The Matthews Correlation Coefficient (MCC) is more Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation," *BioData Mining*, vol. 14, no. 1, pp. 1-22, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[37] Safa Mohamed, and Ridha Ejbali, "Deep SARSA-based Reinforcement Learning Approach for Anomaly Network Intrusion Detection System," *International Journal of Information Security*, vol. 22, no. 1, pp. 235-247, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Mohamed Amine Merzouk et al., "Adversarial Robustness of Deep Reinforcement Learning-based Intrusion Detection," *International Journal of Information Security*, vol. 23, no. 6, pp. 3625-3651, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[39] Lekhetho Joseph Mpoporo, Pius Adewale Owolawi, and Chunling Tu, "Deep Reinforcement Learning Algorithms for Intrusion Detection: A Bibliometric Analysis and Systematic Review," *Applied Sciences*, vol. 16, no. 2, pp. 1-25, 2026. [CrossRef] [Google Scholar] [Publisher Link]

[40] Muhammad Umer et al., "Network Intrusion Detection Model using Wrapper based Feature Selection and Multi Head Attention Transformers," *Scientific Reports*, vol. 15, no. 1, pp. 1-15, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[41] Jian Wang et al., "Fractional-Order Gradient Descent Learning of BP Neural Networks with Caputo Derivative," *Neural Networks*, vol. 89, pp. 19-30, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[42] Vincent Zibi Mohale, and Ibidun Christiana Obagbuwa, "A Systematic Review on the Integration of Explainable Artificial Intelligence in Intrusion Detection Systems to Enhancing Transparency and Interpretability in Cybersecurity," *Frontiers in Artificial Intelligence*, vol. 8, pp. 1-10, 2025. [CrossRef] [Google Scholar] [Publisher Link]