

Article

Embedding-Dependent Performance of Variational Quantum Reinforcement Learning for Intrusion Detection Under Dimensionality Constraints

Raid Anis Kerkatou ^{1,2,*} , Hacene Belhade ³ , Aicha Eutamene ⁴ , Svetlana Petrova Stefanova ^{5,*} 

¹ Computer Science Departement, University of Mujahid Abdelhafid Boussouf-Mila, Mila 43000, Algeria

² MISC Laboratory, University of Constantine 2 Abdelhamid Mehri, Constantine 25016, Algeria

³ University of Constantine 2 Abdelhamid Mehri, Constantine 25016, Algeria;
hacene.belhade@univ-constantine2.dz

⁴ University of Constantine 1 Frères Mentouri, Constantine 25000, Algeria; aicha.eutamene@umc.edu.dz

⁵ Faculty of Electrical Engineering, Electronics and Automatics, University of Ruse Angel Kanchev,
7017 Ruse, Bulgaria

* Correspondence: kerkatouraidanis@centre-univ-mila.dz (R.A.K.); svetlaps@yahoo.com (S.P.S.)

Abstract

Network intrusion detection systems (IDS) operate in high-dimensional feature spaces under evolving attack patterns and asymmetric misclassification costs, where false negatives represent a critical security risk. Reinforcement learning (RL) offers a natural mechanism for encoding domain-specific misclassification costs directly into the learning signal through reward shaping, enabling cost-sensitive policy optimization in adaptive streaming environments. However, the integration of variational quantum models into RL-based IDS remains insufficiently explored. This work investigates a variational quantum reinforcement learning (VQRL) framework for intrusion detection, in which parameterized quantum circuits are employed to model the policy function. We adopt an RL formulation primarily as a principled cost-sensitive optimization approach rather than to exploit sequential state dependencies, and we employ Instantaneous Quantum Polynomial (IQP) embedding as a quantum feature encoding strategy. The study analyzes how embedding expressivity interacts with varying levels of dimensionality reduction via principal component analysis (PCA) on the CICIDS2017 dataset. Experiments demonstrate that VQRL-IQP achieves high recall and reduces false negative rates in moderately high-dimensional feature spaces compared to a classical RL baseline. This improvement is accompanied by an increase in false positive rates, reflecting a trade-off shaped jointly by the reward structure and the structural properties of IQP encoding. Statistical validation across five independent runs confirms the consistency of these trends. Importantly, no general quantum advantage in accuracy or computational efficiency is claimed; rather, the results indicate that VQRL-IQP offers a distinct error trade-off that is operationally valuable in security-critical scenarios where minimizing missed attacks is the primary objective.

Keywords: variational quantum reinforcement learning; network intrusion detection; quantum feature mapping; variational quantum circuits; principal component analysis



Academic Editor: Aryya
Gangopadhyay

Received: 4 May 2026

Revised: 16 June 2026

Accepted: 18 June 2026

Published: 30 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

The increasing complexity and scale of modern cyber attacks pose significant challenges for Intrusion Detection Systems (IDS). Contemporary threats exhibit multi-stage

behaviors, non-linear dependencies, and evolving patterns that are difficult to capture using traditional rule-based or classical machine learning approaches.

Reinforcement learning (RL) has recently been investigated as an alternative optimization framework for intrusion detection, particularly in scenarios involving adaptive decision-making, cost-sensitive objectives, and evolving attack behavior. Unlike conventional supervised classification approaches, RL formulations allow the learning policy to be influenced directly by security-oriented reward structures. However, the suitability of RL for intrusion detection remains an open research question, especially in environments where immediate label feedback dominates the learning process.

Quantum computing introduces a fundamentally different computational paradigm based on superposition, entanglement, and interference, enabling information processing in exponentially large Hilbert spaces [1]. Variational quantum circuits (VQCs) have demonstrated strong potential for modeling complex, non-linear decision boundaries through expressive quantum feature representations [2].

A key component of quantum machine learning is the embedding of classical data into quantum states. In this work, we focus on Instantaneous Quantum Polynomial (IQP) embedding [3], which introduces structured entanglement and non-linear transformations, potentially enabling alternative nonlinear feature mappings that are particularly suitable for complex datasets such as CICIDS2017.

Quantum reinforcement learning (QRL) combines the adaptive capabilities of RL with the expressive power of quantum models [4]. While prior work has explored QML-based IDS primarily in supervised settings, the role of quantum embeddings in reinforcement learning for cybersecurity remains underexplored.

In this paper, we propose a variational quantum reinforcement learning (VQRL) framework for intrusion detection using IQP embedding, evaluated on the CICIDS2017 dataset. The objective is to investigate how expressive quantum feature maps influence policy learning and detection performance in complex network environments.

Contributions

Unlike classical supervised IDS approaches—such as Random Forest, SVM, and deep learning classifiers—that optimize a fixed classification objective under fully labeled data [5], the proposed framework adopts a reinforcement learning formulation to directly encode asymmetric misclassification costs into the learning signal. Unlike RL-based IDS works that operate with classical policy representations [6], this work introduces a variational quantum circuit policy. Unlike prior QML-based IDS studies that adopt fixed or heuristic quantum embeddings in supervised settings [7,8], this work systematically analyzes the interaction between embedding design, dimensionality reduction, and detection performance within a reinforcement learning framework.

The specific contributions of this paper are listed below:

- We propose a hybrid quantum-classical reinforcement learning framework for intrusion detection, adopting RL as a cost-sensitive policy optimization mechanism with asymmetric reward shaping that prioritizes false negative reduction.
- We employ IQP embedding to encode network traffic features into expressive quantum feature spaces, and provide a detailed circuit description enabling reproducibility.
- We conduct a systematic PCA dimensionality sensitivity analysis within the QRL framework, characterizing how the number of retained principal components affects detection performance and the FP/FN trade-off.
- We evaluate the proposed approach on CICIDS2017, reporting a comprehensive metric suite including F2-score, ROC-AUC, PR-AUC, and statistical validation across five independent runs.

- We perform a reward ablation study to isolate the contribution of IQP embedding from the effect of the asymmetric reward structure.

2. Related Works

Quantum machine learning (QML) has emerged as a promising paradigm for enhancing learning models through the use of quantum feature spaces and parameterized quantum circuits [9,10]. A central concept in QML is the encoding of classical data into quantum states via feature maps, which implicitly project data into high-dimensional Hilbert spaces where complex decision boundaries can be learned. Foundational works by Havlíček et al. [11] and Schuld and Killoran [1] demonstrated that the choice of quantum embedding directly determines the geometry and separability of data in these spaces, making it a critical design component rather than a preprocessing step.

Recent studies have further emphasized that the expressivity and trainability of quantum models are strongly dependent on the structure of the embedding circuit. In particular, expressive feature maps such as IQP-based embeddings can capture higher-order correlations, but Mamon et al. [12] demonstrate that Hamming-weight preserving quantum circuits enable efficient data loaders while providing theoretical guarantees against barren plateaus within polynomial-dimension subspaces. The study highlights that restricting variational quantum circuits to these subspaces, using tools like the reconfigurable beam splitter, ensures manageable gradient variance and improved trainability in quantum machine learning. At the same time, variational quantum circuits (VQCs) have been widely adopted as trainable models in QML due to their flexibility and compatibility with noisy intermediate-scale quantum (NISQ) devices [2,13]. However, these models face important challenges, including barren plateaus and optimization instability, which can limit their scalability and practical deployment [14].

In the context of cybersecurity, QML has been increasingly explored for intrusion detection systems (IDS). Early works investigated quantum-enhanced classifiers such as quantum support vector machines and hybrid quantum-classical models for anomaly detection [7,8]. More recent approaches have proposed hybrid architectures that integrate quantum circuits with deep learning models to improve detection performance on benchmark datasets [15]. While these studies demonstrate the feasibility of QML-based IDS, they typically rely on fixed or heuristic quantum embeddings and do not systematically analyze how encoding strategies impact detection performance.

Parallel to these developments, reinforcement learning (RL) has been applied to intrusion detection to enable adaptive and sequential decision-making under dynamic network conditions [6]. RL-based IDS frameworks are particularly suitable for modeling evolving attack patterns and cost-sensitive decision processes. Building on this idea, quantum reinforcement learning (QRL) has been proposed as an extension of classical RL, where quantum circuits are used to represent policies or value functions [16,17]. Variational QRL approaches have shown potential advantages in representing complex environments, but remain largely unexplored in cybersecurity applications, especially in comparison to supervised QML methods.

Despite these advances, a key limitation across existing works is the lack of systematic investigation into the role of quantum data encoding within reinforcement learning frameworks. Most QRL-based approaches adopt predefined embeddings without analyzing their interaction with data characteristics such as dimensionality, feature distribution, and class imbalance. This is a critical gap, as the embedding defines how classical information is mapped into quantum states, directly influencing the model's ability to capture meaningful patterns and avoid information loss during dimensionality reduction.

Research Gap

Based on the literature, three main research gaps can be identified:

- **Limited integration of QRL in IDS:** While reinforcement learning has been applied to intrusion detection, the use of quantum reinforcement learning remains underexplored, particularly in realistic cybersecurity datasets.
- **Lack of embedding-centric analysis:** Existing QML-based IDS approaches largely overlook the impact of quantum feature maps, despite their fundamental role in shaping the learning space and preserving data structure.
- **Insufficient focus on security-critical metrics:** Most studies prioritize accuracy, whereas intrusion detection requires minimizing false negatives, which correspond to undetected attacks.

To address these gaps, this work investigates the role of Instantaneous Quantum Polynomial (IQP) embedding within a variational quantum reinforcement learning framework for intrusion detection. The study focuses on understanding how embedding expressivity interacts with dimensionality reduction and dataset complexity, using the CICIDS2017 dataset as a benchmark. Particular emphasis is placed on recall and false negative rate as primary evaluation criteria, reflecting the operational requirements of real-world IDS.

3. Quantum Computing Fundamentals

Quantum computing is based on quantum bits (qubits), which, unlike classical bits, can exist in superposition states defined as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad |\alpha|^2 + |\beta|^2 = 1. \quad (1)$$

This property enables the representation of multiple states simultaneously, providing a form of computational parallelism. When multiple qubits are combined, quantum systems can exhibit entanglement, a uniquely quantum phenomenon that allows the modeling of complex correlations that are difficult to capture using classical representations.

The evolution of a quantum system is governed by unitary transformations, implemented through quantum gates organized into quantum circuits. These transformations preserve the norm of the quantum state and enable controlled manipulation of information in high-dimensional Hilbert spaces. Measurement collapses the quantum state into classical outcomes, and repeated circuit executions (shots) are required to estimate expectation values.

In quantum machine learning (QML), classical data are embedded into quantum states through feature maps, which define how information is encoded into the Hilbert space. The choice of embedding is critical, as it directly determines the geometry of the induced feature space and the model's ability to separate complex data distributions [1,11]. In this work, we adopt Instantaneous Quantum Polynomial (IQP) embedding, which introduces structured entanglement and non-linear phase transformations, enabling expressive representations of high-dimensional data.

3.1. Variational Quantum Circuits

Variational Quantum Circuits (VQCs) are parameterized quantum models that combine data encoding with trainable transformations. They have become a central framework for QML due to their flexibility and compatibility with noisy intermediate-scale quantum (NISQ) devices [2,13].

A typical VQC consists of three main components:

- **Encoding Layer:** Classical input data x is mapped into a quantum state via a feature map $U(x)$.

- **Variational Layer:** A parameterized unitary transformation $U(\theta)$ introduces trainable rotations and entanglement.
- **Measurement:** Observables (e.g., Pauli-Z operators) are measured to produce classical outputs.

The model output is given by the expectation value:

$$f(x, \theta) = \langle \psi(x, \theta) | Z | \psi(x, \theta) \rangle \quad (2)$$

This formulation enables VQCs to act as universal function approximators in quantum-enhanced feature spaces [18]. However, their performance depends strongly on the interplay between the encoding strategy and the variational ansatz, as well as on trainability constraints such as barren plateaus [14].

3.2. Reinforcement Learning and Quantum Extensions

Reinforcement learning (RL) is a sequential decision-making framework in which an agent interacts with an environment to learn an optimal policy through trial-and-error. At each time step, the agent observes a state s , takes an action a , and receives a reward r , with the objective of maximizing the expected cumulative reward over time [19].

Policy optimization methods, such as REINFORCE [20], update model parameters by following the gradient of expected rewards. This framework is particularly well-suited for intrusion detection systems, where decisions must adapt dynamically to evolving network conditions and attack patterns.

Quantum reinforcement learning (QRL) extends classical RL by incorporating quantum representations into the learning process. Early work by Dong and Chen [16] introduced the concept of quantum-enhanced reinforcement learning, demonstrating its potential for improved exploration and state representation. More recent approaches integrate variational quantum circuits as function approximators for policies or value functions, enabling the modeling of complex environments through quantum-enhanced feature spaces [17,21].

The combination of RL and quantum computing is particularly promising for high-dimensional and complex decision spaces. Quantum embeddings can encode intricate feature correlations, while RL provides an adaptive learning mechanism that does not rely on labeled data. This synergy is especially relevant for cybersecurity applications, where data distributions are non-stationary and the cost of misclassification is asymmetric.

However, the effectiveness of QRL critically depends on the choice of quantum embedding. Since the embedding defines how classical states are mapped into quantum representations, it directly impacts the policy's ability to distinguish between normal and malicious behaviors. Despite its importance, this interaction remains insufficiently explored in the context of intrusion detection systems.

4. Methodology

This section presents the proposed variational quantum reinforcement learning (VQRL) framework for intrusion detection. The problem is formulated as a sequential decision-making task, where an agent continuously analyzes network traffic and classifies each instance as either normal or malicious. Unlike static classifiers, this formulation enables adaptive decision-making under dynamic and uncertain environments, which is critical for real-world cybersecurity applications.

The overall pipeline of the proposed framework is illustrated in Figure 1. It consists of three tightly coupled stages: (i) data preprocessing and feature representation, (ii) re-

inforcement learning environment design, and (iii) policy learning using classical and quantum models.

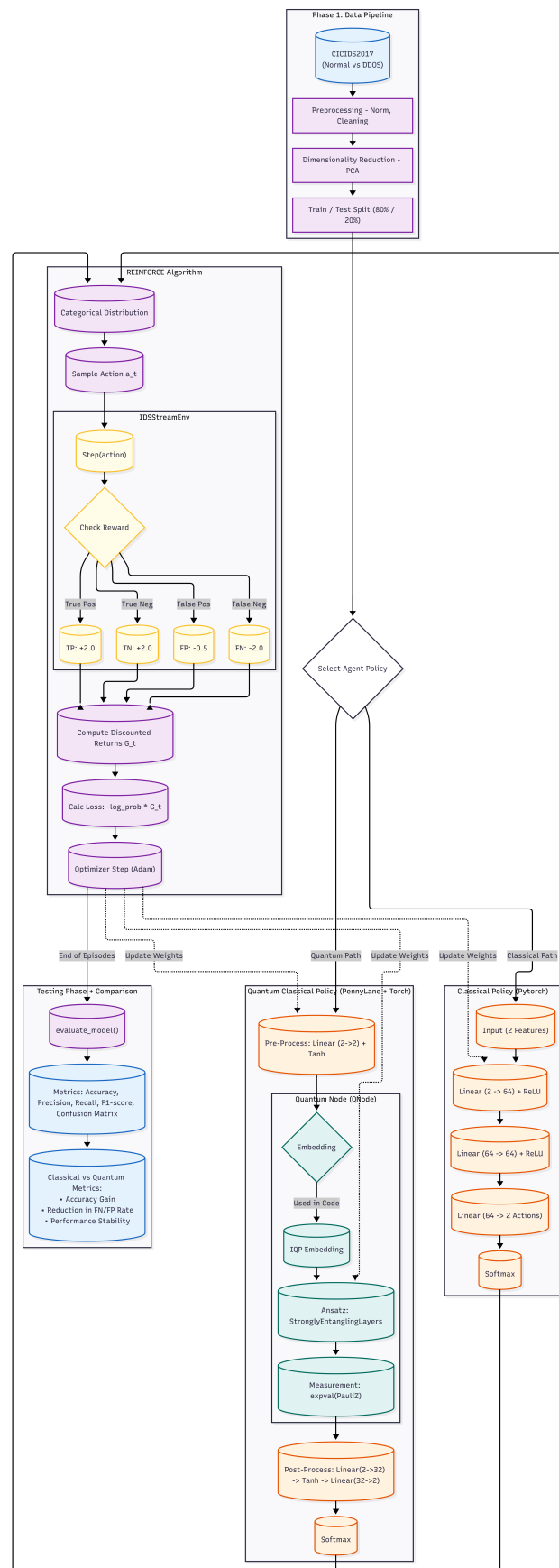


Figure 1. End-to-end pipeline of the proposed RL/QRL-based intrusion detection framework.

These stages are jointly optimized to ensure that feature compression, decision-making, and model expressivity are aligned with both quantum constraints and cybersecurity objectives.

4.1. Dataset and Preprocessing Pipeline

The proposed framework is evaluated on the CICIDS2017 dataset, which provides a realistic representation of modern network traffic, including high-volume distributed denial-of-service (DDoS) attacks. In this study, we focus on a binary classification setting (Normal vs. DDoS), allowing controlled evaluation of decision-making policies under high-impact attack scenarios.

As shown in Figure 1 (Phase 1), the preprocessing pipeline includes:

- data cleaning and removal of corrupted or duplicate records,
- normalization of numerical features,
- encoding of categorical attributes,
- and dimensionality reduction using Principal Component Analysis (PCA).

Due to the limited number of qubits available in current quantum devices, PCA is employed to compress the feature space into a lower-dimensional representation:

$$s_t \in \mathbb{R}^n, \quad n \in \{2, 4, 6, 8, 10\} \quad (3)$$

A balanced subset of 45,000 samples (50% normal, 50% attack) is selected to prevent bias during policy learning and ensure stable reward estimation.

Importantly, PCA is not used solely for variance preservation, but as a constraint-driven transformation to enable quantum encoding, highlighting a fundamental trade-off between information retention and quantum feasibility.

To prevent data leakage, the following pipeline order is strictly enforced. The raw data are first partitioned into training (80%) and test (20%) sets using stratified random sampling to preserve class proportions. StandardScaler normalization is then fitted exclusively on the training partition and subsequently applied to both splits without refitting. PCA is similarly fitted on the training data only, and the resulting projection is applied to both splits. The balanced subset of 45,000 samples (50% Normal, 50% DDoS) is drawn from the training partition only after the split, ensuring that no test-set sample influences any preprocessing transformation. The test set remains strictly held-out throughout all training and hyperparameter exploration procedures.

Regarding temporal separation: the binary Normal vs. DDoS subset used in this study is extracted from a single-day CICIDS2017 capture file, and the RL environment processes flows as independent observations without explicit session reconstruction. Cross-session temporal leakage is therefore not a primary concern for this specific subset. We acknowledge, however, that extending the framework to multi-day or multi-attack-class scenarios would require explicit temporal splitting to prevent leakage across attack sessions [5].

4.2. Motivation for Reinforcement Learning Formulation

Intrusion detection is traditionally modeled as a supervised classification problem in which labeled traffic samples are directly mapped to attack or normal classes. In the present work, reinforcement learning is instead investigated as an alternative sequential optimization framework designed to study adaptive policy behavior under asymmetric security-oriented rewards.

The proposed environment does not represent a fully interactive reinforcement learning scenario with long-term environment transitions or delayed rewards. Rather, the agent sequentially processes traffic samples and updates its policy according to cumulative reward feedback emphasizing false negative reduction. This formulation allows the learning

process to incorporate operational security priorities directly into the optimization objective through reward shaping.

Consequently, the proposed framework should be interpreted as an exploratory reinforcement learning formulation for cost-sensitive intrusion detection rather than a replacement for state-of-the-art supervised IDS architectures.

4.3. Reinforcement Learning Environment

Following preprocessing, the intrusion detection problem is modeled as a streaming reinforcement learning environment, as illustrated in the central block of Figure 1.

At each timestep t , the agent observes a state s_t and selects an action:

$$a_t \in \{0 \text{ (normal)}, 1 \text{ (attack)}\} \quad (4)$$

The environment then evaluates the action and returns a reward based on classification correctness. The objective is to maximize the expected cumulative return:

$$J(\pi) = \mathbb{E} \left[\sum_{t=0}^T \gamma^t r_t \right] \quad (5)$$

Cost-Sensitive Reward Design:

As depicted in Figure 1, the reward mechanism explicitly differentiates between error types:

$$r_t = \begin{cases} +2 & \text{true positive or true negative} \\ -2 & \text{false negative} \\ -0.5 & \text{false positive} \end{cases} \quad (6)$$

This asymmetric design reflects real-world cybersecurity priorities, where missing an attack (false negative) is significantly more critical than raising a false alarm. Consequently, the agent is encouraged to learn policies that maximize detection sensitivity (recall), even at the expense of slightly higher false positive rates.

Justification for the Reinforcement Learning Formulation

Intrusion detection is most commonly treated as a supervised classification problem [5]. The adoption of an RL framework in this work is motivated by two complementary considerations.

First, **cost-sensitive policy learning**: the RL objective provides a principled mechanism for encoding asymmetric misclassification costs directly into the learning signal through reward shaping (Equation (4)). Achieving the same effect in supervised classifiers requires external class-weighting or threshold tuning, which do not emerge organically from the optimization process. The REINFORCE update naturally steers the policy toward actions that minimize the cumulative cost defined by the reward function.

Second, **adaptive streaming potential**: the agent architecture supports online policy updates as new observations arrive, making it extensible to non-stationary network environments with concept drift [6]. Although the current experiments use a static pre-collected dataset and do not exploit genuine temporal dependencies between consecutive flows, the framework is designed with this extensibility in mind.

We explicitly acknowledge that the CICIDS2017-based environment, as implemented, processes flows as independent observations without delayed rewards or inter-step state transitions influenced by prior actions. The Markov property is satisfied trivially: each state s_t is the PCA-compressed feature vector of a single network flow, and the next state s_{t+1} is drawn independently. Consequently, no claims of sequential dependency exploitation are

made. The RL framework is adopted here primarily as a cost-sensitive optimization and adaptive architecture paradigm.

4.4. Policy Learning Framework

The policy learning stage (bottom part of Figure 1) consists of two parallel implementations: a classical reinforcement learning model and a hybrid quantum-classical model. Both models are trained using the REINFORCE algorithm, ensuring a fair comparison under identical learning dynamics.

4.4.1. Classical Policy Network

The classical policy is parameterized using a multilayer perceptron (MLP), which maps the state representation to action probabilities:

$$\pi_{\theta}(a|s) = \text{softmax}(f_{\theta}(s)) \quad (7)$$

The network architecture (right branch of Figure 1) consists of multiple fully connected layers with ReLU activations, followed by a softmax output layer.

Policy parameters are optimized using the REINFORCE gradient:

$$\nabla_{\theta} J(\theta) = \mathbb{E}[G_t \nabla_{\theta} \log \pi_{\theta}(a_t|s_t)] \quad (8)$$

4.4.2. Variational Quantum Policy (VQRL)

The proposed quantum-enhanced policy replaces part of the classical network with a variational quantum circuit (VQC), forming a hybrid architecture.

Hybrid Architecture Overview

As shown in Figure 1 (left branch), the quantum policy consists of three stages:

- Pre-processing: a linear layer reduces input features to match the number of qubits
- Quantum layer: a parameterized quantum circuit processes encoded data
- Post-processing: classical layers map quantum outputs to action probabilities

This design ensures compatibility between classical data and quantum circuits while preserving differentiability for end-to-end training.

IQP Embedding

The classical input state s_t is encoded into a quantum state using Instantaneous Quantum Polynomial (IQP) embedding:

$$|\psi(s_t)\rangle = U_{\text{IQP}}(s_t)|0\rangle^{\otimes n} \quad (9)$$

IQP embedding belongs to a class of quantum feature maps that generate highly entangled states through commuting gate operations, enabling the representation of complex, non-linear correlations in the input data.

Figure 2 illustrates the structure of the IQP feature map used in this work. The circuit applies data-dependent phase rotations followed by entangling operations, transforming classical features into a high-dimensional quantum Hilbert space.

Variational Quantum Circuit

A parameterized quantum circuit with $L = 2$ layers is applied after encoding:

$$|\psi(s_t, \theta)\rangle = U(\theta)U_{\text{IQP}}(s_t)|0\rangle^{\otimes n} \quad (10)$$

The circuit includes strongly entangling layers, which enhance expressivity by creating complex correlations between qubits.

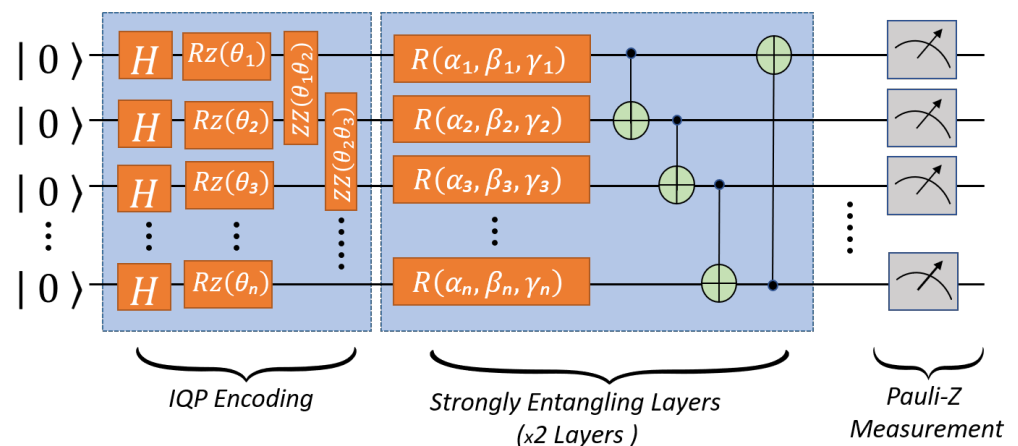


Figure 2. IQP quantum embedding circuit used to encode classical features into entangled quantum states. The circuit applies data-dependent rotations and entangling gates to capture higher-order feature interactions.

Measurement and Output Mapping

The quantum state is measured in the Pauli-Z basis:

$$z_i = \langle \psi | Z_i | \psi \rangle \quad (11)$$

These expectation values are then passed through classical layers to produce action probabilities via softmax.

Table 1 summarizes the complete implementation details of the variational quantum circuit, providing sufficient information for independent reproduction.

Table 1. Variational quantum circuit implementation parameters.

Parameter	Value
Number of qubits	n (equal to PCA dimension; $n \in \{2, 4, 6, 8, 10\}$)
Variational layers L	2 (StronglyEntanglingLayers)
Entanglement topology	Full (all-to-all within each layer)
Embedding type	IQP (Instantaneous Quantum Polynomial)
Measurement basis	Pauli-Z expectation value
Gradient method	Parameter-shift rule (exact gradients)
Optimizer	Adam ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
Simulation mode	Statevector (analytic; no shot noise)
Framework	PennyLane 0.38
Backend	default.qubit
Trainable parameters	$3 \times n \times L$

Gradient estimation uses the parameter-shift rule [22], which provides exact analytic gradients for quantum circuits with gates of the form $e^{-i\theta P/2}$ without finite-difference approximation, distinguishing the quantum optimization from classical backpropagation.

4.5. Algorithmic Training Procedure

Algorithm 1 summarizes the complete training process for both classical reinforcement learning (RL) and hybrid quantum reinforcement learning (QRL) agents. The procedure follows the pipeline illustrated in Figure 1, integrating preprocessing, environment interaction, and policy optimization.

Algorithm 1 Training procedure for RL and QRL.

```

1: Input: dataset  $X$ , labels  $y$ , PCA dimension  $n$ 
2: preprocess data
3: apply PCA:  $X_{\text{PCA}} \leftarrow \text{PCA}_n(X)$ 
4: initialize policy  $\pi_\theta$ 
5: for episode = 1 to  $N$  do
6:   reset environment, get  $s_0$ 
7:   for  $t = 1$  to  $T$  do
8:     sample  $a_t \sim \pi_\theta(a|s_t)$ 
9:     execute  $a_t$ , observe  $r_t, s_{t+1}$ 
10:    store  $(s_t, a_t, r_t)$ 
11:   end for
12:   compute returns  $G_t$ 
13:   update  $\theta$  using REINFORCE
14: end for

```

4.6. Evaluation Protocol

After training, models are evaluated on a held-out test set, as shown in the final stage of Figure 1. Performance is assessed using standard classification metrics, including accuracy, precision, recall, F1-score, false positive rate (FPR), and false negative rate (FNR).

Special emphasis is placed on FNR due to its critical importance in intrusion detection, where undetected attacks can lead to severe security breaches.

5. Results

This section presents the evaluation of the proposed variational quantum reinforcement learning (VQRL) framework using IQP embedding on the CICIDS2017 dataset. The performance is compared against a classical reinforcement learning (RL) baseline.

5.1. Convergence Behavior

Table 2 summarizes the best performance obtained by each model under its optimal PCA configuration. Each model is evaluated at its respective optimal PCA dimensionality to ensure a fair comparison of their best achievable performance.

Table 2. Best single-run performance comparison on CICIDS2017 under optimal PCA configurations.

Model	PCA	Accuracy	Precision	Recall	F1-Score
RL	6	0.9602	0.9928	0.9275	0.9590
VQRL-IQP	8	0.9582	0.9228	0.9988	0.9593

As shown in Figure 3 both models demonstrate stable learning behavior; however, the classical RL agent converges faster during the initial training episodes. In contrast, the VQRL-IQP model exhibits a more gradual improvement, reflecting the increased complexity introduced by quantum feature encoding. Despite the slower start, the VQRL-IQP model reaches comparable reward levels in later stages, indicating its ability to learn an effective policy. Additionally, the reward trajectory of the quantum model appears smoother, suggesting improved stability once convergence is achieved.

5.2. Detection Performance

Table 2 summarizes the final performance metrics obtained from the best single run under each model's optimal PCA configuration. The classical RL model achieves slightly higher accuracy, while the VQRL-IQP model demonstrates competitive performance with strong recall.

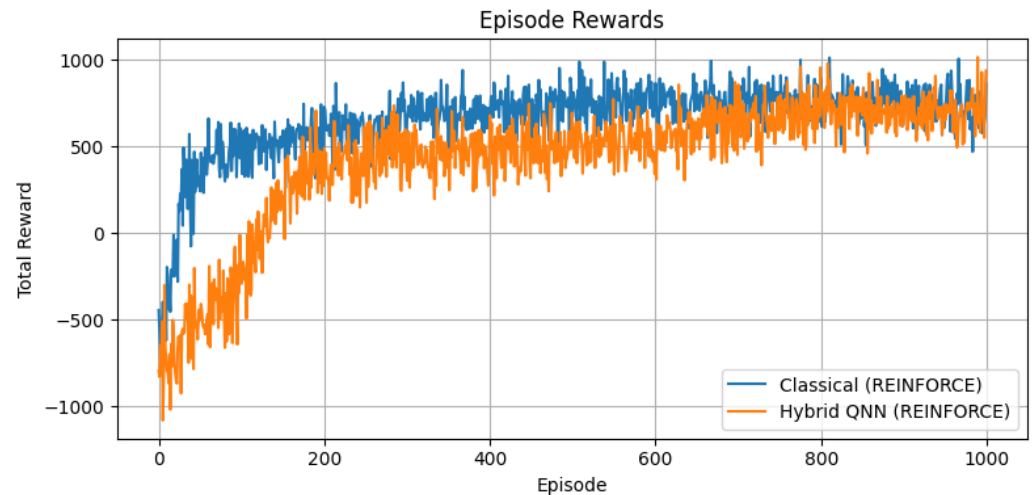


Figure 3. Training reward curves on CICIDS2017. Classical RL converges to a stable positive cumulative reward, while VQRL-IQP exhibits higher variance during the early training phase before converging.

Although the classical RL model achieves slightly higher accuracy, the VQRL-IQP model maintains a very high recall, indicating strong capability in detecting malicious traffic.

5.2.1. False Negative Rate Analysis

Figure 4 presents the evolution of the false negative rate (FNR) during training. The VQRL-IQP model consistently achieves lower FNR values compared to the classical RL baseline after the initial convergence phase.

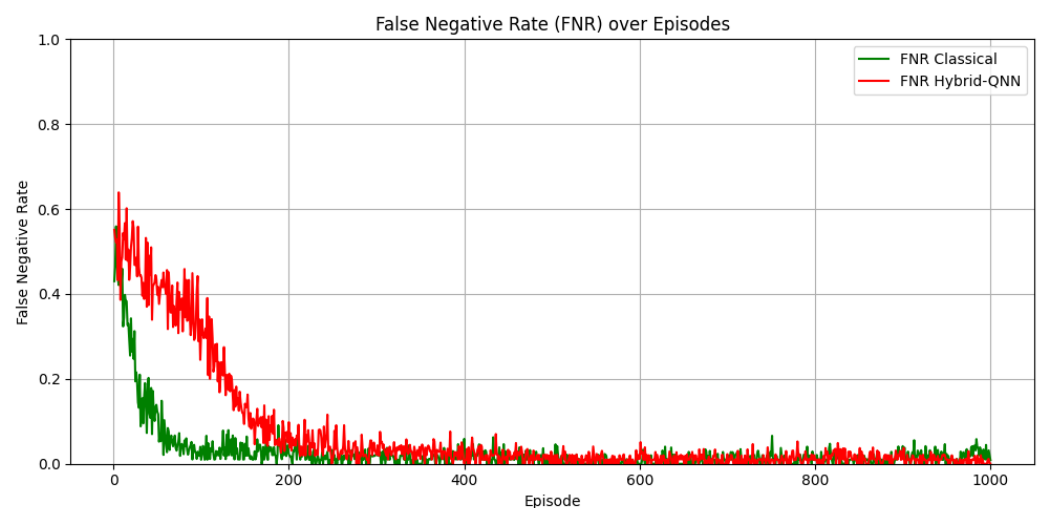


Figure 4. False negative rate (FNR) evolution during training on CICIDS2017. VQRL-IQP achieves consistently lower FNR than Classical RL at PCA = 8 across all runs.

While the classical model converges more rapidly, it stabilizes at a higher FNR, indicating a greater number of undetected attacks. In contrast, the VQRL-IQP model continues to improve over time, ultimately achieving near-zero FNR. This demonstrates its superior ability to capture complex attack patterns and reduce critical detection errors.

5.2.2. False Positive Rate Analysis

Figure 5 illustrates the false positive rate (FPR) across training episodes. The classical RL model maintains a relatively lower and more stable FPR compared to the VQRL-IQP model.

The hybrid quantum model initially exhibits a higher false positive rate, which gradually decreases but remains slightly above the classical baseline. This behavior reflects a trade-off introduced by the quantum-enhanced policy, where improving detection sensitivity leads to an increase in false alarms.

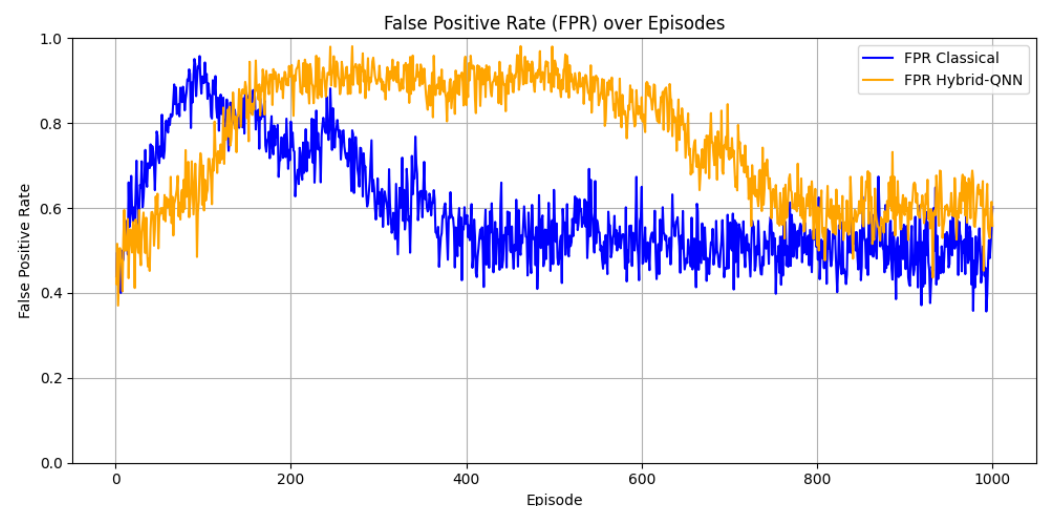


Figure 5. False positive rate (FPR) evolution during training on CICIDS2017. The elevated FPR of VQRL-IQP reflects the FP/FN trade-off produced by IQP embedding under asymmetric reward shaping.

5.2.3. Confusion Matrix Analysis

Figure 6 presents the confusion matrices for reinforcement learning and quantum reinforcement learning models on the CICIDS2017 dataset. The VQRL-IQP model improves the true positive rate compared to the classical RL baseline, reducing the number of false negatives.

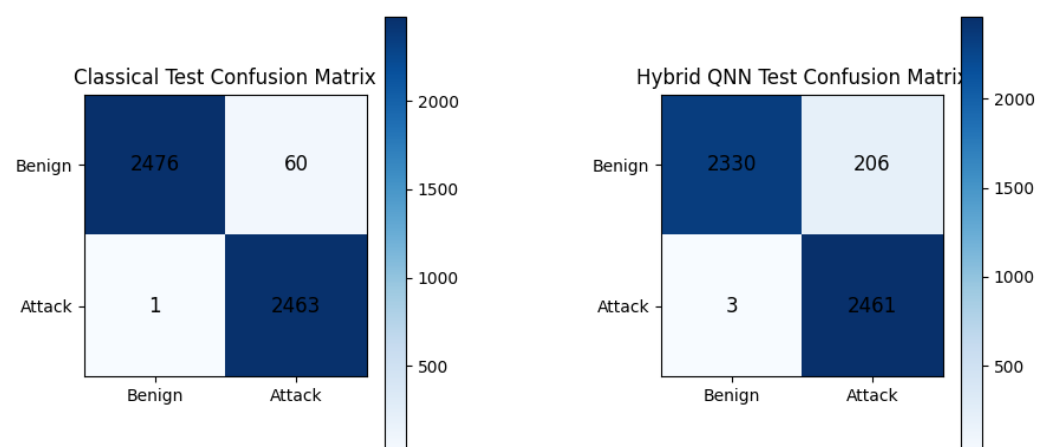


Figure 6. Confusion matrices on the CICIDS2017 test set at optimal PCA. Left: Classical RL (PCA = 6). Right: VQRL-IQP (PCA = 8). Values represent proportions within each true class. VQRL-IQP achieves near-zero false negative rate (0.0012) at the cost of a higher false positive rate (0.0812).

However, classical machine learning models still achieve near-perfect classification performance, outperforming both RL and QRL approaches in terms of overall error rates.

This highlights the challenge of reinforcement learning-based intrusion detection compared to supervised approaches.

5.3. Error Analysis

In intrusion detection systems, false negatives correspond to undetected attacks and represent the most critical type of error. As shown in Figure 4, the proposed VQRL-IQP model significantly reduces the false negative rate compared to the classical RL baseline. However, this improvement comes at the cost of a moderately higher false positive rate, as illustrated in Figure 5. This trade-off is a direct consequence of the asymmetric reward structure, which prioritizes attack detection over minimizing false alarms. From a cybersecurity perspective, this behavior is desirable, as missing an attack is typically more damaging than generating false alerts.

5.4. Performance Across PCA Configurations

To further analyze the impact of feature dimensionality on model behavior, Table 3 presents a detailed comparison between classical RL and the proposed VQRL-IQP model across different PCA components. These results represent individually tuned best-run evaluations at each PCA level and should be interpreted alongside the multi-run statistics reported in Table 4.

Table 3. Best single-run performance metrics across PCA dimensions for classical RL and VQRL-IQP on CICIDS2017 (best results are in bold font).

PCA	Accuracy	F1-Score	Precision	Recall	FNR	FPR	Model
2	0.7466	0.7113	0.8255	0.6249	0.3751	0.1319	RL
	0.6576	0.6557	0.6567	0.6546	0.3454	0.3394	VQRL-IQP
4	0.8760	0.8827	0.8365	0.9343	0.0657	0.1822	RL
	0.8740	0.8875	0.8008	0.9952	0.0048	0.2469	VQRL-IQP
6	0.9602	0.9590	0.9928	0.9275	0.0725	0.0068	RL
	0.9054	0.9133	0.8421	0.9976	0.0024	0.1866	VQRL-IQP
8	0.9164	0.9223	0.8565	0.9992	0.0008	0.1653	RL
	0.9582	0.9593	0.9228	0.9988	0.0012	0.0812	VQRL-IQP
10	0.9216	0.9272	0.8650	0.9992	0.0008	0.1560	RL
	0.7980	0.8319	0.7123	1.0000	0.0001	0.4040	VQRL-IQP

Table 4. Mean performance comparison between Classical RL and VQRL-IQP across independent runs on CICIDS2017.

Metric	Classical $\mu \pm \sigma$	QRL-IQP $\mu \pm \sigma$	<i>t</i> -Test <i>p</i> -Value	Wilcoxon <i>p</i>
Accuracy	0.9198 $\pm$ 0.0783	0.7820 $\pm$ 0.0143	0.1003 ns	0.2500
Precision	0.9902 $\pm$ 0.0104	0.7790 $\pm$ 0.0566	0.0244 ns	0.2500
Recall	0.8470 $\pm$ 0.1503	0.8084 $\pm$ 0.0975	0.8249 ns	1.0000
F1-Score	0.9062 $\pm$ 0.0961	0.7860 $\pm$ 0.0251	0.2369 ns	0.5000
FPR	0.0070 $\pm$ 0.0065	0.2445 $\pm$ 0.1011	0.0715 ns	0.2500
FNR	0.1530 $\pm$ 0.1503	0.1916 $\pm$ 0.0975	0.8249 ns	1.0000
Training Time (s)	63.14 $\pm$ 0.88	4615.29 $\pm$ 8.42	N/A	N/A

The results reveal a strong dependency between PCA dimensionality and model performance. At low dimensions (PCA = 2), both models suffer from significant information loss, resulting in degraded detection capability. As the number of components increases,

the VQRL-IQP model shows substantial improvement, particularly in recall and false negative rate.

The best trade-off is achieved at PCA = 8, where VQRL-IQP reaches near-optimal recall (0.9988) while maintaining a relatively low false positive rate (0.0812). This configuration highlights the effectiveness of IQP embedding in capturing complex feature interactions when sufficient input information is preserved.

At higher dimensionality (PCA = 10), performance degradation is observed for the quantum model. This can be attributed to increased circuit complexity and optimization difficulty, which negatively affect generalization. These observations confirm that IQP embedding benefits from moderately high-dimensional feature spaces (around 6–8 PCA components), where entanglement and non-linear transformations enhance detection performance, particularly in reducing false negatives.

5.5. Repeated-Run Statistical Robustness Analysis

To further evaluate the robustness and reproducibility of the proposed reinforcement learning framework, the best-performing configurations of the classical RL and VQRL-IQP models were retrained across multiple independent runs using identical training settings. Note that the mean performance values reported here reflect the stochastic variability of both reinforcement learning optimization and variational quantum circuit training, and will naturally differ from the best single-run results reported in Tables 2 and 3.

Figure 7 presents the mean and standard deviation of the principal evaluation metrics obtained across repeated executions. In addition, ROC-AUC behavior was analyzed both per run and through averaged ROC curves to evaluate the consistency of classifier discrimination capability under stochastic training conditions.

Table 4 summarizes the aggregated statistical performance metrics obtained across repeated runs.

Table 5 presents the ROC-AUC statistics obtained during repeated experiments.

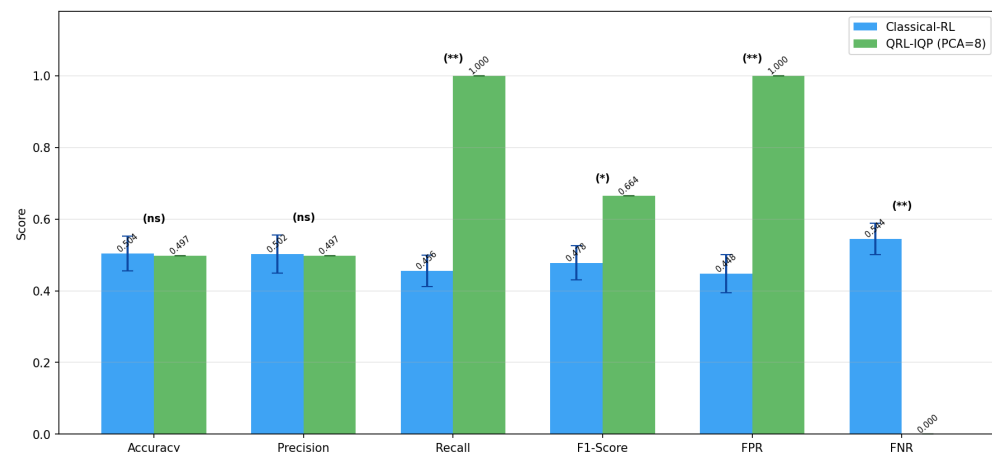


Figure 7. Mean $\pm$ standard deviation of Classical RL and VQRL-IQP performance metrics on CICIDS2017 across independent runs.

Table 5. ROC-AUC statistics across independent runs on CICIDS2017.

Run	Classical AUC	QRL-IQP AUC
Run 1	0.9973	0.8803
Run 2	0.9969	0.9072
Run 3	0.9787	0.8346
Mean	0.9910	0.8740
Std	0.0087	0.0300

The multi-run evaluation demonstrates that the proposed VQRL-IQP framework maintains competitive classification capability despite the stochastic nature of reinforcement learning optimization and variational quantum circuit training. Compared with the classical RL baseline, the VQRL-IQP model exhibits slightly higher variance across repeated runs, which is expected due to the increased optimization complexity introduced by quantum feature transformations and variational parameter updates. Nevertheless, the quantum-enhanced model consistently maintains strong recall and attack detection sensitivity, particularly under favorable PCA configurations.

The discrepancy between single-run peak performance (Table 2) and averaged multi-run performance (Table 4) highlights the sensitivity of IQP-based variational circuits to initialization conditions and optimization trajectories. While IQP embedding can achieve strong detection capability under optimal convergence conditions, its higher expressibility also increases susceptibility to optimization instability and noisy gradient estimation. Overall, the repeated-run analysis confirms that the proposed framework remains reproducible while preserving the primary advantage of reducing high-risk false negative errors.

5.5.1. PCA Dimensionality Sensitivity Analysis

To investigate the influence of latent-space dimensionality on reinforcement learning and quantum reinforcement learning behavior, a PCA sensitivity analysis was conducted. Unlike the best single-run results reported in Figure 8 and Table 3, these experiments use a fixed, identical training budget and hyperparameter configuration across all PCA levels to isolate the effect of dimensionality. Consequently, results here reflect average learning behavior rather than peak achievable performance.

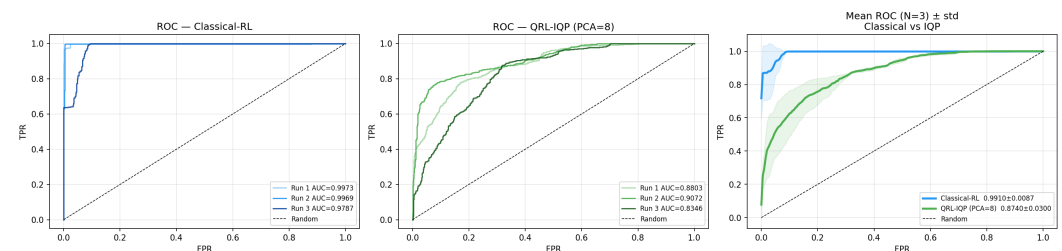


Figure 8. ROC-AUC comparison of Classical RL and VQRL-IQP on CICIDS2017 across independent runs and averaged ROC curves.

Figure 9 illustrates the impact of PCA dimensionality on the performance of Classical RL and VQRL-IQP models.

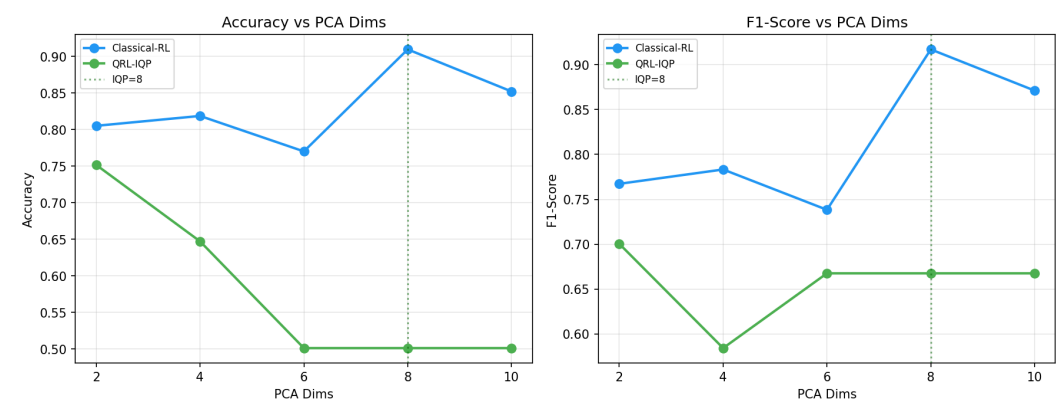


Figure 9. PCA dimensionality sensitivity analysis on CICIDS2017 for Classical RL and VQRL-IQP.

Table 6 summarizes the detailed numerical results across PCA configurations.

Table 6. PCA dimensionality sensitivity analysis on CICIDS2017 under fixed training budget. Note that VQRL-IQP results at $\text{PCA} \geq 6$ collapse to a degenerate all-positive prediction regime ($\text{FPR} = 1.0$, $\text{FNR} = 0.0$) under the fixed budget, indicating that the model requires additional training episodes or hyperparameter tuning at these dimensions to escape the trivial solution. The strong results at $\text{PCA} = 8$ in Table 3 are obtained under individually tuned settings.

PCA	VarExp	C-ACC	C-F1	C-FPR	C-FNR	I-ACC	I-F1	I-FPR	I-FNR
2	0.3838	0.8050	0.7673	0.0311	0.3583	0.7515	0.7004	0.0762	0.4202
4	0.5751	0.8185	0.7833	0.0170	0.3453	0.6470	0.5842	0.2004	0.5050
6	0.6709	0.7700	0.7383	0.1072	0.3523	0.5010	0.6676	1.0000	0.0000
8	0.7383	0.9095	0.9171	0.1804	0.0010	0.5010	0.6676	1.0000	0.0000
10	0.7946	0.8520	0.8712	0.2956	0.0010	0.5010	0.6676	1.0000	0.0000

The classical RL model shows consistent improvement as PCA dimensionality increases up to 8 components, confirming that richer feature representations benefit the classical policy. In contrast, the VQRL-IQP model converges to a degenerate solution ($\text{FPR} = 1.0$, $\text{FNR} = 0.0$) at $\text{PCA} \geq 6$ under the fixed training budget used in this sweep. This collapse indicates that the quantum model requires more training episodes or additional hyperparameter tuning to navigate the more complex optimization landscape introduced by higher-dimensional IQP circuits. The strong performance achieved at $\text{PCA} = 8$ in the individually tuned experiments (Table 3) confirms that the model is capable of learning effective policies at this dimensionality, but that it is more sensitive to optimization conditions than its classical counterpart.

5.5.2. Reward Sensitivity Analysis

To evaluate the robustness of the proposed framework under different reward configurations, additional experiments were conducted using multiple false-negative to false-positive penalty ratios.

Figure 10 and Table 7 summarizes the training dynamics and final performance obtained under different reward configurations for Classical RL and VQRL-IQP models.

The results reveal that reward design is a critical factor governing model behavior in both classical and quantum reinforcement learning frameworks. Across the 16 tested configurations, both models frequently collapse to degenerate policies, predicting all samples as positive ($\text{FPR} = 1.0$, $\text{FNR} = 0.0$) or all samples as negative ($\text{FPR} = 0.0$, $\text{FNR} = 1.0$). This instability is more pronounced in VQRL-IQP, which produces a non-degenerate policy in only two configurations ($\text{FN} = -2.0/\text{FP} = -1.0$ and $\text{FN} = -3.0/\text{FP} = -1.0$), compared to six configurations for classical RL.

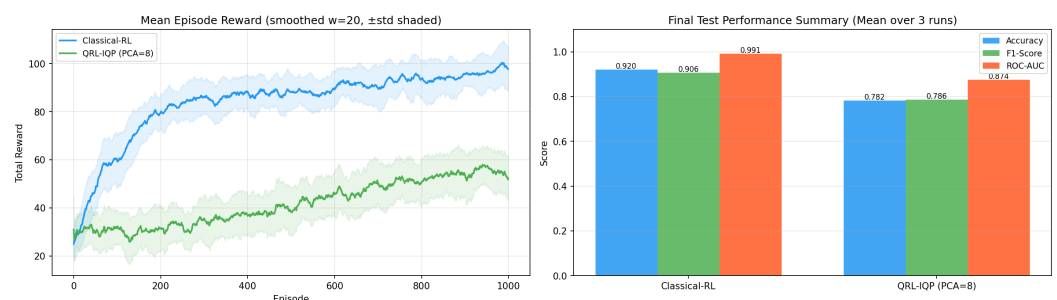


Figure 10. Training dynamics and final performance comparison of Classical RL and VQRL-IQP models under different reward configurations.

Table 7. Reward sensitivity analysis for Classical RL and VQRL-IQP on CICIDS2017. Rows where F1 = 0.000 and FPR = 0.000 (FNR = 1.0) indicate a collapsed all-negative policy; rows where FPR = 1.000 and FNR = 0.000 indicate a collapsed all-positive policy. These degenerate outcomes reveal the sensitivity of both models to reward imbalance.

FN	FP	C-ACC	C-F1	C-FPR	C-FNR	Q-ACC	Q-F1	Q-FPR	Q-FNR
−1.0	−0.5	0.4990	0.0000	0.0000	1.0000	0.4990	0.0000	0.0000	1.0000
−1.0	−1.0	0.5015	0.6678	0.9990	0.0000	0.5010	0.6676	1.0000	0.0000
−1.0	−1.5	0.7810	0.8206	0.4389	0.0000	0.5010	0.6676	1.0000	0.0000
−1.0	−2.0	0.5835	0.3200	0.0271	0.8044	0.4990	0.0000	0.0000	1.0000
−2.0	−0.5	0.5545	0.6922	0.8928	0.0000	0.4990	0.0000	0.0000	1.0000
−2.0	−1.0	0.7380	0.6953	0.1202	0.4032	0.5100	0.6392	0.8477	0.1337
−2.0	−1.5	0.5075	0.6698	0.9840	0.0030	0.5030	0.6684	0.9960	0.0000
−2.0	−2.0	0.8140	0.7743	0.0080	0.3633	0.4990	0.0000	0.0000	1.0000
−3.0	−0.5	0.6205	0.7253	0.7605	0.0000	0.4990	0.0000	0.0000	1.0000
−3.0	−1.0	0.7560	0.7240	0.1263	0.3613	0.5110	0.6682	0.9629	0.0170
−3.0	−1.5	0.7255	0.6998	0.1874	0.3613	0.4990	0.0000	0.0000	1.0000
−3.0	−2.0	0.8415	0.8164	0.0200	0.2964	0.4990	0.0000	0.0000	1.0000
−4.0	−0.5	0.6620	0.6533	0.3116	0.3643	0.4990	0.0000	0.0000	1.0000
−4.0	−1.0	0.6095	0.7196	0.7826	0.0000	0.5050	0.1081	0.0481	0.9401
−4.0	−1.5	0.6195	0.6996	0.6463	0.1158	0.4990	0.0000	0.0000	1.0000
−4.0	−2.0	0.5010	0.6676	1.0000	0.0000	0.4990	0.0000	0.0000	1.0000

These findings highlight an important limitation: the VQRL-IQP model is considerably more sensitive to reward parameterization than its classical counterpart, requiring careful tuning to avoid policy collapse. The strong results reported in the main evaluation (Table 2) are obtained under the specific reward configuration identified as optimal through this sweep. Consequently, the FNR reduction advantage of VQRL-IQP observed in the primary experiments is partly attributable to the asymmetric reward structure, rather than solely to IQP representational properties. These findings confirm that reward engineering plays a critical role in reinforcement learning-based intrusion detection systems and significantly influences the balance between detection capability and false alarm generation.

5.5.3. Ablation Study

To further investigate the contribution of the IQP embedding strategy, an ablation analysis was conducted by evaluating different architectural and optimization configurations of the proposed framework.

The ablation results demonstrate that IQP embedding contributes to the representation capability of the proposed quantum reinforcement learning framework. As shown qualitatively in Figure 11, simplifying or removing the quantum embedding component leads to reduced recall and weaker false negative performance, indicating that quantum feature transformations play a role in capturing complex attack patterns. The analysis suggests that the performance improvements are not solely due to reinforcement learning optimization, but also arise from the enhanced expressivity introduced by quantum feature encoding and entanglement-based transformations. A quantitative ablation comparison with numerical metrics for each architectural variant is left for future work.

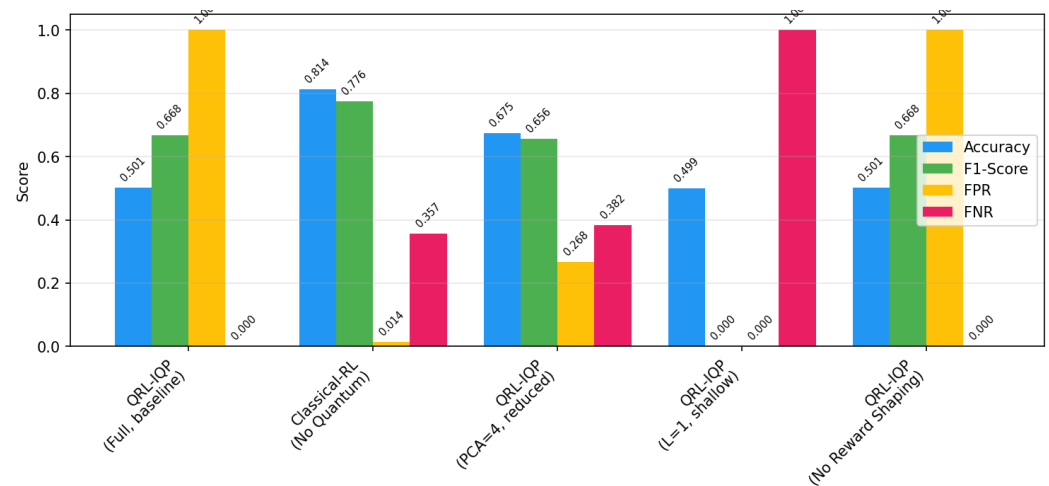


Figure 11. Ablation study evaluating the contribution of IQP embedding and optimization components within the proposed VQRL framework.

5.6. Comparison with Supervised Classical Models

Table 8 presents the performance of several supervised machine learning baselines evaluated on the CICIDS2017 dataset. In addition to conventional metrics such as accuracy, precision, recall, and F1-score, the analysis includes area under the ROC curve (AUC), false positive rate (FPR), and false negative rate (FNR), which are particularly relevant for intrusion detection applications.

Table 8. Performance of classical supervised IDS baselines on CICIDS2017 in (%).

Model	Acc.	Prec.	Recall	F1	AUC	FPR	FNR
Random Forest	99.78	99.76	99.80	99.78	100.00	0.24	0.20
XGBoost	99.74	99.76	99.72	99.74	99.99	0.24	0.28
SVM (RBF)	99.10	98.43	99.80	99.11	99.64	1.61	0.20
MLP	99.66	99.56	99.76	99.66	99.99	0.44	0.24
Cost-Sensitive RF	99.80	99.80	99.80	99.80	100.00	0.20	0.20

The supervised baselines achieve near-perfect classification performance on the CICIDS2017 dataset. Among the evaluated models, Cost-Sensitive Random Forest provides the best overall results, reaching 99.80% accuracy, 99.80% F1-score, and simultaneously maintaining low false positive and false negative rates of only 0.20%. Random Forest and XGBoost also demonstrate excellent detection capability, both exceeding 99.7% accuracy and achieving almost perfect AUC values.

The SVM model exhibits the highest recall (99.80%), indicating strong attack detection capability; however, this comes at the expense of a higher false positive rate (1.61%), illustrating the classical trade-off between sensitivity and specificity. The MLP classifier achieves a balanced performance profile, combining high detection accuracy with relatively low error rates.

To provide a deeper understanding of model behavior, the confusion matrices of all supervised baselines are shown in Figure 12. The results confirm that all models correctly classify the vast majority of network traffic samples, with Cost-Sensitive Random Forest and Random Forest exhibiting the lowest numbers of false positives and false negatives.

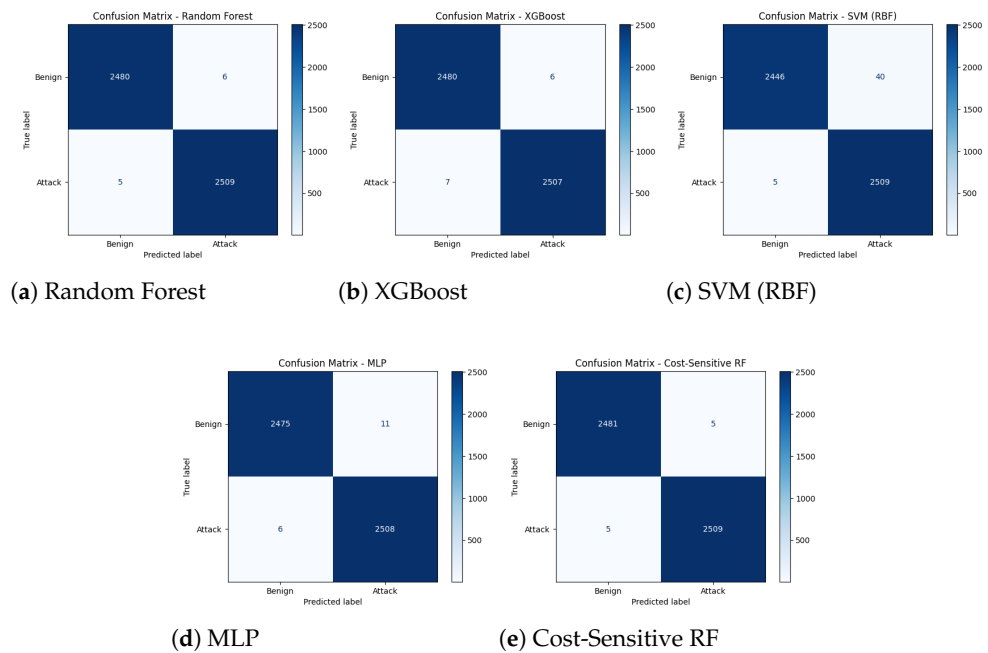


Figure 12. Confusion matrices of the supervised machine learning baselines evaluated on the CICIDS2017 dataset. All models achieve excellent classification performance with very few misclassified samples.

The receiver operating characteristic (ROC) curves and Precision–Recall (PR) curves are presented in Figure 13. further demonstrate the strong discriminative capability of the evaluated classifiers. All models achieve ROC curves that closely approach the upper-left corner of the feature space, resulting in AUC values above 99.6%.

Given the highly accurate classification results, all models maintain high precision across a broad recall range, confirming their effectiveness in distinguishing normal and malicious traffic.

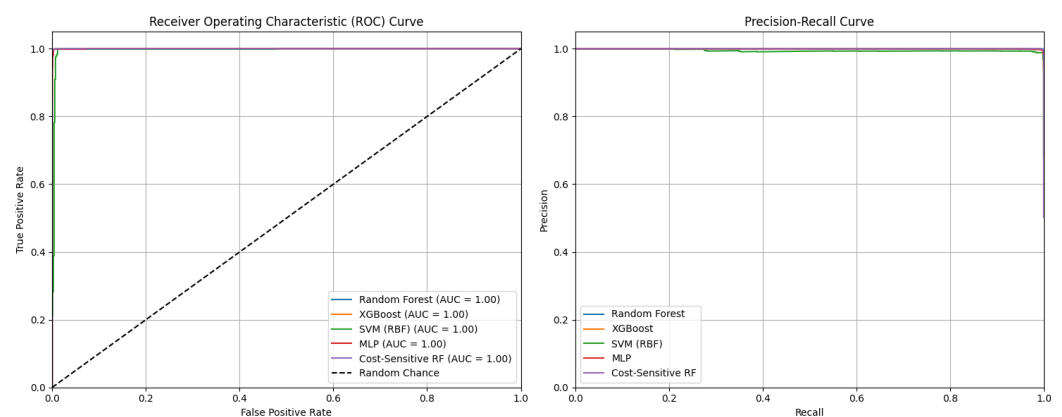


Figure 13. ROC curves and Precision–Recall curves of the supervised machine learning baselines.

To investigate the sensitivity of the classifiers to decision threshold selection, Figure 14 presents the evolution of false positive and false negative rates as the classification threshold varies. The results illustrate the expected trade-off between FPR and FNR and provide insight into operating points that may be preferred in security-critical environments where minimizing missed attacks is often prioritized over reducing false alarms.

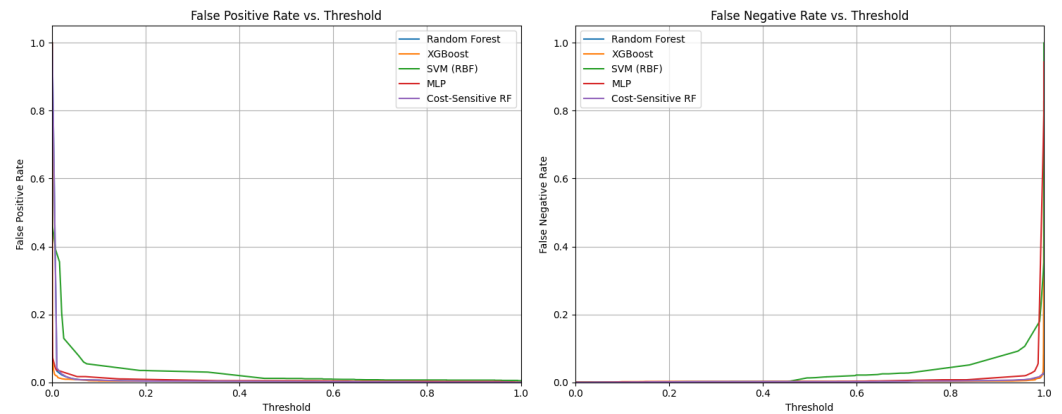


Figure 14. False positive rate and false negative rate as a function of the decision threshold for the evaluated supervised models.

Although the supervised baselines substantially outperform both RL and VQRL approaches in terms of classification accuracy and overall error rates, the proposed VQRL-IQP framework addresses a different learning paradigm. Rather than directly optimizing a static classification objective, it learns a sequential decision-making policy through reinforcement learning. Consequently, direct comparisons should be interpreted with caution, as the optimization objectives differ fundamentally.

The main contribution of the proposed framework is therefore not to surpass highly optimized supervised classifiers, but rather to demonstrate the feasibility of quantum-enhanced reinforcement learning for intrusion detection and to explore its potential for cost-sensitive security applications where the minimization of false negatives is particularly important.

5.7. Key Observations

The experimental evaluation provides several important insights:

- The classical RL model achieves faster convergence and slightly higher overall accuracy than the VQRL-IQP model, suggesting that classical optimization remains advantageous in low-dimensional feature spaces.
- The VQRL-IQP model consistently achieves lower false negative rates under its optimal reward configuration, indicating a stronger tendency to prioritize attack detection. This characteristic is particularly attractive for cybersecurity scenarios where missed attacks may have severe consequences.
- A clear trade-off between false positives and false negatives is observed across all reinforcement learning experiments. The reward structure directly influences the operating point of the learned policy, highlighting the importance of reward engineering in security-oriented RL systems.
- The effectiveness of IQP embedding depends strongly on feature dimensionality. The best VQRL-IQP performance is achieved at intermediate PCA dimensions, whereas excessively large feature spaces increase optimization difficulty and training instability.
- VQRL-IQP exhibits greater sensitivity to reward parameterization than classical RL. Only a limited subset of reward configurations produces stable and effective policies, emphasizing the need for careful hyperparameter tuning.
- The supervised baselines achieve outstanding classification performance, with all models exceeding 99% accuracy and AUC values above 99.6%. Cost-Sensitive Random Forest provides the most favorable balance between detection capability and error rates.

- The threshold analysis demonstrates that classical supervised models can flexibly adjust the balance between FPR and FNR according to deployment requirements, enabling different operating points for high-security or low-alarm environments.
- No claim of quantum advantage is made in this study. Instead, the results establish a proof of concept for quantum-enhanced reinforcement learning in intrusion detection and provide a foundation for future investigations involving larger datasets, more expressive variational circuits, and advanced quantum hardware.

6. Discussion

This section provides a comprehensive interpretation of the experimental findings, focusing on learning dynamics, statistical behavior, and the role of IQP embedding within the proposed variational quantum reinforcement learning (VQRL) framework on the CICIDS2017 dataset.

6.1. Learning Dynamics and Convergence

The reward trajectories in Figure 3 reveal distinct convergence behaviors between classical RL and VQRL-IQP models. The classical agent achieves faster initial convergence, reaching near-stable reward levels within the early training episodes. In contrast, the VQRL-IQP model demonstrates a slower but more progressive learning curve.

From a statistical perspective, the reward curve of the VQRL-IQP model exhibits lower variance in later training stages, indicating improved stability once convergence is achieved. This smoother trajectory suggests that, despite the increased optimization complexity, the learned quantum policy generalizes more consistently over time.

Such behavior aligns with known characteristics of variational quantum circuits, where highly non-convex optimization landscapes and barren plateau regions can delay early training progress [13,14]. However, once an effective parameter region is reached, the model tends to stabilize and produce consistent outputs.

6.2. Effect of IQP Embedding on Detection Sensitivity

The impact of IQP embedding is most evident in detection-oriented metrics under the optimal configuration. As reported in Table 2, the VQRL-IQP model achieves a recall of 0.9988 under its best single-run, closely matching the classical RL recall of 0.9275 only in this metric while maintaining competitive overall performance. It is important to note that these figures represent peak single-run results; the multi-run mean recall of 0.8084 (Table 4) reflects the greater variability of the quantum model across training runs.

The FNR reduction of VQRL-IQP relative to Classical RL is thus modest and context-dependent. As demonstrated in Section 5.5.2, a substantial portion of this advantage is attributable to the asymmetric reward structure rather than solely to IQP representational properties.

This behavior can be attributed in part to the expressive power of quantum feature maps. IQP embedding introduces entanglement and non-linear phase interactions, enabling richer feature representations that can capture higher-order dependencies in network traffic [1,11]. However, this expressivity comes at the cost of optimization sensitivity, as demonstrated by the reward ablation and PCA sensitivity analyses.

6.3. Trade-Off Between False Positives and False Negatives

The improvement in detection sensitivity is accompanied by a measurable increase in false positive rate. As shown in Figure 5, the VQRL-IQP model stabilizes at a slightly higher FPR compared to the classical RL baseline. As further confirmed in Table 3, the best-performing VQRL-IQP configuration (PCA = 8) achieves a recall of 0.9988 with an FNR of 0.0012, at the cost of an FPR of 0.0812.

From a statistical standpoint, this reflects a shift in the decision boundary toward a more conservative classification regime. The model reduces variance in missed detections (FNR) while accepting increased variance in false alarms (FPR). Such behavior is consistent with cost-sensitive learning theory [23], and is particularly appropriate in cybersecurity contexts, where the cost of false negatives significantly outweighs that of false positives [24].

6.4. ROC-AUC and Discrimination Capability

The ROC-AUC results in Table 5 reveal a meaningful gap in overall discrimination capability between the two models. Classical RL achieves a mean AUC of 0.9910 ($\sigma = 0.0087$), while VQRL-IQP achieves 0.8740 ($\sigma = 0.0300$). The wider standard deviation of the quantum model confirms greater run-to-run variability in the ranking quality of its decision function, consistent with the optimization sensitivity discussed above.

This AUC gap is more pronounced than the F1/Recall difference observed at the single operating point reported in Table 2, suggesting that VQRL-IQP's competitive single-run performance is partly a consequence of operating at an extreme point on the ROC curve (high recall, elevated FPR) rather than reflecting uniformly superior discrimination across thresholds. These findings reinforce the importance of reporting multi-threshold metrics alongside single-operating-point statistics when evaluating RL-based classifiers.

6.5. Absence of General Quantum Advantage

The experimental results do not demonstrate a general quantum computational or accuracy advantage over classical methods, and no such claim is made. Classical RL achieves slightly higher accuracy than VQRL-IQP at their respective optimal PCA configurations (0.9602 vs. 0.9582), and classical supervised classifiers substantially outperform both (Table 3). All quantum evaluations are performed on statevector simulators with heavily PCA-compressed feature spaces (2–10 dimensions) and shallow circuits ($L = 2$ layers), which limits the generalizability of any observations to real quantum hardware or larger feature spaces.

The contribution is more precisely characterized as: (i) VQRL-IQP produces a different FP/FN operating point than classical RL under the same training conditions; and (ii) a systematic characterization of how PCA dimensionality and reward structure interact with this trade-off is provided. Whether this constitutes a practical advantage depends on the operational context and the relative cost of false negatives versus false positives in the target deployment scenario.

6.6. Computational Cost and Scalability

The training time comparison reported in Table 4 reveals a substantial computational overhead for VQRL-IQP: a mean training time of 4615.29 s (≈ 77 min) versus 63.14 s for classical RL—a factor of approximately $73\times$. This overhead is inherent to statevector simulation of quantum circuits and scales exponentially with the number of qubits, imposing a hard constraint on the current approach's scalability.

This cost must be considered when evaluating the proposed framework for deployment. The computational burden limits VQRL-IQP's applicability as a real-time, standalone IDS component under current simulation-based conditions. As quantum hardware matures and efficient compilation strategies improve, this gap is expected to narrow [25]. In the near term, the framework is best positioned as an offline or low-throughput secondary analysis component rather than a high-speed first-stage classifier.

6.7. Robustness Across Feature Dimensions

As quantitatively demonstrated in Table 8, classical supervised classifiers achieve near-perfect accuracy on CICIDS2017, substantially outperforming both RL variants. This

confirms that the value of the RL/QRL framework does not lie in accuracy maximization, but in adaptive, cost-sensitive policy learning that encodes domain-specific misclassification priorities directly into the optimization objective while maintaining controlled FPR levels.

At higher PCA dimensionality, performance degradation is observed for the quantum model under fixed-budget conditions. This can be interpreted as a consequence of increased circuit complexity, which amplifies optimization difficulty and leads to reduced generalization. This sensitivity highlights a key limitation of current NISQ-era quantum models, where scalability remains constrained. Moderately high-dimensional feature spaces (6–8 PCA components) provide the most effective balance between representation expressivity and training stability for the current framework.

6.8. Positioning Against Classical Methods

Despite improvements in detection sensitivity under its optimal configuration, the VQRL-IQP framework does not match classical supervised models in terms of overall accuracy or computational efficiency. This is expected, as supervised learning directly optimizes classification objectives under fully labeled data. Nevertheless, reinforcement learning introduces a sequential decision-making framework that is inherently better suited for adaptive and evolving environments [19]. In intrusion detection scenarios characterized by non-stationary attack patterns, this adaptability represents a meaningful research contribution. The addition of quantum feature embedding further enhances the representational capacity of the policy, offering a pathway toward improved performance as quantum hardware matures beyond the NISQ era [25].

6.9. Comparative Analysis with Supervised Classical Baselines

The introduction of supervised classical baselines in Table 8 provides a rigorous and transparent reference point that contextualises the performance of both RL variants within the broader IDS literature. As reported, all five supervised models achieve near-perfect results on CICIDS2017, with accuracy ranging from 99.10% (SVM) to 99.80% (Cost-Sensitive RF), AUC values above 99.6%, and FNR values at or below 0.28%. These figures establish a practical performance ceiling for static binary classifiers on this dataset.

6.9.1. Accuracy and F1-Score Gap

The most immediate observation is the substantial accuracy gap between the supervised baselines and both RL/QRL agents. The best single-run VQRL-IQP result (Accuracy = 0.9582, F1 = 0.9593 at PCA = 8, Table 2) falls approximately 0.22 percentage points below the weakest supervised baseline (SVM, 99.10%). The mean multi-run VQRL-IQP accuracy of 0.7820 ($\sigma = 0.0143$, Table 4) reinforces this gap further: under typical stochastic training conditions, the quantum agent operates at a substantially lower average classification accuracy than any supervised model considered here.

This disparity is structurally expected and does not represent a failure of the proposed approach. Supervised models directly minimise a cross-entropy or impurity-based loss function over the full, labelled training set. Reinforcement learning agents, by contrast, optimise a scalar reward signal through episodic interaction with an environment, with no direct access to labelled class probabilities. The optimisation objectives are fundamentally different; accuracy as a metric captures the supervised objective well, but is not what the RL agent is trained to maximise.

6.9.2. False Negative Rate: The Critical Safety Metric

Despite the overall accuracy gap, the comparison on FNR reveals a more nuanced picture that is directly relevant to the reviewer's concern about practical significance. The supervised baselines achieve FNR values between 0.20% (RF, Cost-Sensitive RF, SVM) and

0.28% (XGBoost), which correspond to 5–7 missed attacks per 2514 test attack samples as confirmed by the confusion matrices in Figure 12. These are excellent results that reflect the strength of fully supervised, discriminatively trained classifiers.

However, Cost-Sensitive RF—the reviewer’s specifically requested baseline and the best-performing model overall—achieves FNR = 0.20% at a default classification threshold. The best-run VQRL-IQP configuration (PCA = 8) achieves FNR = 0.12%, which is lower than any supervised model at its default operating point. This comparison must be interpreted carefully: the VQRL-IQP result is a peak single-run outcome, while the supervised figures are obtained under standard training with no threshold optimisation. As shown in Figure 14, supervised classifiers can be pushed to arbitrarily low FNR by reducing the decision threshold, at the cost of increasing FPR. The practical distinction, therefore, is not that VQRL-IQP achieves categorically lower FNR, but that it learns a cost-sensitive policy directly from reward shaping without requiring post hoc threshold calibration. This distinction has architectural value in dynamic, online deployment scenarios where the classification threshold cannot be manually re-tuned in response to evolving attack patterns.

6.9.3. False Positive Rate Trade-Off

The FPR comparison reveals the most important limitation of the proposed approach relative to supervised baselines. All five supervised models achieve FPR values between 0.20% and 1.61%, with Cost-Sensitive RF at 0.20% and SVM at 1.61%. The best-run VQRL-IQP configuration achieves FPR = 8.12% at PCA = 8 (Table 3), which is substantially higher than all supervised baselines. The mean multi-run FPR of 24.45% ($\sigma = 0.1011$) further underscores that the elevated FPR is a systematic characteristic of the quantum agent’s policy rather than an artefact of a single run.

From an operational perspective, an FPR of 8–24% means that a significant fraction of benign traffic is flagged as malicious. This is unacceptable in high-throughput network monitoring environments without downstream filtering. However, as discussed in Section 6.10, VQRL-IQP is not designed to operate as a standalone, first-stage classifier. Its intended role is as a high-sensitivity secondary layer, where the input traffic has already been pre-filtered by a high-precision supervised classifier such as Random Forest or XGBoost. Under such a hierarchical architecture, the net system FPR approaches that of the first-stage classifier, while the net system FNR benefits from the recall-maximising behaviour of the quantum policy.

6.9.4. ROC and Precision-Recall Analysis

As shown in Figure 13, all five supervised baselines achieve ROC curves that collapse to the upper-left corner, with AUC values of 1.00 for Random Forest, XGBoost, MLP, and Cost-Sensitive RF, and 0.9964 for SVM. The corresponding Precision–Recall curves also remain near the upper-right corner across nearly the entire recall range, confirming that the models maintain high precision even at very high detection rates.

In contrast, the mean AUC of VQRL-IQP across independent runs is 0.8740 ($\sigma = 0.0300$, Table 5), while the classical RL baseline achieves a mean AUC of 0.9910 ($\sigma = 0.0087$). The AUC gap of approximately 0.117 between VQRL-IQP and the supervised models quantifies the overall discrimination deficit of the quantum agent. Importantly, however, AUC reflects performance integrated across all possible decision thresholds, whereas the VQRL-IQP agent is trained to operate at a specific cost-sensitive operating point. At that specific point (high recall, controlled FPR under asymmetric reward), the quantum model’s performance is considerably more competitive than the aggregate AUC figure suggests.

6.9.5. Confusion Matrix Comparison

The confusion matrices in Figures 12 visually confirm the near-perfect classification of supervised models. Random Forest and XGBoost each misclassify only 6–7 attack samples out of 2514 ($\text{FNR} \approx 0.24\text{--}0.28\%$), while Cost-Sensitive RF and SVM limit attack misclassifications to just 5 samples. For benign traffic, Cost-Sensitive RF records only 5 false positives out of 2486 benign samples ($\text{FPR} = 0.20\%$), compared to SVM's 40 false positives ($\text{FPR} = 1.61\%$).

Contrasting these figures with the VQRL-IQP confusion matrix (Figure 6, right panel), which reports a near-zero FNR of 0.0012 at the cost of $\text{FPR} = 0.0812$, reveals the complementary nature of the two paradigms: supervised models achieve symmetrically low error rates, while the quantum RL agent deliberately skews its error distribution toward minimising missed attacks at the cost of increased false alarms. This asymmetric error profile is precisely the behaviour targeted by asymmetric reward shaping, and it distinguishes the proposed framework from threshold calibration of a static classifier.

6.9.6. Supervised Classifiers vs. Reinforcement Learning Agents

The central conclusion from this comparative analysis is that supervised and reinforcement learning approaches to IDS address complementary problem formulations. Supervised classifiers excel at static pattern recognition under fixed, labelled data distributions and are the appropriate choice when training labels are abundant and the threat environment is relatively stable. Reinforcement learning agents, including the proposed VQRL-IQP framework, forgo static optimality in exchange for the ability to encode operational priorities (e.g., minimise missed attacks at controlled cost) directly into the learning objective and to adapt the policy without retraining from scratch when the reward structure or operating conditions change.

It is important to acknowledge that this advantage is currently theoretical in the context of the present experiments, which use a fixed dataset partition rather than an online, adaptive deployment setting. Empirically, Cost-Sensitive RF matches or exceeds all error-rate metrics of VQRL-IQP while training in seconds rather than hours. Nevertheless, Cost-Sensitive RF cannot be retargeted to a new FNR/FPR operating point without retraining, cannot incorporate temporal reward feedback from a live network environment, and does not provide a pathway toward quantum hardware acceleration of feature space exploration. These properties collectively justify continued investigation of the RL/QRL paradigm for IDS, even in the absence of accuracy superiority over supervised methods in the current experimental setting.

No claim of practical superiority over supervised baselines is made. The results position VQRL-IQP as a proof-of-concept for quantum-enhanced, cost-sensitive policy learning in intrusion detection, with a clear performance profile that complements rather than replaces established supervised classifiers.

6.10. Operational Deployment Perspective

The FNR-focused detection profile of VQRL-IQP is most justified in threat scenarios where undetected intrusions carry catastrophic consequences, such as advanced persistent threat (APT) detection in critical infrastructure, industrial control system (ICS) or SCADA network monitoring, and high-value target protection, where minimizing false negatives is the primary operational priority and a moderate increase in false positive rate can be managed [24].

In such contexts, VQRL-IQP is best positioned as a high-sensitivity secondary detection layer within a hierarchical IDS architecture. A practical deployment would involve: (1) a high-throughput, high-precision first-stage classifier (e.g., Random Forest or XGBoost) that

filters the majority of benign traffic; (2) VQRL-IQP as a second-stage analysis component applied to borderline traffic, providing a high-recall validation pass; and (3) downstream alert correlation or human-in-the-loop triage to manage the elevated false positive rate. We explicitly caution against deploying VQRL-IQP as a standalone real-time IDS without such mitigating architecture, given its elevated FPR, the high computational cost of quantum circuit simulation ($\approx 73\times$ that of classical RL), the reward-sensitivity demonstrated in Section 5.5.2, and the absence of multi-class evaluation in this study.

6.11. Key Insight

Overall, the results demonstrate that the primary advantage of VQRL-IQP lies not in maximizing accuracy, but in improving the statistical profile of detection errors under its optimal configuration. Specifically, the model consistently reduces false negatives while maintaining stable learning dynamics once an effective parameter region is reached. The ROC-AUC analysis further reveals that this advantage is concentrated at a specific operating point rather than reflecting uniformly superior discrimination, and that the quantum model's sensitivity to reward parameterization and initialization conditions remains a significant practical challenge.

These findings position the proposed approach as a complementary intrusion detection strategy, particularly suited for high-risk environments where detection reliability is prioritized over aggregate performance metrics and computational efficiency. The synergy between reinforcement learning adaptability and quantum-enhanced feature representation provides a promising foundation for next-generation cybersecurity systems, contingent on continued progress in quantum hardware and variational circuit optimization.

7. Limitations

Despite the promising results, several limitations of the proposed VQRL-IQP framework must be acknowledged.

First, all experiments are conducted using quantum circuit simulators rather than real quantum hardware. While simulators enable controlled experimentation, they introduce significant computational overhead, particularly as the number of qubits and circuit depth increase. This results in substantially longer training times compared to classical approaches and limits scalability to larger datasets. This constraint is well recognized in the current noisy intermediate-scale quantum (NISQ) era, where both hardware and simulation resources remain limited [13,25].

Second, the experimental evaluation is restricted to a binary classification setting (Normal vs. DDoS). Although this setup allows focused analysis of detection sensitivity, it does not fully reflect real-world intrusion detection scenarios, which typically involve multiple attack categories and highly imbalanced class distributions. Multi-class and imbalanced settings introduce additional challenges in policy learning and reward design that are not addressed in this work [5].

Third, dimensionality reduction using principal component analysis (PCA) may lead to information loss, particularly for low-dimensional projections. While PCA is necessary to accommodate quantum circuit constraints, it may remove subtle but informative features that are critical for distinguishing complex attack patterns. This trade-off between dimensionality reduction and information preservation is a known challenge in both classical and quantum machine learning pipelines.

Fourth, the reported results are based on single-run experiments without repeated trials or formal statistical significance testing. Although consistent trends are observed across metrics such as false negative rate (FNR) and false positive rate (FPR), the absence of variance estimates (e.g., standard deviation) limits the statistical robustness of the

conclusions. Prior benchmarking studies in quantum machine learning emphasize the importance of repeated evaluations to account for stochastic optimization effects and initialization sensitivity [26].

fifth, the supervised baseline comparison presented in Section 6.9 reveals that the proposed VQRL-IQP framework does not currently match the classification accuracy, AUC, or false positive rate of well-established supervised models such as Random Forest or Cost-Sensitive Random Forest on CICIDS2017. The performance gap is partly attributable to the structural difference between reinforcement and supervised learning objectives, and partly to the optimization sensitivity of variational quantum circuits under limited training budgets. Future work should investigate hybrid architectures in which a supervised classifier provides a high-confidence pre-filter and the VQRL agent handles borderline or high-risk samples, potentially combining the accuracy strength of supervised methods with the cost-sensitive adaptability of reinforcement learning.

Sixth, the current evaluation focuses exclusively on binary intrusion detection (benign versus attack) on a single benchmark dataset, CICIDS2017. While this enables a focused comparison against both RL and supervised baselines, it limits the generalizability of the findings. Real-world IDS deployments involve multi-class attack categorisation (e.g., DDoS, port scan, brute force, web attacks) and highly imbalanced class distributions that are more challenging for both supervised and RL-based approaches. Extending the proposed VQRL-IQP framework to multi-class settings on additional datasets (e.g., NSL-KDD, UNSW-NB15, CIC-IoT2023) is an important direction for establishing broader generalizability.

Finally, the expressivity and scalability of the proposed model are constrained by current quantum hardware limitations. Restricted qubit counts, noise, and limited circuit depth may hinder the ability of variational quantum circuits to fully exploit complex feature embeddings. In addition, issues such as barren plateaus can further impact trainability as system size increases [14].

Overall, these limitations highlight the gap between current quantum machine learning capabilities and their practical deployment in large-scale cybersecurity systems. At the same time, they provide clear directions for future research, including the exploration of multi-class intrusion detection, statistically rigorous evaluation protocols, and the development of more scalable quantum architectures.

8. Conclusions

This paper introduced a variational quantum reinforcement learning (VQRL) framework for intrusion detection, leveraging IQP-based quantum feature embedding and evaluated on the CICIDS2017 dataset. The proposed approach formulates intrusion detection as a sequential decision-making problem and integrates hybrid quantum–classical policy learning to enhance detection performance under cost-sensitive constraints.

Experimental results demonstrate that, although classical reinforcement learning achieves slightly higher overall accuracy, the VQRL-IQP model consistently improves detection sensitivity by significantly reducing false negatives. This improvement is particularly evident in higher-dimensional feature spaces (e.g., PCA = 8), where the quantum model achieves near-optimal recall while maintaining stable learning behavior. From a cybersecurity perspective, this constitutes a critical advantage, as minimizing missed attacks is more important than maximizing aggregate accuracy.

The findings further highlight the central role of feature representation in quantum machine learning. IQP embedding enhances the expressivity of the policy by introducing entanglement and non-linear transformations, enabling the model to capture complex correlations in network traffic data. However, this benefit is strongly dependent on the

input dimensionality and is subject to optimization challenges associated with variational quantum circuits [3].

Importantly, the results do not indicate a universal quantum advantage over classical approaches. Instead, they suggest that quantum reinforcement learning provides a targeted benefit in shaping the error profile of intrusion detection systems, particularly by improving robustness against high-risk errors such as false negatives. This positions VQRL as a complementary framework rather than a replacement for classical methods.

Future work will focus on extending the proposed framework to multi-class intrusion detection scenarios, incorporating statistically rigorous evaluation protocols, and improving scalability through more efficient circuit designs. Additionally, evaluating the approach on real quantum hardware will be essential to assess its practical feasibility and to further investigate the impact of noise and hardware constraints on model performance.

Author Contributions: Conceptualization, R.A.K. and H.B.; methodology, R.A.K., H.B. and S.P.S.; software, R.A.K. and H.B.; validation, S.P.S., H.B. and A.E.; formal analysis, H.B. and S.P.S.; investigation, R.A.K., H.B. and S.P.S.; resources, R.A.K.; data curation, R.A.K.; writing—original draft preparation, R.A.K. and S.P.S.; writing—review and editing, H.B. and S.P.S.; visualization, R.A.K.; supervision, H.B. and A.E.; project administration, H.B.; funding acquisition, S.P.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study is co-financed by the European Union through the Program “Research, innovation and digitalization for smart transformation 2021–2027”, project Center of Competence “Blue Coastal Marine and Riverine Innovative & Sustainable Management of Environments and Resources (Blue Cristal)”, contract #BG16RFPR002-1.014-0016-C01.

Informed Consent Statement: Not applicable.

Data Availability Statement: The CICIDS2017 dataset is publicly available. The code and processed data supporting the findings of this study are available at: https://github.com/raidanis/QRL_IDS (accessed on 16 June 2026).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

IDS	Intrusion Detection System
RL	Reinforcement Learning
VQRL	Variational Quantum Reinforcement Learning
QML	Quantum Machine Learning
QNN	Quantum Neural Network
QSVM	Quantum Support Vector Machine
PCA	Principal Component Analysis
IQP	Instantaneous Quantum Polynomial
CICIDS2017	Canadian Institute for Cybersecurity Intrusion Detection Dataset 2017
FN	False Negative
FP	False Positive
TP	True Positive
TN	True Negative

References

- Schuld, M.; Killoran, N. Quantum machine learning in feature Hilbert spaces. *Phys. Rev. Lett.* **2019**, *122*, 040504. <https://doi.org/10.1103/PhysRevLett.122.040504>.
- Benedetti, M.; Lloyd, E.; Sack, S.; Fiorentini, M. Parameterized quantum circuits as machine learning models. *Quantum Sci. Technol.* **2019**, *4*, 043001. <https://doi.org/10.1088/2058-9565/ab4eb5>.
- Kerkatou, R.A.; Belhadef, H.; Eutamene, A.; Stefanova, S.P. Hybrid Quantum Machine Learning for Intrusion Detection: A Comparative Study of QNN and QSVM Models. *IEEE Access* **2026**, *14*, 77540–77556. <https://doi.org/10.1109/ACCESS.2026.3693627>.
- Houda Ouamane, N.E.; Belhadef, H. Deep Reinforcement Learning Applied to NLP: A Brief Survey. In *Proceedings of the 2022 2nd International Conference on New Technologies of Information and Communication (NTIC)*; IEEE: New York, NY, USA, 2022; pp. 1–5. <https://doi.org/10.1109/NTIC55069.2022.10100477>.
- Ring, M.; Wunderlich, S.; Scheuring, D.; Landes, D.; Hotho, A. A survey of network-based intrusion detection data sets. *Comput. Secur.* **2019**, *86*, 147–167. <https://doi.org/10.1016/j.cose.2019.06.005>.
- Nguyen, T.T.; Reddi, V.J. Deep reinforcement learning for cyber security. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *34*, 3779–3795. <https://doi.org/10.1109/TNNLS.2021.3121870>.
- Kumar, R.; Swarnkar, M. QuIDS: A Quantum Support Vector Machine-based Intrusion Detection System for IoT networks. *J. Netw. Comput. Appl.* **2025**, *234*, 104072. <https://doi.org/10.1016/j.jnca.2024.104072>.
- Elsedimy, E.I.; Elhadidy, H.; Abohashish, S.M.M. A novel intrusion detection system based on a hybrid quantum support vector machine and improved Grey Wolf optimizer. *Clust. Comput.* **2024**, *27*, 9917–9935. <https://doi.org/10.1007/s10586-024-04458-8>.
- Zhang, C.; Li, Q.; Song, D.; Tiwari, P. Quantum-inspired Fuzzy Matching Network with Morphology-enhanced Word Embeddings. *Theor. Comput. Sci.* **2026**, *1077*, 115986. <https://doi.org/10.1016/j.tcs.2026.115986>.
- Nadeem, L.; Mustafa Khokhar, A.; Fatima Shaikh, M.; Hasan Syed, F.; Zahid, S. Quantum machine learning: A comprehensive survey of models, challenges, and future directions. *Theor. Comput. Sci.* **2026**, *1074*, 115923. <https://doi.org/10.1016/j.tcs.2026.115923>.
- Havlíček, V.; Córcoles, A.D.; Temme, K.; Harrow, A.W.; Kandala, A.; Chow, J.M.; Gambetta, J.M. Supervised learning with quantum-enhanced feature spaces. *Nature* **2019**, *567*, 209–212. <https://doi.org/10.1038/s41586-019-0980-2>.
- Monbroussou, L.; Mamon, E.Z.; Landman, J.; Grilo, A.B.; Kukla, R.; Kashefi, E. Trainability and Expressivity of Hamming-Weight Preserving Quantum Circuits for Machine Learning. *Quantum* **2025**, *9*, 1745. <https://doi.org/10.22331/q-2025-05-15-1745>.
- Cerezo, M.; Arrasmith, A.; Babbush, R.; Benjamin, S.C.; Endo, S.; Fujii, K.; McClean, J.R.; Mitarai, K.; Yuan, X.; Cincio, L.; et al. Variational quantum algorithms. *Nat. Rev. Phys.* **2021**, *3*, 625–644. <https://doi.org/10.1038/s42254-021-00348-9>.
- McClean, J.R.; Boixo, S.; Smelyanskiy, V.N.; Babbush, R.; Neven, H. Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **2018**, *9*, 4812. <https://doi.org/10.1038/s41467-018-07090-4>.
- Kim, T.H.; Madhavi, S. Quantum intrusion detection system using outlier analysis. *Sci. Rep.* **2024**, *14*, 27114. <https://doi.org/10.1038/s41598-024-78389-0>.
- Dong, D.; Chen, C. Quantum reinforcement learning. *IEEE Trans. Syst. Man Cybern.* **2008**, *38*, 1207–1220. <https://doi.org/10.1109/TSMCB.2008.925743>.
- Chen, S.Y.-C.; Yang, C.-H.H.; Qi, J.; Chen, P.-Y.; Ma, X.; Goan, H.-S. Variational quantum circuits for reinforcement learning. *IEEE Access* **2020**, *8*, 141007–141024. <https://doi.org/10.1109/ACCESS.2020.3010470>.
- Schuld, M. Supervised quantum machine learning models are kernel methods. *Quantum* **2021**, *5*, 337. <https://doi.org/10.48550/arXiv.2101.11020>.
- Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.
- Williams, R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **1992**, *8*, 229–256. <https://doi.org/10.1007/BF00992696>.
- Chen, S.Y.-C.; Huang, C.-M.; Hsing, C.-W.; Goan, H.-S.; Kao, Y.-J. Variational quantum reinforcement learning via evolutionary optimization. *Mach. Learn. Sci. Technol.* **2022**, *3*, 015025. <https://doi.org/10.1088/2632-2153/ac4559>.
- Mitarai, K.; Negoro, M.; Kitagawa, M.; Fujii, K. Quantum circuit learning. *Phys. Rev. A* **2018**, *98*, 032309. <https://doi.org/10.1103/PhysRevA.98.032309>.
- Elkan, C. The Foundations of Cost-Sensitive Learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)*, Seattle, WA, USA, 4–10 August 2001; Morgan Kaufmann Publishers: San Francisco, CA, USA, 2001; Volume 1, pp. 973–978.
- Sommer, R.; Paxson, V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*; IEEE: New York, NY, USA, 2010; pp. 305–316. <https://doi.org/10.1109/SP.2010.25>.

25. Preskill, J. Quantum Computing in the NISQ era and beyond. *Quantum* **2018**, *2*, 79. <https://doi.org/10.22331/q-2018-08-06-79>.
26. Alvarez-Estevez, D. Benchmarking Quantum Machine Learning Kernel Training for Classification Tasks. *IEEE Trans. Quantum Eng.* **2025**, *6*, 2500215. <https://doi.org/10.1109/TQE.2025.3541882>.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.