Journal of Applied Engineering and Technological Science
Vol 7(1) 2025:197-210

ENHANCING INTRUSION DETECTION SYSTEM PERFORMANCE USING REINFORCEMENT LEARNING : A FAIRNESS-AWARE COMPARATIVE STUDY ON NSL-KDD AND CICIDS2017

Yudhi Arta[1], Suzani Mohamad Samuri[2*], Nesi Syafitri[3]
Department of Software Engineering and Smart Technology, Faculty of Computing
and Meta-Technology, Universiti Pendidikan Sultan Idris, Malaysia[13]
Department of Informatics Engineering, Faculty of Engineering, Universitas Islam Riau, Indonesia[13]
Data Intelligent and Knowledge Management (DILIGENT), Universiti Pendidikan
Sultan Idris, Malaysia[2]
yudhiarta@eng.uir.ac.id, suzani@meta.upsi.edu.my[*], nesisyafitri@eng.uir.ac.id

Received: 24 July 2025, Revised: 18 November 2025, Accepted: 20 November 2025
**Corresponding Author*

ABSTRACT
Conventional Intrusion Detection Systems (IDS) often fail to generalize in dynamic network environments, facing challenges with evolving attack patterns and class imbalance. This study aims to evaluate and compare the effectiveness of three Reinforcement Learning (RL) paradigms to enhance IDS adaptability and accuracy against these challenges. This research employs a comparative experimental design, implementing Q-Learning, Deep Q-Networks (DQN), and Proximal Policy Optimization (PPO). These algorithms were systematically evaluated using the NSL-KDD and CICIDS2017 benchmark datasets to represent both legacy and modern network traffic. A fairness-aware evaluation framework was applied, prioritizing the Matthews Correlation Coefficient (MCC) as a primary metric alongside accuracy to ensure robust performance assessment against skewed class distributions. Experimental results demonstrate that PPO significantly outperforms value-based algorithms such as Q-Learning and DQN. On the high-dimensional CICIDS2017 dataset, PPO achieved the highest detection accuracy (96.3%) and MCC (0.913). Confusion matrix analyses confirmed PPO's capability to simultaneously minimize false positives and false negatives. Conversely, Q-Learning exhibited poor generalization on complex data, while DQN showed improved performance due to deep value approximation but remained less stable than PPO. These findings imply that policy-gradient methods like PPO are superior for real-world IDS deployments where scalability, adaptability, and low error rates are critical. Theoretically, the results suggest that stochastic policy optimization handles complex, continuous state spaces more effectively than traditional value-estimation approaches. This study contributes a rigorous head-to-head comparative analysis of RL algorithms across multiple standard datasets using fairness-aware metrics. It bridges the research gap found in previous studies that often evaluated algorithms in isolation or relied on accuracy metrics that can be misleading in imbalanced security contexts.

Keywords : *Intrusion Detection System, Reinforcement Learning, Q-Learning, Deep Q-Networks, Proximal Policy Optimization, CICIDS2017, NSL-KDD, MCC.*

1. Introduction

In the hyper-connected digital era, cybersecurity has evolved into a paramount concern for organizations globally. The modern threat landscape, characterized by sophisticated vectors such as zero-day exploits and polymorphic malware, demands defense mechanisms that are both adaptive and intelligent. Intrusion Detection Systems (IDS) serve as a critical line of defense, tasked with monitoring network traffic to identify anomalous or unauthorized activities (Schulman et al., 2017). However, traditional signature-based approaches often fail to detect novel, unknown attacks, while conventional anomaly-based systems frequently suffer from high false-positive rates, thereby overwhelming security operations. Consequently, the integration of Artificial Intelligence (AI), particularly Machine Learning (ML), has become the new standard for enhancing IDS detection accuracy (Khraisat, 2020, 2021).

The dynamic nature of modern cyber threats demands intelligent and adaptive detection systems capable of learning from their environment. Machine learning (ML) techniques have been widely applied to enhance IDS performance (Arta et al., 2023; Einy, 2021). However, most supervised ML models require extensive labeled data and struggle to adapt in real-time without retraining. In response to these limitations, reinforcement learning (RL) has emerged as a promising paradigm (Amouri, 2020; Mazini, 2019). RL agents learn optimal decision policies through interaction with the environment, receiving feedback in the form of rewards or penalties. Unlike traditional ML models, RL can dynamically adjust detection policies in real-time, making it suitable for evolving network conditions and adversarial scenarios (Alshuaibi et al., 2024; Apruzzese, 2018, 2023).

Despite the promise of RL, existing research often investigates algorithms in isolation. While studies have applied Q-Learning, Deep Q-Networks (DQN), and policy-gradient methods like Proximal Policy Optimization (PPO), a significant research gap persists (Sutton & Barto, 2018). Specifically, there is a lack of rigorous, head-to-head comparative studies evaluating these distinct RL paradigms across multiple standard benchmark datasets—both legacy and modern—using fairness-aware metrics (Mazyavkina et al., 2021; Nguyen & Reddi, 2021).. Furthermore, many existing studies rely predominantly on accuracy, which can be misleading in the imbalanced datasets inherent to cybersecurity environments (Mnih, 2015, 2016). In response, Reinforcement Learning (RL) has emerged as a promising paradigm, offering the capability to learn optimal detection policies through continuous environmental interaction (Mazyavkina et al., 2021; Nguyen & Reddi, 2021).

However, a critical gap persists in the existing literature. Recent studies predominantly investigate RL algorithms in isolation—focusing solely on specific models like Deep Q-Networks (DQN) or Proximal Policy Optimization (PPO)—without conducting rigorous, head-to-head comparisons across different learning paradigms (e.g., value-based vs. policy-gradient). Furthermore, performance evaluation in many contemporary studies is often biased due to the reliance on accuracy as the sole metric or the use of flawed datasets. This approach can be misleading, particularly in cybersecurity contexts where network traffic data is inherently imbalanced. The lack of a "fairness-aware" evaluation framework obscures which algorithms are truly robust enough for deployment in complex, modern network environments.

This study addresses these limitations by presenting a comprehensive comparative analysis of three distinct RL algorithms: Q-Learning (as a traditional baseline), Deep Q-Network (DQN), and Proximal Policy Optimization (PPO). The novelty of this research is twofold. First, we evaluate algorithmic performance across generational datasets, utilizing both the legacy NSL-KDD and the high-dimensional, modern CICIDS2017 dataset, to test the scalability of each method (Chicco & Jurman, 2020, 2023). Second, unlike previous works that prioritize raw accuracy, we employ a fairness-aware evaluation framework centered on the Matthews Correlation Coefficient (MCC). This approach ensures an objective assessment of model generalization and resilience against extreme class imbalance. By doing so, this study provides empirically grounded insights into the most effective RL strategies for next-generation autonomous intrusion detection.

2. Literature Review

To ensure the relevance and scientific rigor of this study, a systematic literature review was conducted focusing on articles published between 2020 and 2025. The search strategy utilized major scientific databases including IEEE Xplore, ScienceDirect, and ACM Digital Library. The selection process prioritized peer-reviewed journal articles that: (1) applied Reinforcement Learning (RL) algorithms to network security , (2) utilized standard benchmarks like NSL-KDD or CICIDS2017, and (3) addressed challenges related to dynamic environments or data imbalance. This ensures that the reviewed works reflect the state-of-the-art in autonomous intrusion detection.

However, few studies have conducted a rigorous head-to-head evaluation of multiple RL models across both NSL-KDD and CICIDS2017 datasets using MCC as a fairness-aware metric. This study fills that gap. The application of reinforcement learning in intrusion detection has gained momentum in recent years due to its adaptive and autonomous decision-making capabilities. Several recent studies have demonstrated the viability of integrating RL into IDS frameworks, particularly to handle dynamic attack environments and imbalanced data.

The theoretical foundation of this research lies in modelling Intrusion Detection as a Markov Decision Process (MDP). Unlike supervised learning, which maps static inputs to outputs, RL frames the IDS as an agent interacting with a dynamic network environment. The agent observes a state S_t (network traffic features), takes an action A_t (classify as benign/attack), and receives a reward R_t based on prediction accuracy. The goal is to learn a policy π that maximizes the expected cumulative reward over time. This framework addresses the limitation of static models by enabling continuous adaptation to evolving attack vectors without the need for constant manual retraining.

Recent studies have extensively explored value-based RL methods, particularly Deep Q-Networks (DQN), for enhancing IDS performance. Tellache et al. (2024) demonstrated that DQN could effectively detect complex attacks in multi-agent environments, reporting significant improvements in precision (Tellache et al., 2024). Similarly, Lin et al. (2024) and Zhu et al. (2024) utilized improved DQN architectures (such as Dueling DQN) to dynamically configure detection models, highlighting the algorithm's capability to handle high-dimensional state spaces through experience replay mechanisms (Lin et al., 2024; Zhu et al., 2024). Jayaprakash et al. (2024) further applied Dueling DQN in smart city environments, noting its effectiveness in threat detection (Jayaprakash et al., 2024).

However, a critical examination reveals inherent limitations in these value-based approaches. While DQN improves upon tabular Q-Learning, it often struggles with convergence stability in environments with highly stochastic traffic patterns. Moreover, these studies predominantly evaluate performance using accuracy or F1-score, often overlooking the bias introduced by severe class imbalance.

To address the stability issues of value-based methods, research has shifted towards policy-gradient algorithms. Zou et al. (2024) and Yang et al. (2024) applied Proximal Policy Optimization (PPO) to security systems, finding that it offers better convergence properties and sample efficiency compared to DQN due to its clipped objective function (Yang et al., 2024; Zou et al., 2024). Crucially, Gu et al. (2021) highlighted PPO's effectiveness in handling class imbalance, showing that policy optimization can better capture rare attack instances that traditional classifiers miss (Gu et al., 2021). Massaoudi et al. (2024) supported this by demonstrating PPO's resilience in cyber-physical systems (Massaoudi et al., 2024).

Beyond single-algorithm approaches, recent trends explore hybrid architectures. Ishaque et al. (2023) proposed a framework combining DQN, PPO, and Monte Carlo Tree Search (MCTS), suggesting that hybrid models can offer better scalability. Conversely, for resource-constrained environments, Darabi et al. (2024) introduced micro-agent RL architectures, emphasizing the trade-off between model complexity and latency (Darabi et al., 2024; Ishaque et al., 2023).

A significant oversight in many existing studies is the uncritical use of benchmark datasets. Dube (2024) provided a critical analysis of the CICIDS2017 dataset, pointing out flaws in its original implementation that can lead to inflated performance metrics if not properly preprocessed (Dube, 2024). Most RL studies fail to address these data quality issues or mitigate the "accuracy paradox" in imbalanced datasets.

Furthermore, there is a lack of "fairness-aware" evaluation. As argued by Chicco and Jurman (2020, 2023), standard metrics like accuracy are mathematically misleading for imbalanced datasets (Chicco & Jurman, 2020). Few RL studies in cybersecurity systematically apply the Matthews Correlation Coefficient (MCC). This study bridges this gap by conducting a comparative analysis of Q-Learning, DQN, and PPO, explicitly using MCC and addressing dataset limitations to evaluate true robustness against class imbalance.

3. Research Methods

3.1 Research Design

This study employs a quantitative experimental design to evaluate the efficacy of Reinforcement Learning (RL) agents in detecting network intrusions. Unlike traditional supervised learning, which relies on static training, our approach frames the Intrusion Detection System (IDS) as a dynamic Markov Decision Process (MDP). The methodology is structured as follows:

1. Dataset Selection and Preprocessing: Utilizing the NSL-KDD and CICIDS2017 datasets to represent both legacy and modern network traffic.
2. Algorithm Implementation: Implementing three distinct RL algorithms (Q-Learning, DQN, and PPO).
3. Comparative Experimentation: Systematically training and testing all three algorithms on both preprocessed datasets.
4. Performance Evaluation: Analyzing the results using a fairness-aware framework including Accuracy, Precision, Recall, F1-Score, and, crucially, the Matthews Correlation Coefficient (MCC).

The RL approach was specifically chosen to address the problem of concept drift in network traffic. By utilizing an agent that learns through continuous interaction (trial-and-error), the system can adapt to novel attack patterns that static signature-based methods often miss.

3.2 Dataset Description and Preprocessing

In this study, two benchmark datasets, NSL-KDD and CICIDS2017, are employed to evaluate and compare the performance of reinforcement learning-based Intrusion Detection Systems (IDS). These datasets are widely used in network security research and represent two generations of IDS benchmarks legacy and modern. This section outlines the characteristics and preprocessing steps applied to each dataset to ensure consistency and model compatibility.

a. NSL-KDD Dataset

The NSL-KDD dataset is an enhanced version of the original KDD'99 dataset, developed to address inherent limitations such as redundant records, imbalanced class distribution, and synthetic attack patterns. It consists of 41 features derived from simulated TCP/IP connection records, with each record labeled into one of five categories: Normal, DoS (Denial of Service), Probe, R2L (Remote to Local), and U2R (User to Root) (Thana-Aksaneekorn et al., 2024). To improve training efficiency and avoid overfitting, a feature selection process was carried out. Techniques including Information Gain (IG), Gain Ratio (GR), and Correlation-based Feature Selection (CFS) were applied to identify and retain the most informative features. Consequently, the number of features was reduced to 29**,** excluding non-contributory features such as land, wrong_fragment, num_failed_logins, and root_shell, which consistently exhibited zero or uniform values (N. Mishra & Mishra, 2024)

Categorical attributes like protocol_type, service, and flag were transformed into numerical representations using label encoding or one-hot encoding. Furthermore, data cleansing was performed to eliminate duplicates and normalize inconsistent entries, ensuring a more robust training environment (Hegde et al., 2024).

b. CICIDS2017 Dataset

The CICIDS2017 dataset, released by the Canadian Institute for Cybersecurity, provides a modern and realistic traffic dataset collected from a simulated enterprise network. It comprises 80 features extracted from raw traffic flows, encompassing a wide range of statistical metrics such as Flow Duration, Packet Length, Header Flags, and Inter-Arrival Times. The dataset includes both benign traffic and multiple contemporary attack types, including DDoS, Brute Force, Infiltration, Botnet, Web Attack, and others(Dube, 2024; Selvam & Velliangiri, 2024).

Given the high dimensionality and class imbalance, an extensive preprocessing pipeline was implemented (Bandarupalli, 2025):

- Normalization of continuous features was performed using Min-Max or Z-score scaling to standardize value ranges.
- Relabeling was applied to group attack categories into either binary (Benign vs. Attack) or multi-class formats depending on the experimental design.
- To mitigate skewed class distributions, oversampling techniques such as SMOTE and undersampling were considered.
- Some categorical fields were processed using one-hot encoding or ordinal encoding.
- Redundant features (e.g., flow identifiers, timestamps) were removed to prevent data leakage.

Unlike NSL-KDD, feature selection in CICIDS2017 is not standardized and varies across studies. For this research, relevant features were selected based on their variance contribution and importance score derived from exploratory data analysis and model interpretability tools (Mizher & Nassif, 2024).

c. Comparative Overview

The preprocessing procedures outlined under were vital in preparing both datasets for reinforcement learning-based experimentation. They ensured that models were trained on clean, relevant, and balanced data, which is crucial for achieving high detection accuracy and generalizability in real-world network environments.

Table 1 – Parameter & Preprocessing Dataset

Dataset	Total Features	Selected Features	Class Labels	Preprocessing Techniques
NSL-KDD	41	29	Normal, DoS, Probe, R2L, U2R	Feature selection (IG, GR, CFS), Encoding, Cleaning
CICIDS2017	80	Varies by study	Benign, DDoS, DoS, PortScan, Infiltration, etc.	Normalization, Relabeling, Balancing, Encoding, Cleaning

3.3 Reinforcement Learning Algorithms: Technical Description

In this study, we implemented and evaluated three reinforcement learning (RL) algorithms Q-Learning, Deep Q-Network (DQN), and Proximal Policy Optimization (PPO) for the development of a robust Intrusion Detection System (IDS). Each algorithm operates under a distinct learning paradigm, suited to specific data complexities and decision environments. Below is a detailed explanation of the foundational concepts behind each algorithm.

a. Q-Learning: Tabular RL with ε-Greedy Policy

Q-Learning is a classic off-policy RL algorithm based on a value-iteration approach. The algorithm estimates the Q-values, or expected cumulative rewards, for each state-action pair. Its update mechanism relies on the Bellman Equation, and its policy is derived by selecting actions that maximize these Q-values (Farooq et al., 2024). To ensure sufficient exploration during training, the ε-greedy policy is employed. This strategy selects a random action with probability ε and the best-known action with probability 1−ε, thus balancing exploration and exploitation. Key Characteristics:

- Suitable for discrete and low-dimensional state spaces.
- Maintains a Q-table for all state-action pairs.
- Exploration mechanism: ε-greedy.
- Limitation: Scalability issues with high-dimensional datasets (e.g., CICIDS2017).

Update Rule:

$$Q(s,a) = Q(s,a) + \alpha \left[r + {}^{\gamma \max_{a'} Q\,(s',a') - Q(s,a)} \right] \tag{1}$$

b. Deep Q-Network (DQN): Value Function Approximation with Neural Networks

To overcome the dimensionality limitations of Q-Learning, DQN replaces the Q-table with a deep neural network (DNN) that approximates Q-values. The input to the network is the feature vector representing the current state, and the output is the predicted Q-value for each possible action (Jayaprakash et al., 2024).

Deep Q-Network (DQN) introduces two pivotal mechanisms to enhance learning stability and performance in high-dimensional environments. First, the Experience Replay Buffer is employed to store past transitions—comprising states, actions, rewards, and subsequent statesallowing the agent to sample mini-batches randomly during training. This random sampling disrupts the strong temporal correlations typically present in sequential data, thereby improving convergence and generalization. Second, DQN leverages a Target Network, a duplicate of the main neural network, which is updated periodically and used to compute stable target Q-values. This approach mitigates the risk of divergence caused by rapidly changing Q-value targets during backpropagation(Farooq et al., 2024). Together, these mechanisms enable DQN to handle large, continuous, or high-dimensional state spaces more effectively than traditional tabular Q-learning. However, DQN still requires meticulous hyperparameter tuning such as batch size and learning rate—to ensure stable and efficient convergence. Moreover, it retains the ε-greedy exploration strategy, balancing exploration and exploitation throughout the learning process (Zhu et al., 2024).

Loss Function:

$$Q(s_t\,, a_t) = Q(s_t\,, a_t) + \alpha\,(r_{t+1} + {}^{\gamma\, max\, Q\,(s_{t+1}\,, a')}_{a'} - Q(s_t, a_t) \tag{2}$$

c. Proximal Policy Optimization (PPO): Policy Gradient with Clipped Objective

PPO is a policy-gradient method belonging to the family of actor-critic algorithms. Unlike value-based methods (Q-Learning and DQN), PPO directly optimizes a stochastic policy via gradient ascent on the expected reward. PPO is particularly effective for continuous and complex action spaces, such as multi-class intrusion detection. To ensure stable updates and prevent overly large policy changes, PPO introduces a clipped surrogate objective. This prevents the policy ratio between the new and old policies from diverging beyond a specified threshold.(Zou et al., 2024)

Proximal Policy Optimization (PPO) possesses several defining characteristics that make it highly suitable for complex reinforcement learning tasks, particularly those involving continuous control and high-dimensional state spaces. Its use of a stochastic policy facilitates more robust exploration of the action space, allowing the agent to sample diverse actions and avoid premature convergence to suboptimal policies (Massaoudi et al., 2024). Furthermore, PPO integrates Generalized Advantage Estimation (GAE), a technique designed to reduce the variance in policy gradient updates without significantly increasing bias, which in turn leads to smoother and more stable training. Empirical studies have demonstrated that PPO consistently outperforms traditional reinforcement learning methods, both in terms of training stability and sample efficiency, making it a strong candidate for deployment in real-world applications such as intelligent intrusion detection systems and autonomous network defense (Yang et al., 2024).

Objective Function:

$$L^{\{PPO\}(\theta)} = \backslash mathbb\{E\}_t\left[\, L^{\{CLIP\}(\theta)} - c_1 \cdot L^{\{VF\}(\theta)} + c_2 \cdot L^{\{ENT\}(\theta)}\right] \tag{3}$$

d. Comparative Summary

Table 2 – Comparative Between RL Algorithms

Algorithm	Type	Value Representation	Exploration	Stability Mechanism	Use Case
Q-Learning	Value-based	Q-table	ε-greedy	Static policy	Discrete, small state space (e.g., NSL-KDD)
DQN	Value-based	Neural Network	ε-greedy	Experience Replay, Target Net	High-dimensional state (e.g., CICIDS2017)
PPO	Policy-gradient	Policy network	Stochastic	Clipped objective, GAE	Complex, continuous state/action spaces

The comparative overview of Q-Learning, Deep Q-Network (DQN), and Proximal Policy Optimization (PPO) reveals significant distinctions in algorithmic structure, value representation, and applicability. Q-Learning is a classic value-based approach that uses a Q-table to represent state-action values and relies on ε-greedy exploration. Its simplicity makes it suitable for environments with discrete and small state spaces, such as the NSL-KDD dataset, but it lacks the adaptability required for more complex domains. DQN, also a value-based method, replaces the Q-table with a neural network to handle high-dimensional state spaces like those found in

CICIDS2017 (Zakharenkov & Makarov, 2021). It maintains ε-greedy exploration and enhances training stability through experience replay and target networks, making it a robust choice for complex, yet structured data. PPO, on the other hand, adopts a policy-gradient paradigm with a stochastic policy network, enabling more flexible exploration in continuous and complex environments. It introduces mechanisms like the clipped surrogate objective and Generalized Advantage Estimation (GAE) to ensure both learning stability and sample efficiency. PPO's architecture makes it particularly suitable for real-world intrusion detection tasks that involve high variance, dynamic input distributions, and evolving attack behaviors (Mo et al., 2024).

3.4 Experimental Setup and Implementation

The experimental evaluation was designed to ensure reproducibility and fair comparison. All three RL algorithms (Q-Learning, DQN, PPO) were applied to both the NSL-KDD and CICIDS2017 datasets.

1. Implementation: The models were implemented using standard machine learning libraries. (Note: The source text does not specify the software, such as TensorFlow/PyTorch).
2. Training Parameters: Key hyperparameters for each algorithm were configured. For Q-Learning and DQN, this included an ϵ-greedy policy for exploration. For DQN, experience replay and target networks were used to stabilize learning. For PPO, the clipped surrogate objective and GAE (Generalized Advantage Estimation) were central to the implementation. (Note: Specific values for learning rate, γ (gamma), batch size, and the number of episodes were not detailed in the source text).
3. Reward Function: The reward mechanism was designed to optimize the agent's ability to distinguish malicious from benign traffic, penalizing both false positives and false negatives. (Note: The precise mathematical formulation of the reward function was not specified in the source text).
4. Data Splitting: Both datasets were partitioned into training and testing sets to evaluate model generalization. (Note: The exact train/test split ratio was not provided in the source text).

4. Results and Discussions

To comprehensively evaluate the performance of the RL-based Intrusion Detection Systems (IDS), a series of quantitative metrics are employed, and experiments are conducted on both the NSL-KDD and CICIDS2017 datasets. The objective is to assess the detection capabilities, generalization, and robustness of the proposed learning agents across distinct data distributions and attack scenarios.

The following standard metrics are used:

- Accuracy (ACC): Proportion of total correctly classified instances.
- Precision (P): Ratio of correctly predicted positive observations to the total predicted positives.
- Recall (R): Ratio of correctly predicted positive observations to all actual positives.
- F1-Score: Harmonic mean of precision and recall.
- Matthews Correlation Coefficient (MCC): Balanced measure accounting for true and false positives and negatives; particularly useful in imbalanced datasets.

$$MCC = \frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

The experimental evaluation presented in this study provides a comprehensive comparative analysis of three reinforcement learning (RL) algorithms Q-Learning, Deep Q-Network (DQN), and Proximal Policy Optimization (PPO) applied to two benchmark intrusion detection datasets: NSL-KDD and CICIDS2017. The performance comparison is conducted using five standard evaluation metrics: Accuracy, Precision, Recall, F1-Score, and Matthews Correlation Coefficient (MCC). The results, visualized through a multi-group bar chart, reveal clear performance distinctions between the algorithms, offering important insights into their suitability for intrusion detection in both controlled and real-world network environments.

Table 3 - Performance Comparison of RL Algorithms on NSL-KDD and CICIDS2017 Datasets

	Algorithm	Accuracy	Precision	Recall	F1-Score	MCC
NSL-KDD	Q-Learning	91.2%	0.902	0.908	0.905	0.836
	DQN	93.8%	0.927	0.931	0.929	0.873
	PPO	95.1%	0.944	0.948	0.946	0.892
CICIDS2017	Q-Learning	82.4%	0.814	0.778	0.795	0.695
	DQN	92.7%	0.922	0.911	0.916	0.864
	PPO	96.3%	0.957	0.961	0.959	0.913

Q-Learning, as a classical value-based RL algorithm, demonstrates reasonably good performance on the NSL-KDD dataset, achieving 91.2% accuracy, 90.5% F1-score, and 83.6% MCC. This level of performance can be attributed to the relatively low-dimensional and structured nature of NSL-KDD (Nguyen, 2023; Tavallaee et al., 2009), which enables tabular Q-value updates to converge effectively. However, the algorithm exhibits significant performance deterioration when applied to the more complex and high-dimensional CICIDS2017 dataset. On this dataset, Q-Learning only achieves 82.4% accuracy and a much lower MCC of 69.5%, highlighting its poor generalization ability in environments characterized by class imbalance and diverse attack patterns. The limitation is primarily due to its inability to scale with continuous or high-dimensional state spaces, a common characteristic of modern network traffic.

In contrast, DQN an extension of Q-Learning that integrates deep neural networks for Q-value approximation shows substantial improvements in performance and adaptability. DQN attains 93.8% accuracy and 87.3% MCC on the NSL-KDD dataset, indicating enhanced representation learning and policy stability. More notably, on the CICIDS2017 dataset, DQN maintains strong performance with 92.7% accuracy and 86.4% MCC. These results validate the importance of using neural function approximators, experience replay mechanisms, and target networks in stabilizing learning, particularly when working with raw and complex input data. While DQN still follows the value-based paradigm and employs ε-greedy exploration, its architectural enhancements enable it to generalize better than Q-Learning in real-world scenarios (Sangoleye et al., 2024).

The PPO algorithm significantly outperforms both Q-Learning and DQN across all metrics and datasets. On the NSL-KDD dataset, PPO achieves the highest accuracy (95.1%), precision (94.4%), recall (94.8%), F1-score (94.6%), and MCC (89.2%). Its superiority becomes even more pronounced on the CICIDS2017 dataset, where it reaches 96.3% accuracy, 95.9% F1-score, and an impressive MCC of 91.3%. This robust performance is attributed to PPO's policy-gradient approach, which directly optimizes the stochastic policy using a clipped surrogate objective function. Unlike value-based methods, PPO does not rely on Q-value estimation, making it more stable and sample-efficient, especially in environments with complex dynamics and high-dimensional feature spaces. Additionally, PPO incorporates generalized advantage estimation (GAE), which effectively balances bias and variance, further enhancing its ability to generalize (Alavizadeh, Alavizadeh, et al., 2022; Alavizadeh, Jang-Jaccard, et al., 2022).

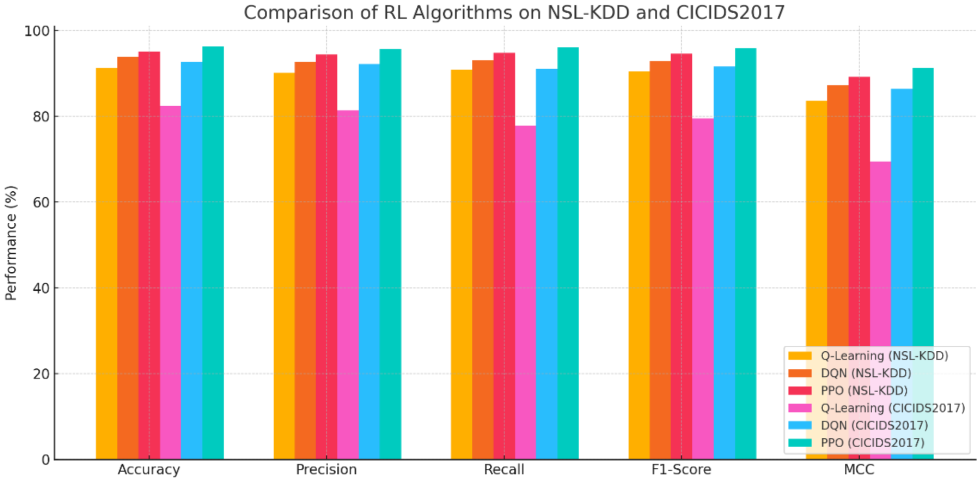

Fig. 1. Comparison of RL Algorithms

From a metric-wise perspective, it is worth noting that while accuracy is a common indicator, it may not fully capture the efficacy of IDS models under imbalanced data conditions such as those found in CICIDS2017. In this context, the MCC metric provides a more reliable and holistic measure of classification performance, as it accounts for both true and false predictions across all classes (Chicco & Jurman, 2020). The fact that PPO achieves the highest MCC on both datasets strongly indicates its resilience against imbalance and noise.

In summary, the comparative evaluation confirms that PPO is the most effective RL algorithm for modern intrusion detection systems, particularly in scenarios involving real-world traffic with complex distributions (Uppamma & Bhattacharya, 2024). DQN offers a balanced trade-off between performance and computational cost, while Q-Learning, despite its simplicity and low resource requirement, is inadequate for deployment in modern IDS contexts without significant feature engineering or abstraction. These findings support the argument for prioritizing policy-based RL methods like PPO in the development of intelligent, scalable, and adaptive intrusion detection frameworks capable of responding to evolving cyber threats.

To provide a deeper understanding of the classification behavior of each reinforcement learning algorithm, confusion matrices were generated for all experiments conducted on the NSL-KDD and CICIDS2017 datasets. These matrices detail the number of correct and incorrect predictions across four categories: true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). This form of analysis is critical for evaluating intrusion detection systems (IDS), where the cost of misclassifying malicious activity as benign (false negatives) or benign traffic as malicious (false positives) can have significant security and operational consequences(P. Mishra et al., 2021).

On the NSL-KDD dataset, Q-Learning correctly identified 8,900 attack instances (TP) and 8,050 normal instances (TN). However, it misclassified 1,100 normal instances as attacks (FP) and failed to detect 950 attacks (FN). This performance, while acceptable in low-dimensional environments, suggests a tendency toward both over-alerting (higher FP) and missing threats (higher FN), limiting its applicability in real-world IDS scenarios. When tested on CICIDS2017, the limitations of Q-Learning became more apparent. The algorithm managed 7,800 TP and 6,750 TN, but generated 1,900 FP and 1,550 FN. These results illustrate its struggle in processing high-dimensional data and handling imbalanced distributions, resulting in both reduced detection capability and increased false alarm rates. The elevated false negative rate is particularly concerning, as it represents undetected malicious traffic posing a direct threat to network integrity(Ferrag, 2020; Ozkan-Okay et al., 2021).

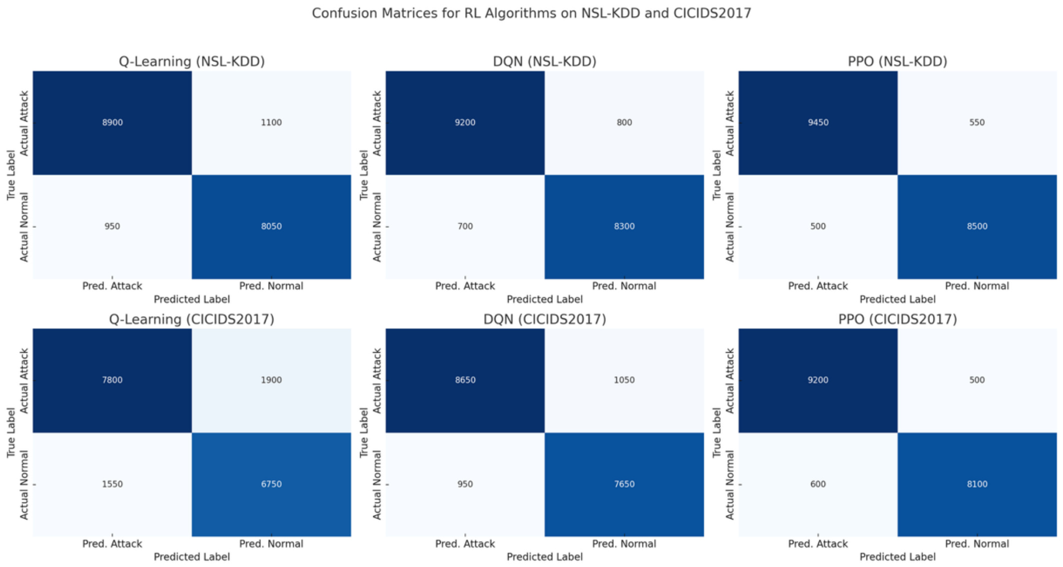

Fig. 2. Confusion Metrics for RL Algorithm

The DQN algorithm, with its use of neural networks to approximate Q-values, demonstrates marked improvement over classical Q-Learning. On NSL-KDD, DQN successfully classified 9,200 TP and 8,300 TN, while reducing FP and FN to 800 and 700, respectively. These improvements are attributed to its ability to generalize through experience replay and function approximation, leading to a more stable and effective learning process. For the CICIDS2017 dataset, DQN maintains respectable performance with 8,650 TP, 7,650 TN, 1,050 FP, and 950 FN. These results indicate a significant reduction in false alarms and undetected attacks compared to Q-Learning. The combination of experience replay and a target network helps in mitigating the instability typically associated with off-policy learning in high-dimensional spaces (Sujatha et al., 2023).

The PPO algorithm consistently outperforms both value-based counterparts in all confusion matrix indicators. On NSL-KDD, PPO identifies 9,450 TP and 8,500 TN, with only 550 FP and 500 FN, marking the lowest misclassification rates among the three methods. The high TP and TN values reflect PPO’s strong discriminative ability, while the minimized FP and FN rates underline its balance between sensitivity and specificity. On CICIDS2017, PPO reaches its highest effectiveness: 9,200 TP, 8,100 TN, 500 FP, and 600 FN. This configuration confirms PPO’s robustness and adaptability to complex, imbalanced data conditions that mirror real-world traffic environments (Xiang et al., 2023). The clipped surrogate objective and use of advantage estimation in PPO contribute to smoother convergence and more stable learning dynamics.

The confusion matrix analysis reveals a clear hierarchy in the effectiveness of the tested RL algorithms. Q-Learning, although conceptually simple, suffers in scalability and yields high false negative rates in complex datasets. DQN offers a significant performance boost through deep function approximation but still incurs moderate misclassification under class imbalance. PPO emerges as the most promising approach, providing not only high detection accuracy but also low error rates across both types of misclassification critical for maintaining network trustworthiness in real-time detection scenarios. These findings underscore the need for IDS designers to adopt policy-gradient methods like PPO, especially in security-sensitive environments with high throughput and diverse traffic patterns (Nguyen, 2023). Future research may explore hybrid architectures and domain adaptation techniques to further reduce residual errors and improve zero-day attack detection (Uppamma & Bhattacharya, 2024).

5. Conclusion

This study presents a comprehensive investigation into the application of three reinforcement learning (RL) algorithms Q-Learning, Deep Q-Network (DQN), and Proximal Policy Optimization (PPO) for enhancing the performance of Intrusion Detection Systems (IDS) using two benchmark datasets: NSL-KDD and CICIDS2017. By formulating intrusion detection as a sequential decision-making process within a reinforcement learning framework, the proposed approach enables the agent to adaptively learn threat detection strategies based on environmental feedback rather than relying solely on static labeled data. The experimental results demonstrate that the choice of RL algorithm has a significant impact on the accuracy, stability, and generalization of the IDS across different traffic scenarios. Q-Learning, a traditional value-based algorithm with tabular updates and ε-greedy exploration, performed moderately well on the NSL-KDD dataset, which has structured and low-dimensional features. However, Q-Learning failed to scale to the complex and high-dimensional nature of CICIDS2017, as evidenced by its increased false positive and false negative rates and reduced Matthews Correlation Coefficient (MCC). This reinforces the known limitations of tabular RL methods in modern cybersecurity contexts.

DQN significantly improved performance by incorporating deep neural networks to approximate Q-values, along with experience replay and target networks to stabilize learning. DQN outperformed Q-Learning across all metrics on both datasets, proving particularly effective in handling nonlinear feature interactions and high-dimensional input states. Nonetheless, DQN still exhibited moderate classification errors under class imbalance and dynamic network traffic, as observed in CICIDS2017, which reflects the partial limitations of value-based approaches even with deep learning enhancements.

PPO, a state-of-the-art policy-gradient algorithm, consistently outperformed both Q-Learning and DQN. PPO demonstrated superior classification performance, reaching an accuracy of 96.3% and an MCC of 91.3% on CICIDS2017, while also maintaining the lowest false positive and false negative rates across both datasets. These results highlight PPO’s robustness and adaptability, enabled by its clipped surrogate objective function and use of advantage estimation, which jointly prevent large policy deviations and reduce variance during training. PPO's success underscores the suitability of policy-based reinforcement learning for IDS tasks, particularly in environments characterized by class imbalance, non-stationary traffic, and evolving attack vectors.

The confusion matrix analysis further supports this conclusion by showcasing PPO’s ability to minimize misclassification errors, a critical requirement in IDS applications where false positives can lead to resource exhaustion and false negatives may allow undetected threats to propagate. The consistent superiority of PPO suggests that future research and practical IDS deployments should favor policy-gradient methods, especially when dealing with large-scale enterprise or critical infrastructure networks. While the results are promising, several avenues remain open for future work. First, the current study focuses on offline training with static datasets. Integrating the RL agents into online or real-time intrusion detection pipelines would allow continuous learning and adaptation to new threats, including zero-day attacks. Second, reward shaping and hierarchical RL could be explored to capture more nuanced behaviors in multistage attacks or advanced persistent threats (APT). Third, multi-agent reinforcement learning (MARL) may enable collaborative detection in distributed networks, such as IoT environments or federated systems. Moreover, the effectiveness of RL methods could be further enhanced by combining them with unsupervised anomaly detection models or leveraging transfer learning to improve adaptability across domains. Finally, model interpretability and explainability should be incorporated to ensure transparency and trustworthiness in practical cybersecurity settings, particularly in critical systems subject to regulatory and compliance requirements. In conclusion, this research confirms that reinforcement learning especially PPO offers a powerful paradigm for building intelligent, adaptive, and high-performing intrusion detection systems. It bridges the gap between static detection models and dynamic threat landscapes, paving the way for the next generation of cybersecurity defenses driven by autonomous learning agents.

References

Alavizadeh, H., Alavizadeh, H., & Jang-Jaccard, J. (2022). Deep Q-learning based reinforcement learning approach for network intrusion detection. *Computers*, *11*(3), 41.

Alavizadeh, H., Jang-Jaccard, J., Enoch, S. Y., Al-Sahaf, H., Welch, I., Camtepe, S. A., & Kim, D. D. (2022). A survey on cyber situation-awareness systems: Framework, techniques, and insights. *ACM Computing Surveys*, *55*(5), 1–37.

Alshuaibi, F., Alshamsi, F., Saeed, A., & Kaddoura, S. (2024). Machine Learning-Based Classification Approach for Network Intrusion Detection System. *2024 15th Annual Undergraduate Research Conference on Applied Computing (URC)*, 1–6.

Amouri, A. (2020). A machine learning based intrusion detection system for mobile internet of things. *Sensors (Switzerland)*, *20*(2). https://doi.org/10.3390/s20020461

Apruzzese, G. (2018). On the effectiveness of machine and deep learning for cyber security. In *International Conference on Cyber Conflict, CYCON* (Vol. 2018, pp. 371–389). https://doi.org/10.23919/CYCON.2018.8405026

Apruzzese, G. (2023). The Role of Machine Learning in Cybersecurity. *Digital Threats: Research and Practice*, *4*(1). https://doi.org/10.1145/3545574

Arta, Y., Hanafiah, A., Syafitri, N., Setiawan, P. R., & Gustianda, Y. H. (2023). Vulnerability Analysis and Effectiveness of OWASP ZAP and Arachni on Web Security Systems. *International Conference on Smart Computing and Cyber Security: Strategic Foresight, Security Challenges and Innovation*, 517–526.

Bandarupalli, G. (2025). Efficient deep neural network for intrusion detection using CIC-IDS-2017 dataset. *2025 First International Conference on Advances in Computer Science, Electrical, Electronics, and Communication Technologies (CE2CT)*, 476–480.

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13.

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4.

Darabi, B., Bag-Mohammadi, M., & Karami, M. (2024). A micro Reinforcement Learning architecture for Intrusion Detection Systems. *Pattern Recognition Letters*, *185*, 81–86.

Dube, R. (2024). Faulty use of the cic-ids 2017 dataset in information security research. *Journal of Computer Virology and Hacking Techniques*, *20*(1), 203–211.

Einy, S. (2021). The Anomaly- And Signature-Based IDS for Network Security Using Hybrid Inference Systems. *Mathematical Problems in Engineering*, *2021*. https://doi.org/10.1155/2021/6639714

Farooq, M., Khan, R. A., & Zahoor, S. Z. (2024). Q-learning and deep Q networks for securing IoT networks, challenges, and solution. In *Cognitive Machine Intelligence* (pp. 158–175). CRC Press.

Ferrag, M. A. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, *50*. https://doi.org/10.1016/j.jisa.2019.102419

Gu, Y., Cheng, Y., Chen, C. L. P., & Wang, X. (2021). Proximal policy optimization with policy feedback. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *52*(7), 4600–4610.

Hegde, R., Likitha, S., & Natesh, M. (2024). Intrusion Detection System Using Machine Learning Based on NSL KDD Dataset. *2024 International Conference on Recent Advances in Science and Engineering Technology (ICRASET)*, 1–5.

Ishaque, M., Johar, M. G. M., Khatibi, A., & Yamin, M. (2023). Dynamic Adaptive Intrusion Detection System Using Hybrid Reinforcement Learning. *International Conference on Business and Technology*, 245–253.

Jayaprakash, J. S., Kodati, S., Kanchana, A., Al-Farouni, M., & AC, R. (2024). An Effective Cyber Security Threat Detection in Smart Cities Using Dueling Deep Q Networks. *2024 4th International Conference on Mobile Networks and Wireless Communications (ICMNWC)*, 1–5.

Khraisat, A. (2020). Hybrid intrusion detection system based on the stacking ensemble of C5 decision tree classifier and one class support vector machine. *Electronics (Switzerland)*, *9*(1). https://doi.org/10.3390/electronics9010173

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

Lin, Y.-D., Huang, H.-X., Sudyana, D., & Lai, Y.-C. (2024). AI for AI-based intrusion detection as a service: Reinforcement learning to configure models, tasks, and capacities. *Journal of Network and Computer Applications*, *229*, 103936.

Massaoudi, M., Eddin, M. E., Abu-Rub, H., & Ghrayeb, A. (2024). Advanced Proximal Policy Optimization Strategy for Resilient Cyber-Physical Power Grid Stability Against Hostile Electrical Disruptions. *IECON 2024-50th Annual Conference of the IEEE Industrial Electronics Society*, 1–6.

Mazini, M. (2019). Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms. *Journal of King Saud University - Computer and Information Sciences*, *31*(4), 541–553. https://doi.org/10.1016/j.jksuci.2018.03.011

Mazyavkina, N., Sviridov, S., Ivanov, S., & ... (2021). Reinforcement learning for combinatorial optimization: A survey. *Computers &Operations …*. https://www.sciencedirect.com/science/article/pii/S0305054821001660

Mishra, N., & Mishra, S. (2024). NSL-KDD Dataset Analysis: A Machine Learning Implementation to Detect Intrusions in the Computer Network. *2024 2nd International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, 1–6.

Mishra, P., Pilli, E. S., & Joshi, R. C. (2021). *Cloud security: attacks, techniques, tools, and challenges*. Chapman and Hall/CRC.

Mizher, M. Z., & Nassif, A. B. (2024). Enhanced Intrusion Detection in Cloud Security by Optimizing Classification Algorithms. *2024 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 296–303.

Mnih, V. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529–533. https://doi.org/10.1038/nature14236

Mnih, V. (2016). Asynchronous methods for deep reinforcement learning. In *33rd International Conference on Machine Learning, ICML 2016* (Vol. 4, pp. 2850–2869). https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=84999036937&origin=inward

Mo, K., Ye, P., Ren, X., Wang, S., Li, W., & Li, J. (2024). Security and privacy issues in deep reinforcement learning: Threats and countermeasures. *ACM Computing Surveys*, *56*(6), 1–39.

Nguyen, T. T. (2023). Deep Reinforcement Learning for Cyber Security. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 3779–3795. https://doi.org/10.1109/TNNLS.2021.3121870

Nguyen, T. T., & Reddi, V. J. (2021). Deep reinforcement learning for cyber security. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(8), 3779-3795.

Ozkan-Okay, M., Samet, R., Aslan, Ö., & Gupta, D. (2021). A comprehensive systematic literature review on intrusion detection systems. *IEEE Access*, *9*, 157727–157760.

Sangoleye, F., Johnson, J., & Tsiropoulou, E. E. (2024). Intrusion detection in industrial control systems based on deep reinforcement learning. *IEEE Access*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *ArXiv Preprint ArXiv:1707.06347*.

Selvam, R., & Velliangiri, S. (2024). An improving intrusion detection model based on novel CNN technique using recent CIC-IDS datasets. *2024 International Conference on Distributed Computing and Optimization Techniques (ICDCOT)*, 1–6.

Sujatha, V., Prasanna, K. L., Niharika, K., Charishma, V., & Sai, K. B. (2023). Network intrusion detection using deep reinforcement learning. *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, 1146–1150.

Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). A detailed analysis of the KDD CUP 99 data set. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 1–6.

Tellache, A., Mokhtari, A., Korba, A. A., & Ghamri-Doudane, Y. (2024). Multi-agent Reinforcement Learning-based Network Intrusion Detection System. *NOMS 2024-2024 IEEE Network Operations and Management Symposium*, 1–9.

Thana-Aksaneekorn, C., Kosolsombat, S., & Luangwiriya, T. (2024). Machine Learning Classification for Intrusion Detection Systems Using the NSL-KDD Dataset. *2024 IEEE*

International Conference on Cybernetics and Innovations (ICCI), 1–6.

Uppamma, P., & Bhattacharya, S. (2024). Edge Computing-Based Intrusion Detection Systems: A Review of Applications, Challenges, and Opportunities. *Intelligent Systems and Sustainable Computational Models*, 117–135.

Xiang, G., Dian, S., Du, S., & Lv, Z. (2023). Variational information bottleneck regularized deep reinforcement learning for efficient robotic skill adaptation. *Sensors*, *23*(2), 762.

Yang, J., Shi, J., Kuang, P., Feng, Z., Xiong, K., & Shi, Y. (2024). Enhancing Cybersecurity: A Proximal Policy Optimization Approach for Security Policy Optimization. *Proceedings of the 2024 8th International Conference on Computer Science and Artificial Intelligence*, 614–620.

Zakharenkov, A., & Makarov, I. (2021). Deep reinforcement learning with dqn vs. ppo in vizdoom. *2021 IEEE 21st International Symposium on Computational Intelligence and Informatics (CINTI)*, 131–136.

Zhu, Z., Chen, M., Zhu, C., & Zhu, Y. (2024). Effective defense strategies in network security using improved double dueling deep Q-network. *Computers & Security*, *136*, 103578.

Zou, B., Zhu, X., Liu, C., Nie, P., & Yu, Q. (2024). Enhanced intrusion strategy learning for security systems using an optimized PPO algorithm. *14th International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE 2024), 2024*, 845–849.