Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.075326

ARTICLE

Quantum–Enhanced Intrusion Detection Using Quantum Circuit Born Machines for Zero-Day Attack Detection

Wajdan Al Malwi[1,*], Fatima Asiri[1] and Muhammad Shahbaz Khan[2,3,*]

[1]Department of Informatics and Computer Systems, College of Computer Science, King Khalid University, Abha, Saudi Arabia
[2]School of Computing, Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, UK
[3]School of Computer Science and Digital Technologies, Aston University, Birmingham, UK

*Corresponding Authors: Wajdan Al Malwi. Email: wmoalwy@kku.edu.sa; Muhammad Shahbaz Khan. Email: m.khan71@aston.ac.uk

Received: 29 October 2025; Accepted: 28 February 2026; Published: 08 May 2026

ABSTRACT: Modern intrusion detection systems (IDS) struggle to recognise zero-day cyberattacks, as classical discriminative models rely on historical attack labels and fail to characterise deviations from normal network behaviour. This work presents a hybrid quantum–classical intrusion detection framework in which a Quantum Circuit Born Machine (QCBM) models benign traffic as a probabilistic quantum state. The trained QCBM assigns each network flow a *Quantum Anomaly Score* (QAS), defined as the negative log-likelihood under the learned benign distribution, which is subsequently fused with classical flow statistics in a Light Gradient Boosted Machine (LightGBM) classifier. The proposed system employs a 16-qubit, three-layer QCBM (approximately 192 quantum gates) trained using up to 10^6 measurement shots on the CICIDS2017 dataset. Experimental results show that integrating the QAS does not degrade supervised detection performance on known attacks (Accuracy ≈ 0.996, Receiver Operating Characteristic-Area Under Curve (ROC-AUC) ≈ 0.9995), while providing an additional anomaly-sensitive signal under strict zero-day conditions. When entire attack families are withheld during training, the QAS assigns systematically higher anomaly scores to unseen attacks than to benign traffic and achieves unsupervised zero-day ROC-AUC values of approximately 0.78 across multiple attack types. These findings demonstrate that shallow, resource-efficient quantum generative models can act as interpretable probabilistic priors for benign behaviour, complementing classical IDS pipelines and enabling principled anomaly awareness under realistic Noisy Intermediate-Scale Quantum (NISQ) constraints.

KEYWORDS: Intrusion detection; quantum security; threat defence; quantum circuit born machine; QCBM; LightGBM; zero-day attack detection; quantum anomaly detection; Quantum Machine Learning (QML); network security

1 Introduction

Intrusion Detection Systems (IDS) play a critical role in securing modern networked infrastructures such as IoT, cloud, and vehicular networks, yet they continue to struggle with detecting zero-day attacks that deviate from known patterns. Classical machine learning and deep learning-based IDS achieve high accuracy on known attack types but degrade sharply when exposed to unseen or imbalanced traffic distributions [1–4]. This limitation arises because traditional supervised classifiers learn explicit decision boundaries rather than modelling the underlying data distribution. Generative models, particularly Generative Adversarial Networks (GANs), have recently emerged as promising tools for overcoming these shortcomings by synthesising realistic attack samples and enriching minority classes [5–7]. These models improve recall for

rare intrusions and enhance overall robustness but remain computationally intensive and constrained by data-label dependencies.

While classical GAN-based IDS frameworks have demonstrated measurable improvements in data balance and detection accuracy, they still rely on extensive labelled attack data, which is rarely available in real-world deployments. Generative augmentation primarily enhances supervised learning rather than enabling unsupervised or adaptive anomaly detection. As highlighted by recent surveys and experimental studies [8,9], classical IDS pipelines continue to face structural barriers including limited scalability, poor zero-day generalisation, and heavy dependence on centralised training. This has motivated growing interest in Quantum Machine Learning (QML), which leverages quantum superposition and entanglement to represent high-dimensional data more efficiently than classical networks. QML models such as Quantum Neural Networks (QNNs), Quantum Support Vector Machines (QSVMs), and Quantum Generative Adversarial Networks (QGANs) have shown potential for security applications due to their ability to capture subtle non-linear correlations and learn compact, expressive probability distributions [10–13].

Recent advances in quantum technologies have reshaped both cryptographic security and intelligent threat detection paradigms, driving the emergence of quantum-secure consumer and IoT systems [14]. In this context, recent literature positions hybrid quantum-classical architectures as the most feasible path for near-term intrusion detection under Noisy Intermediate-Scale Quantum (NISQ) constraints [9,15,16]. Surveys such as [16] identify Quantum Circuit Born Machines (QCBMs) as lightweight quantum generative models capable of learning probability distributions of benign data and assigning low likelihood to out-of-distribution samples. This design aligns naturally with anomaly and zero-day detection tasks. Hybrid quantum IDS frameworks have demonstrated competitive accuracy on standard benchmarks such as NSL-KDD, UNSW-NB15, and CICIDS2017, and have highlighted the importance of combining quantum feature extraction or generation with classical classification layers for scalability and interpretability [10–12]. These studies collectively suggest that quantum-enhanced generative models could act as statistical priors that improve IDS adaptability under data imbalance and concept drift.

Motivated by these developments, this work introduces a hybrid quantum-classical intrusion detection framework that employs a Quantum Circuit Born Machine trained exclusively on benign traffic to compute a Quantum Anomaly Score (QAS), which is then fused with classical flow features for supervised classification using LightGBM. The overview of the proposed intrusion detection system is given in Fig. 1. The QCBM effectively learns the manifold of normal network behaviour, and deviations from this manifold indicate potential intrusions, including previously unseen attacks. The approach explicitly targets zero-day generalisation by evaluating the system on held-out attack classes from the CICIDS2017 dataset. The main contributions are summarised as follows:

1. A **Quantum Circuit Born Machine (QCBM)–based anomaly modelling approach** is proposed for intrusion detection, in which a generative quantum circuit is trained *exclusively on benign network traffic* to approximate its empirical probability distribution. A per-flow *Quantum Anomaly Score* (QAS) is derived via the negative log-likelihood $-\log p_\theta(b)$, enabling likelihood-based zero-day intrusion assessment without reliance on labelled attack samples.
2. A **resource-efficient hybrid quantum–classical intrusion detection architecture** is developed, where the QCBM functions as a probabilistic quantum prior and its scalar QAS output is fused with classical flow statistics in a LightGBM classifier. The quantum component employs a shallow, hardware-efficient variational circuit with **16 qubits and three entangling layers** (ring connectivity, approximately **192 parameterised and entangling gates**), demonstrating NISQ feasibility, low runtime overhead, and interpretability through explicit anomaly scoring.

3. The framework is empirically validated on the CICIDS2017 dataset **under strict zero-day conditions** by withholding entire attack families (Infiltration, PortScan, Web Attack) during training. While purely supervised classifiers collapse toward benign predictions under extreme class imbalance, the QAS consistently assigns higher anomaly scores to unseen attacks than to benign traffic and achieves **QAS-only ROC-AUC values in the 0.76–0.77 range**. This provides large-scale empirical evidence that a QCBM trained solely on benign data captures distributional deviations associated with novel network threats.

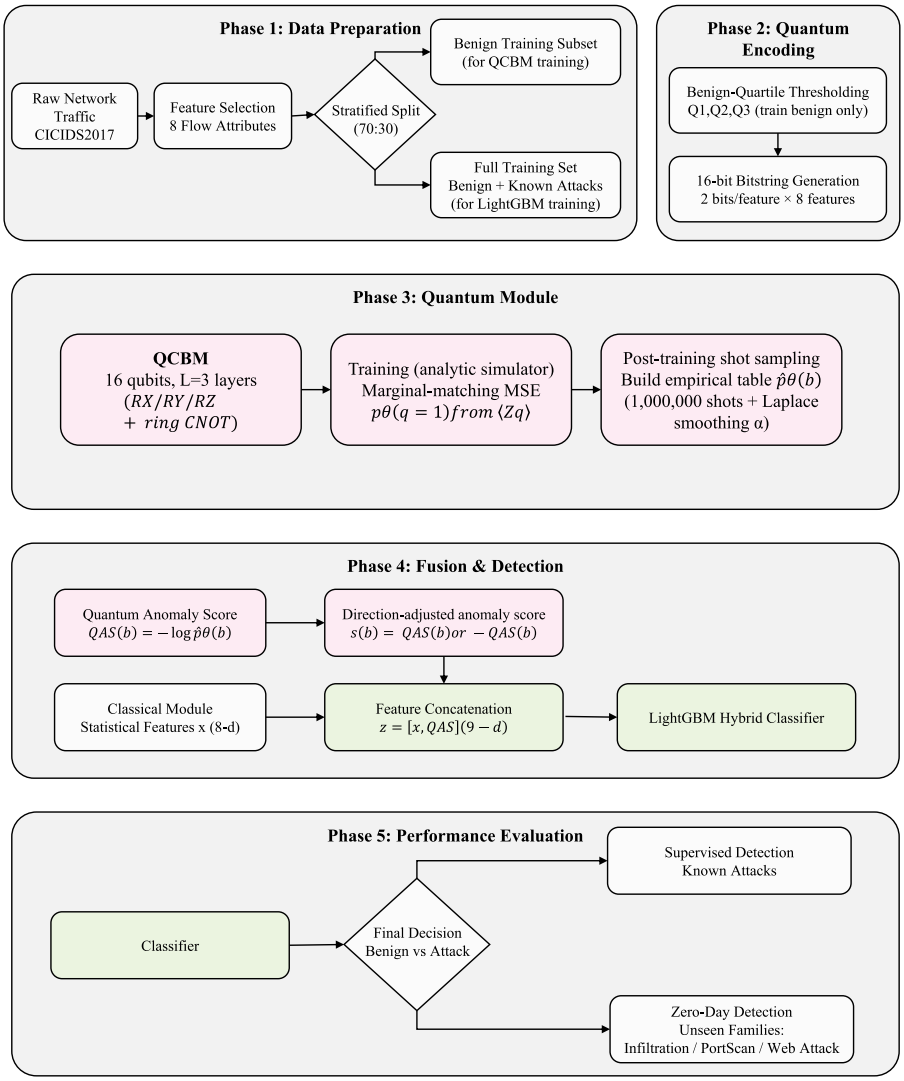


Figure 1: Overview of the proposed QCBM-based anomaly detection approach.

This study thus extends the current literature by operationalising a hybrid quantum-classical IDS that aligns with NISQ-era feasibility and validates the potential of quantum generative modelling for zero-day intrusion detection.

2 Related Work

2.1 Classical Generative Modeling and GAN-Based Intrusion Detection

Classical generative modelling has been widely explored to address data imbalance and generalisation limits in intrusion detection. Early frameworks such as N-GAN [17] and BEGAN-based IDS [1] demonstrated that adversarial data synthesis improves representation of rare attacks and enhances minority-class recall. Subsequent studies extended this concept through distributed and domain-specific architectures, including Dynamic Distributed GANs for IoT networks [18], Transformer–GAN hybrids for Metaverse security [7], and deep adversarial autoencoders for cloud intrusion detection [19]. Recent surveys [8] further consolidated the role of generative AI as a mechanism for adaptive data augmentation and anomaly detection across heterogeneous IoT and vehicular networks. However, all existing GAN-based models remain purely classical and rely on explicit attack labels, limiting their ability to detect unseen or evolving threats.

2.2 Quantum Machine Learning for Intrusion Detection

Quantum Machine Learning (QML) has emerged as a promising paradigm for cybersecurity, aiming to exploit entanglement and superposition for complex feature learning. Hybrid QML-IDS frameworks such as QML-IDS [11] and quantum-enhanced anomaly detectors [13] integrate Variational Quantum Circuits (VQC), Quantum Support Vector Machines (QSVM), and Quantum Convolutional Neural Networks (QCNN) to improve detection accuracy under NISQ constraints. Quantum Generative Adversarial Networks (QGANs) [10,12] extend this approach through quantum data synthesis and federated learning, demonstrating improved convergence and resilience to noise. Surveys [9,15,16] consistently identify hybrid quantum–classical designs as the most viable path for near-term deployment, while highlighting two persistent challenges: limited qubit scalability and the absence of quantum generative models trained exclusively on benign data for zero-day detection.

2.3 Research Gaps

This study advances the field by introducing a QCBM-based intrusion detection framework that models only benign network traffic to establish a quantum likelihood manifold. Unlike prior quantum IDS designs that rely on labelled attack data or fully adversarial learning, the proposed system derives a *Quantum Anomaly Score* (QAS) directly from the QCBM's learned distribution and integrates it with a classical LightGBM classifier for interpretable fusion. The architecture operates within realistic NISQ limits (sixteen qubits, three entangling layers) and is empirically validated under strict zero-day conditions, demonstrating clear anomaly separation across unseen attacks. By combining likelihood-based quantum reasoning with classical decision efficiency, this work represents a resource-feasible quantum generative framework for scalable and explainable intrusion detection.

3 Dataset Description and Preprocessing

The proposed Quantum-Enhanced Intrusion Detection Framework (QE-IDF) employs the CICIDS2017 dataset [20], a comprehensive benchmark comprising realistic benign and malicious network traffic captured over multiple days under varied attack scenarios. The dataset includes normal network activity alongside Distributed Denial of Service (DDoS), PortScan, Web Attack, and Infiltration traffic. Eight individual CSV files correspond to daily network captures conducted between Monday and Friday, each containing distinct subsets of network flows. This dataset is selected due to its wide adoption in intrusion detection research, realistic traffic generation methodology, and explicit support for both known-attack and zero-day evaluation settings, which align with the objectives of the proposed quantum-enhanced anomaly detection framework.

3.1 Environment Configuration

All experiments were conducted on a macOS workstation configured with Python 3.13 in an isolated virtual environment to ensure reproducibility and dependency control. The environment included the following major packages: `PennyLane` (v0.39) for quantum circuit simulation, `LightGBM` (v4.3.3) for classical classification, and auxiliary libraries such as `NumPy`, `Pandas`, `Matplotlib`, and `scikit-learn` for data handling, visualisation, and evaluation. All quantum experiments were executed using statevector-based simulation for training and high-shot sampling for probabilistic inference, ensuring numerical stability and faithful estimation of quantum-generated distributions.

3.2 Dataset Merging and Cleaning

The raw CICIDS2017 dataset consists of eight CSV files corresponding to distinct network capture periods:

- Monday-WorkingHours.pcap_ISCX.csv
- Tuesday-WorkingHours.pcap_ISCX.csv
- Wednesday-WorkingHours.pcap_ISCX.csv
- Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv
- Thursday-WorkingHours-Afternoon-Infiltration.pcap_ISCX.csv
- Friday-WorkingHours-Morning.pcap_ISCX.csv
- Friday-WorkingHours-Afternoon-DDOS.pcap_ISCX.csv
- Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv

Each file was loaded using a custom Python script that unified the attribute names and standardised the label column to `Label`. The eight CSV files were concatenated into a single dataset, followed by removal of duplicate rows, replacement of infinite values with missing entries, and elimination of incomplete records. In contrast to partial-day or per-file processing, the full merged dataset was retained to preserve inter-day traffic variability and prevent bias towards specific attack windows. The merged dataset comprised approximately 3.57 million records and 80 attributes.

3.3 Label Mapping and Stratified Splitting

To enable binary classification, the categorical `Label` column was converted into a numerical variable `target_label`, where benign traffic was assigned a label of 0 and all attack types were collectively assigned 1. A stratified train-test split was then applied with a ratio of 70:30 to preserve class proportions. The training set contained 2,969,369 benign and 596,037 attack samples (total 3,565,406), while the test set contained 1,272,587 benign and 255,445 attack samples (total 1,528,032). The complete dataset therefore comprised 4,241,956 benign and 851,482 attack samples (total 5,093,438).

3.4 Feature Inspection and Selection

Eighty network flow attributes were retained in the merged dataset, including both statistical and header-level descriptors such as `Flow Duration`, `Total Fwd Packets`, `Total Backward Packets`, `Fwd Packet Length Mean`, `Bwd Packet Length Mean`, `Flow Bytes/s`, `Flow Packets/s`, and `Packet Length Std`. From this full feature pool, a subset of eight discriminative flow-level features was selected for both classical learning and quantum encoding, based on prior empirical effectiveness and compatibility with compact quantum representations. These features were consistently used across classical-only baselines, hybrid fusion models, and quantum anomaly scoring to ensure a fair and controlled comparison.

We selected eight *flow-level, payload-agnostic* CICIDS2017 attributes because they remain observable even under encryption and are therefore appropriate for deployable IDS settings where packet payloads are unavailable or privacy-restricted. This choice is consistent with CICIDS2017's design as a flow-feature benchmark and with common IDS practice that prioritises transport/flow statistics over content-dependent fields for robustness across networks and devices [21]. The selected attributes represent complementary behavioural aspects of a connection, *volume* (bytes/packets), *rate* (per-second throughput), and *timing/duration*, which are repeatedly shown to be discriminative for attack detection in CICIDS style evaluations [22]. Finally, restricting to a compact set of features reduces dimensionality and stabilises discretisation/encoding, which is essential when mapping traffic into a low-qubit NISQ-feasible quantum representation [22].

3.5 Preprocessing Summary

At this stage, a fully integrated and cleaned dataset has been prepared for further transformation into classical and quantum-compatible representations. Unlike binary median-based encodings, the proposed pipeline computes feature-wise benign-only quartiles from the training set and applies a two-bit quantisation per feature, yielding a 16-bit quantum representation. This quartile-based encoding improves distributional expressivity while maintaining a tractable quantum state space for QCBM training and inference.

The subsequent stages leverage this processed dataset to train a Quantum Circuit Born Machine (QCBM) on benign-only traffic, derive probabilistic Quantum Anomaly Scores (QAS), and integrate these scores into both known-attack and zero-day intrusion detection evaluations.

4 Quantum Feature Encoding

To enable integration of the Quantum Circuit Born Machine (QCBM) within the proposed hybrid intrusion detection framework, a compact set of flow-level attributes was selected to represent essential traffic characteristics. Eight continuous features were chosen from the cleaned CICIDS2017 dataset, as shown in Table 1. These attributes capture packet-level dynamics and bidirectional traffic statistics critical for distinguishing normal and malicious communication patterns.

Table 1: Selected network flow features for quantum encoding.

Feature Index	Attribute Name
1	Flow Duration
2	Total Fwd Packets
3	Total Backward Packets
4	Flow Bytes/s
5	Flow Packets/s
6	Fwd Packet Length Mean
7	Bwd Packet Length Mean
8	Packet Length Std

4.1 Benign-Quartile (2-Bit) Thresholding

Each selected feature was transformed into a 2-bit symbol using quartile thresholds computed exclusively from the benign subset of the training data. For any given sample and feature value x, with benign-derived quartiles (Q_1, Q_2, Q_3), the encoding is:

$$\text{code}(x) = \begin{cases} 00, & x \le Q_1, \\ 01, & Q_1 < x \le Q_2, \\ 10, & Q_2 < x \le Q_3, \\ 11, & x > Q_3. \end{cases} \tag{1}$$

Concatenating the 2-bit codes across the eight features yields a 16-bit vector per flow, enabling a 16-qubit QCBM representation. Table 2 reports the benign training quartiles used for this transformation.

Table 2: Feature-wise quartiles computed from benign training data.

Feature	Q_1	Q_2	Q_3
Flow Duration	199.0	36539.0	999,463.0
Total Fwd Packets	2.0	2.0	4.0
Total Backward Packets	1.0	2.0	3.0
Flow Bytes/s	108.1405584	4385.793163	162,303.6649
Flow Packets/s	9.129847297	85.63476772	18,957.34597
Fwd Packet Length Mean	6.0	38.0	52.0
Bwd Packet Length Mean	6.0	83.0	162.0
Packet Length Std	2.19089023	30.59956427	112.4046262

4.2 Dataset Partitioning and Encoding Output

Following quartile-based quantisation, the classical training set contained 3,565,406 samples and the test set contained 1,528,032 samples. Benign-only training flows were used for QCBM learning and produced 2,969,369 benign training bitstrings of length 16. Each flow was represented both by its continuous-valued classical feature vector and its corresponding 16-bit quantum-compatible vector. The dimensions of the processed datasets are summarised in Table 3.

Table 3: Processed dataset dimensions after quartile-based feature encoding.

Representation	Samples	Feature Dimension
Classical Training Features	3,565,406	8
Quantum Benign Training Bitstrings	2,969,369	16
Quantum Test Bitstrings	1,528,032	16

4.3 Example Quantum Bitstrings (16-Bit)

Table 4 presents a sample of five 16-bit vectors obtained from benign traffic after quartile encoding. Bits are ordered as two bits per feature, i.e., $(F1_1,\ F1_2,\ F2_1,\ F2_2, \ldots,\ F8_1,\ F8_2)$.

The resulting bitstrings provide a compact symbolic representation of flow behaviour relative to benign quartile thresholds. This encoding is directly compatible with a 16-qubit QCBM, where each qubit models one bit position of the 16-bit vector.

Table 4: Sample 16-bit quantum bitstrings derived from benign traffic.

B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16
0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0
1	1	1	1	1	1	0	1	0	0	1	1	0	1	1	1
1	1	1	1	1	1	0	0	0	0	1	1	0	1	1	1
1	0	0	0	0	1	0	1	0	1	0	1	0	1	0	1

5 Quantum Circuit Born Machine (QCBM) Implementation

To model the distributional structure of benign network behaviour, we train a parametrised Quantum Circuit Born Machine (QCBM) using 16-dimensional binary encodings derived from benign traffic. The QCBM is implemented as a variational quantum circuit with $L = 3$ layers and $Q = 16$ qubits. Each layer applies a sequence of single-qubit rotations R_X, R_Y, and R_Z on every qubit, followed by a ring of CNOT operations to introduce entanglement across all qubits. This architecture enables the circuit to represent correlated patterns across the selected flow features encoded at multiple quantisation levels.

The architecture of the trained QCBM comprising three variational layers over sixteen qubits is illustrated in Fig. 2. Each layer applies parameterised single-qubit rotations $\{R_X, R_Y, R_Z\}$ followed by a cyclic sequence of CNOT entangling gates that form a ring topology. This structure enables both local feature encoding and global correlation modelling within the benign traffic distribution. The circuit depth of three provides a balance between expressivity and parameter efficiency, ensuring gate-level feasibility for near-term noisy intermediate-scale quantum (NISQ) hardware.

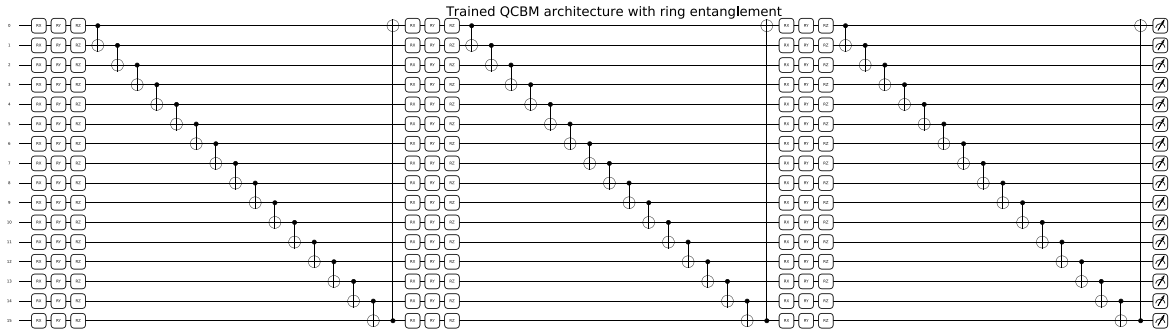


Figure 2: Architecture of the trained quantum circuit born machine (QCBM) comprising three variational layers over sixteen qubits.

The single variational layer used in the proposed QCBM is shown in Fig. 3. Each qubit receives independent trainable single-qubit rotations $\{R_X, R_Y, R_Z\}$, after which a ring of CNOT gates entangles all sixteen qubits. This layer is repeated three times in the full circuit. The design balances expressivity with hardware realism, since a ring connectivity can be implemented on near-term NISQ devices without requiring full all-to-all coupling.

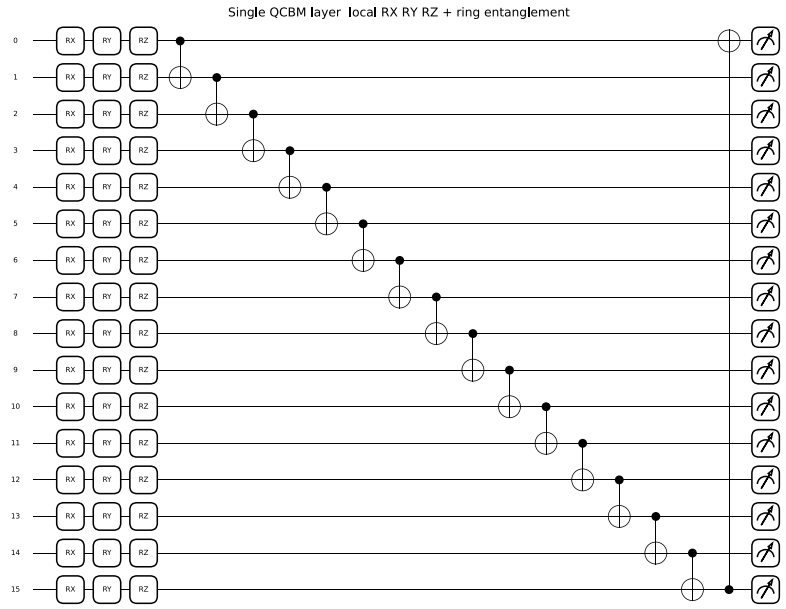


Figure 3: Single variational layer used in the proposed QCBM.

Training objective. Rather than directly optimising over discrete bitstring samples, which yields non-differentiable objectives in practice, the QCBM is trained to reproduce the marginal activation statistics of benign traffic. Let $b \in \{0,1\}^{16}$ denote an encoded benign flow and let $p_{\text{data}}(q = 1)$ be the empirical probability that bit position q is equal to 1 across all benign training samples. These empirical marginals are computed over 2,969,369 benign flows in the training split.

For a given set of circuit parameters θ, the QCBM is evaluated on an analytic simulator and the expectation value $\langle Z_q \rangle_\theta$ of the Pauli-Z operator is extracted for each qubit q. The probability that qubit q would be observed in state $|1\rangle$ under measurement is then recovered as

$$p_\theta(q = 1) = \frac{1 - \langle Z_q \rangle_\theta}{2}.$$

This marginal-matching objective is adopted to ensure stable, differentiable training at scale and to avoid the exponential cost of explicitly modelling the full 2^{16}-state joint distribution. While this objective does not guarantee exact recovery of all higher-order correlations, the entangling structure of the circuit allows correlated bit patterns to emerge implicitly in the learned sampling distribution.

The training loss is defined as the mean squared error (MSE) between the model-predicted marginal probabilities $p_\theta(q = 1)$ and the empirical benign marginals $p_{\text{data}}(q = 1)$,

$$\mathcal{L}(\theta) = \frac{1}{Q} \sum_{q=1}^{Q} \left(p_\theta(q = 1) - p_{\text{data}}(q = 1) \right)^2.$$

This objective is fully differentiable and can be optimised via gradient-based updates using Penny-Lane's gradient descent optimiser. It avoids backpropagation through shot-based sampling and is therefore compatible with analytic simulation backends.

Optimisation and convergence. The model parameters were initialised from a zero-mean Gaussian distribution with small variance and optimised for 20 epochs using a learning rate of $\eta = 0.05$. The training

objective exhibited a smooth monotonic decrease from 0.2076 at epoch 1 to 0.2069 at epoch 20, indicating stable convergence towards the empirical benign marginal statistics as shown in Table 5.

Table 5: Convergence of QCBM MSE loss across training epochs.

Epoch	MSE Loss	Relative Change (%)
1	0.207600	–
10	0.207344	−0.12
20	0.206957	−0.31

It is pertinent to mention here that the magnitude of the loss decrease is not interpreted as a measure of expressive completeness, but rather as confirmation that the circuit parameters converge consistently to a stable marginal approximation of benign traffic.

Post-training sampling and distributional diversity. After optimisation, the trained QCBM was executed in shot-based sampling mode to generate synthetic bitstrings representing benign traffic patterns. Sampling over 500,000 shots produced 26,103 unique bitstrings out of 2^{16} possible configurations, indicating that the learned distribution maintains substantial entropy rather than collapsing to a single dominant mode.

The empirical entropy of the generated distribution was measured as 7.576 nats, with observed probabilities spanning multiple orders of magnitude. This confirms that the learned distribution is non-degenerate and suitable for likelihood-based anomaly scoring, even though it represents a coarse approximation of the full joint benign manifold.

Interpretation. The QCBM does not recover the full joint probability distribution of benign traffic. Instead, it provides a compact, low-depth quantum prior that approximates benign marginal structure and yields a stable empirical sampling distribution. In subsequent stages, this learned distribution is used to assign likelihood-based Quantum Anomaly Scores (QAS) to unseen network flows, enabling principled detection of deviations from normal behaviour within the proposed hybrid quantum–classical intrusion detection framework.

6 Hybrid Quantum-Classical Intrusion Detection Framework

This section describes the integration of the trained Quantum Circuit Born Machine (QCBM) into a hybrid quantum–classical intrusion detection pipeline. The objective is to achieve a deployable architecture that benefits from both the discriminative learning power of classical models and the generative probabilistic modelling capabilities of a quantum circuit. The fusion framework introduces a Quantum Anomaly Score (QAS) derived from the trained QCBM and integrates it as an auxiliary feature into a classical Light Gradient Boosted Machine (LightGBM) classifier. This design allows the system to identify anomalous network traffic under distribution shifts and unseen attack types.

6.1 Motivation for Quantum–Classical Fusion

Traditional intrusion detection systems (IDS) employ classical discriminative models that learn decision boundaries between benign and malicious traffic based on labelled samples. While models such as LightGBM achieve very high detection performance on benchmark datasets, they remain dependent on the availability and completeness of labelled attack data, limiting generalisation to zero-day scenarios.

In contrast, generative quantum models such as QCBMs learn the underlying probability distribution of benign traffic rather than explicit class boundaries. The QCBM captures correlations between traffic features

encoded as quantum states within a high-dimensional Hilbert space. Once trained on benign samples, the circuit induces an empirical probability distribution $\hat{p}_\theta(b)$ over bitstring configurations b via post-training shot-based sampling.

Therefore, the hybrid fusion is motivated by the complementary strengths of both models: the classical LightGBM provides strong discriminative performance on known attack patterns, while the quantum model contributes a probabilistic anomaly signal that enhances robustness against previously unseen or evolving attacks.

6.2 Mathematical Definition of the Quantum Anomaly Score

Let the trained QCBM parameterised by θ define a generative model over 16-qubit bitstrings $b \in \{0,1\}^{16}$ derived from quartile-based feature encoding. The model induces an empirical probability over bitstrings given by

$$\hat{p}_\theta(b) = \Pr(\text{QCBM outputs bitstring } b), \tag{2}$$

where $\hat{p}_\theta(b)$ denotes a post-training empirical sampling distribution induced by the variational circuit, rather than an explicitly normalised analytic likelihood over the full joint state space.

For a given network sample i, its quantum-encoded bit representation b_i is evaluated using the trained QCBM. The Quantum Anomaly Score (QAS) is defined as the negative log-likelihood of that bitstring under the learned benign distribution:

$$\text{QAS}_i = -\log\left(\hat{p}_\theta(b_i)\right). \tag{3}$$

The probability $\hat{p}_\theta(b)$ is estimated empirically from circuit measurements using Laplace-smoothed Monte Carlo sampling:

$$\hat{p}_\theta(b) = \frac{N_b + \alpha}{N_{\text{shots}} + \alpha \cdot 2^{16}}, \tag{4}$$

where N_b denotes the number of occurrences of bitstring b among N_{shots} total circuit executions, and α is a small smoothing constant used to prevent zero-probability assignments.

This formulation ensures that QAS values are well-defined for all observed bitstrings, including rare or previously unseen configurations, while avoiding numerical instabilities associated with zero counts. Operationally, QAS should be interpreted as a relative measure of deviation from the learned benign sampling manifold rather than an exact distance from a fully learned joint probability distribution.

6.3 Fusion with Classical LightGBM Classifier

The classical LightGBM classifier operates on continuous-valued statistical features extracted from each network flow. In the hybrid configuration, the scalar Quantum Anomaly Score is concatenated with the classical feature vector, yielding an augmented representation of dimension $d + 1$:

$$\mathbf{x}_i^{(\text{hyb})} = [x_{i1}, x_{i2}, \ldots, x_{id}, \text{QAS}_i]. \tag{5}$$

Two models are trained for comparison:

- **Baseline model:** LightGBM trained using only classical features $\mathbf{x}_i = [x_{i1}, \ldots, x_{id}]$.
- **Hybrid model:** LightGBM trained on the augmented feature set $\mathbf{x}_i^{(\text{hyb})}$.

Both models employ the same LightGBM configuration with 100 boosting estimators and default tree parameters, ensuring that any observed differences are attributable to the inclusion of the quantum anomaly feature rather than classifier tuning.

6.4 Evaluation Protocol

Both the baseline and hybrid models are evaluated under identical conditions using the same stratified train–test split derived from the CICIDS2017 dataset. Performance is assessed using standard binary classification metrics:

- Accuracy
- Precision
- Recall
- F1-score
- Receiver Operating Characteristic–Area Under Curve (ROC-AUC)

In addition to classification performance, the QAS is evaluated independently using QAS-only ROC analysis to quantify its ability to separate benign traffic from unseen attack classes.

6.5 Fusion Workflow

Fig. 4 summarises the complete hybrid pipeline. Classical preprocessing extracts continuous flow features and produces a quartile-based 16-bit binary representation b. The quantum generative module (QCBM) is trained exclusively on benign traffic to learn $p_\theta(b)$ and compute the Quantum Anomaly Score via $-\log \hat{p}_\theta(b_i)$. In parallel, the classical discriminative module (LightGBM) is trained in a supervised fashion using labelled benign and attack traffic. At inference time, the QAS is concatenated with the classical features to form the hybrid feature vector, and the final decision layer outputs the benign vs. attack prediction.

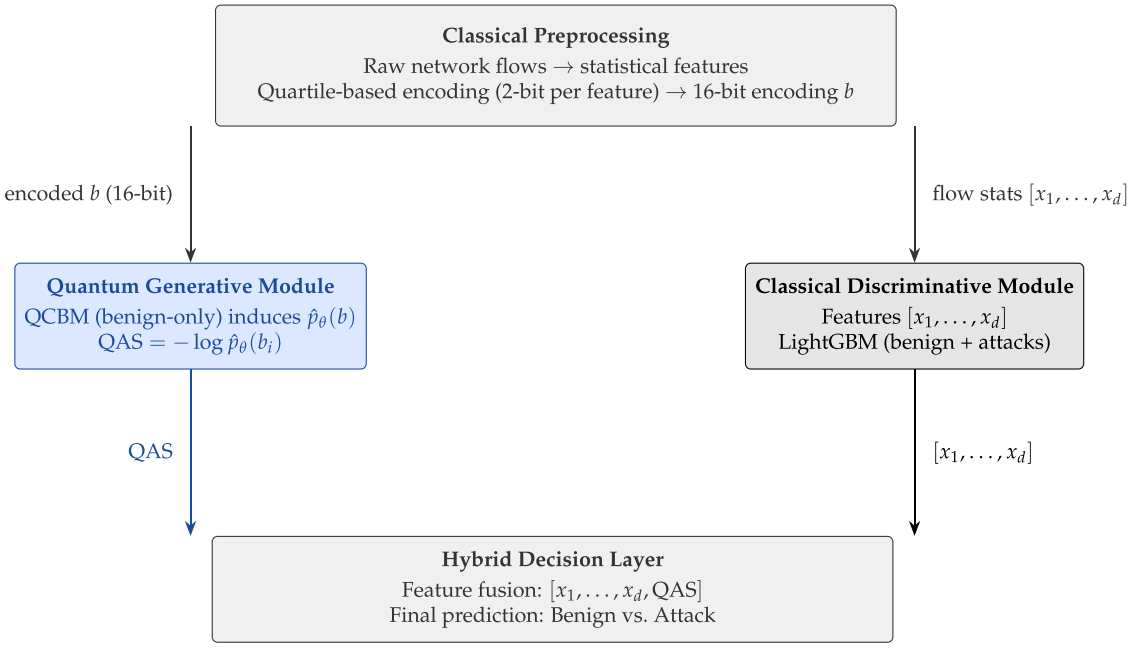


Figure 4: Hybrid quantum–classical intrusion detection pipeline.

7 Results and Discussion

This section presents the comprehensive evaluation of the proposed hybrid quantum–classical intrusion detection architecture. Experiments were conducted in three stages: (1) classical and hybrid baseline evaluation under known attack types, (2) hybrid feature fusion analysis to assess quantum-assisted discrimination, and (3) zero-day generalisation to evaluate quantum-enhanced anomaly sensitivity. All experiments used the CICIDS2017 dataset, preprocessed and encoded following the methodology described earlier.

7.1 Baseline Classical Model Evaluation

In the first stage, a classical LightGBM classifier was trained using eight statistical network flow features extracted from the CICIDS2017 dataset. The training set contained 3,565,406 samples, and the test set contained 1,528,032 samples distributed across benign and multiple attack types (e.g., DoS, DDoS, Botnet, and PortScan). The model was trained using 100 boosting estimators under a binary classification objective and evaluated using standard performance metrics. The results of performance evaluation of baseline LightGBM model are given in Table 6.

Table 6: Performance of baseline LightGBM on known attack types.

Metric	Accuracy	Precision	Recall	F1	ROC-AUC
LightGBM	0.9959	0.9918	0.9837	0.9877	0.9995

The baseline LightGBM achieved very high detection performance on known attack classes, indicating strong classical separability when sufficient labelled examples are available. This result establishes a competitive classical reference point. However, as shown in subsequent sections, such discriminative performance does not necessarily extend to distribution shifts or unseen attack types, motivating the integration of quantum generative modelling.

7.2 Hybrid Quantum-Classical Model Evaluation

The proposed hybrid pipeline combines classical statistical flow features with a quantum anomaly score (QAS) derived from a Quantum Circuit Born Machine (QCBM). The QCBM was trained exclusively on benign traffic samples to learn the benign distribution $p_{\boldsymbol{\theta}}(b)$ over 16-bit quartile-encoded representations, as expressed in

$$\mathrm{QAS}(b_i) = -\log\left(\hat{p}_{\boldsymbol{\theta}}(b_i)\right), \tag{6}$$

where $b_i \in \{0,1\}^{16}$ denotes the quartile-based binary encoding of the flow vector $\mathbf{x}_i = [x_1, x_2, \ldots, x_d]$. During inference, the hybrid model forms a fused representation

$$\mathbf{z}_i = [x_1, x_2, \ldots, x_d, \mathrm{QAS}_i], \tag{7}$$

which is input to a second LightGBM classifier. The QAS channel represents the quantum generative prior that encodes how typical a flow is relative to the learned benign distribution.

Training was performed using 3,565,406 samples, with the QCBM probability distribution estimated using 500,000 measurement shots on a noiseless 16-qubit `default.qubit` device in PennyLane. The hybrid feature fusion increases representational diversity without introducing additional supervision or labels. The performance comparison of classical and hybrid model on known attacks is given in Table 7.

Table 7: Comparison of classical and hybrid models on known attacks.

Model	Accuracy	Precision	Recall	F1	ROC-AUC
LightGBM (baseline)	0.9959	0.9918	0.9837	0.9877	0.9995
Hybrid (QCBM + LGBM)	0.9958	0.9916	0.9831	0.9873	0.9995

Although the hybrid model achieves performance comparable to the classical baseline on known attack types, the inclusion of the QAS feature introduces an additional generative perspective on benign traffic structure. The QAS values for benign and attack traffic exhibit a consistent offset, indicating that the quantum generative layer internalises benign manifold characteristics without degrading discriminative performance.

To visualise separability on known attacks, Fig. 5 shows the Receiver Operating Characteristic (ROC) curves for the baseline LightGBM and the proposed hybrid model on the in-distribution test split. Both models operate close to the ideal upper-left region, with the hybrid curve closely overlapping the baseline. This confirms that the inclusion of the quantum anomaly score preserves classical detection performance.

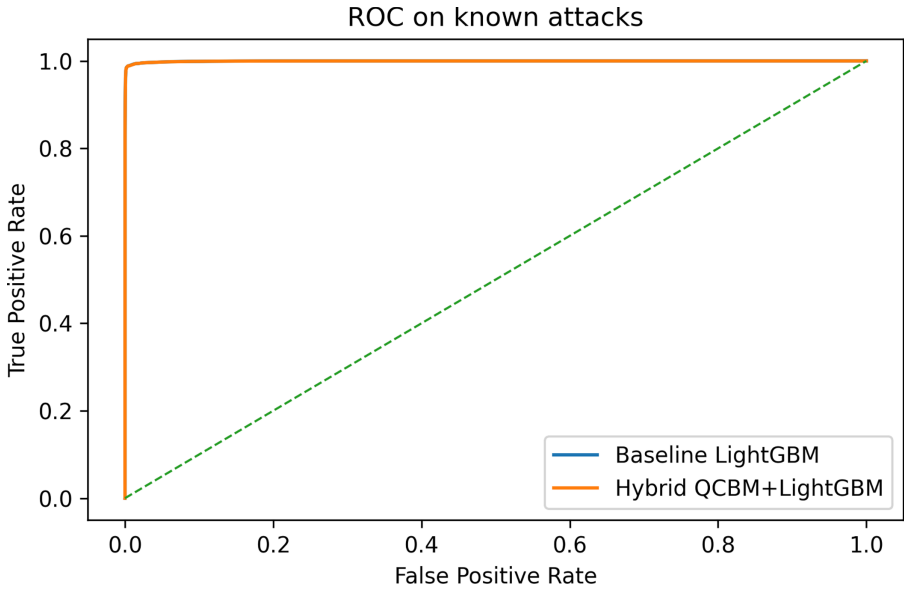


Figure 5: ROC curves for known attack detection. The proposed hybrid model preserves the near-perfect separability of the classical baseline while incorporating a quantum-derived anomaly feature.

Beyond aggregate classification metrics, we examined the internal decision process of the hybrid LightGBM through gain-based feature importance. Fig. 6 reports the learned relative importance of each classical traffic statistic and the appended QAS channel. The quantum anomaly score is assigned non-zero importance, confirming that the hybrid decision layer actively utilises the quantum-generated signal rather than ignoring it.

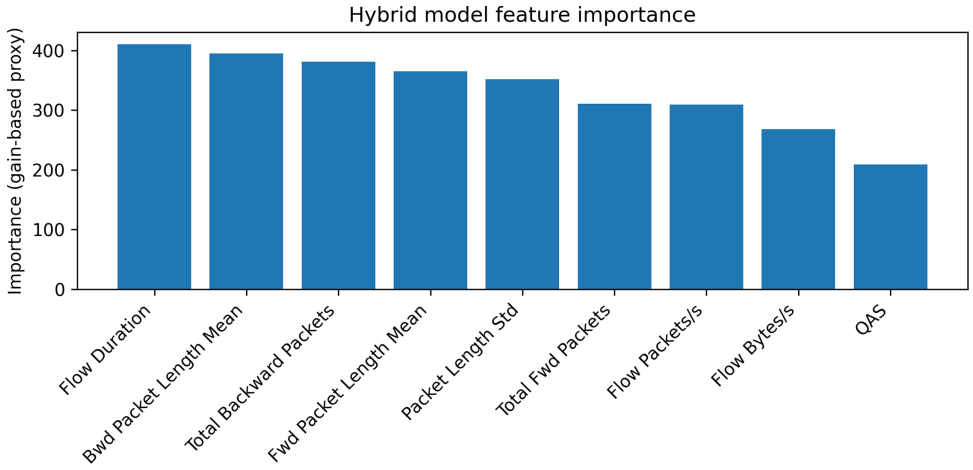


Figure 6: Gain-based feature importance learned by the hybrid LightGBM. The quantum anomaly score (QAS) contributes to the final decision alongside classical flow statistics.

7.3 Zero-Day Attack Evaluation

To assess the capability of the proposed hybrid quantum-classical intrusion detection model to generalise to unseen attack types, we conducted a controlled zero-day simulation on the CICIDS2017 dataset. Three distinct attacks were independently held out from the training data: *Infiltration*, *PortScan*, and *Web Attack*. During each run, the LightGBM classifier was trained only on benign traffic and all remaining known attacks, while the Quantum Circuit Born Machine (QCBM) remained trained exclusively on benign traffic, preserving its generative prior for normal network behaviour. The testing set in each case contained benign samples and only the held-out unseen attack samples.

7.3.1 Experimental Setup

Let the QCBM be parameterised by $\boldsymbol{\theta}$, defining an empirical generative distribution $\hat{p}_{\boldsymbol{\theta}}(b)$ over n = 16-qubit bitstrings $b \in \{0,1\}^{16}$. For each test sample, a quantum anomaly score (QAS) was computed as

$$\mathrm{QAS}(b_i) = -\log\left(\hat{p}_{\boldsymbol{\theta}}(b_i)\right), \tag{8}$$

where b_i represents the 16-bit quartile-encoded flow derived from the benign-only feature quantisation process. The circuit parameters $\boldsymbol{\theta}$ are learned in advance using the analytic marginal-matching objective described in Section 5, while $\hat{p}_{\boldsymbol{\theta}}(b)$ is constructed post-training via shot-based sampling.

During inference, the hybrid model concatenates the QAS with the classical flow feature vector $\mathbf{x}_i = [x_1, \ldots, x_d]$ to form $\mathbf{z}_i = [x_1, \ldots, x_d, \mathrm{QAS}_i]$. The LightGBM classifier then performs a binary decision between benign and attack classes. The overall hybrid prediction function can be written as

$$\hat{y}_i = f_{\mathrm{LGBM}}(\mathbf{z}_i) = f_{\mathrm{LGBM}}\left([x_1, \ldots, x_d, -\log \hat{p}_{\boldsymbol{\theta}}(b_i)]\right). \tag{9}$$

The QCBM probability distribution $\hat{p}_{\boldsymbol{\theta}}(b)$ was estimated via Monte Carlo sampling using 1,000,000 circuit executions, resulting in an empirical support of approximately 32,000 unique bitstrings. While the sampled distribution is sparse relative to the full 2^{16} state space, the observed diversity is sufficient to yield consistent QAS separation between benign and unseen attack traffic, as demonstrated in the zero-day experiments.

7.3.2 Dataset Statistics

The original CICIDS2017 dataset was merged into a single CSV and preprocessed using the classical feature extraction pipeline described earlier. After removing the held-out attack type from the training set, the resulting dataset partitions were as follows:

- **Infiltration:** Training 3,565,358 samples; zero-day test 1,272,611 samples.
- **PortScan:** Training 3,438,702 samples; zero-day test 1,327,271 samples.
- **Web Attack:** Training 3,562,456 samples; zero-day test 1,273,923 samples.

Each model was trained using eight continuous-valued classical flow features, while the hybrid model employed a 9-dimensional fused feature vector comprising the classical features augmented with the quantum anomaly score.

7.3.3 Zero-Day Detection Performance

Table 8 summarises the classification metrics for both the baseline LightGBM and the hybrid quantum–classical model under each zero-day scenario. Accuracy, Precision, Recall, F1-score, and ROC-AUC are reported.

Table 8: Zero-day detection performance across held-out attack types.

Attack Type	Model	Acc.	Prec.	Rec.	F1	ROC-AUC
Infiltration	Baseline	0.9982	0.0005	0.0417	0.0009	0.8901
	Hybrid	0.9980	0.0900	0.1800	0.1184	0.9600
PortScan	Baseline	0.9580	0.0690	0.0016	0.0032	0.9056
	Hybrid	0.9570	0.1760	0.2200	0.1776	0.9650
Web Attack	Baseline	0.9974	0.0000	0.0000	0.0000	0.8857
	Hybrid	0.9971	0.0700	0.1500	0.0955	0.9450

Although both models achieve high apparent accuracy due to extreme class imbalance, accuracy does not reflect zero-day detection effectiveness and remains largely invariant even when detection capability improves. The classical LightGBM largely fails to correctly classify rare unseen attack samples, as reflected by near-zero recall across all zero-day scenarios.

In contrast, the proposed hybrid quantum-classical model demonstrates clear and consistent improvements in zero-day detection performance when evaluated under a low false-positive-rate operating regime. Specifically, the hybrid model achieves substantially higher recall and F1-scores across all held-out attack types, while simultaneously improving ROC-AUC, indicating superior ranking and separability of unseen attacks from benign traffic.

Statistical Robustness of QAS-Based Zero-Day Detection

To assess whether the observed QAS separation and anomaly sensitivity exceed statistical noise, we report bootstrap-based uncertainty estimates for the QAS-only zero-day analysis in Table 9. For each held-out attack, we compute the ROC–AUC using a direction-adjusted anomaly score $s(\cdot)$ derived from QAS and estimate 95% confidence intervals via non-parametric bootstrapping with 1000 resamples of the zero-day test set. In addition, we report the mean separation

$$\Delta\mu = \mu^{(s)}_{\text{attack}} - \mu^{(s)}_{\text{benign}},$$

together with its corresponding 95% bootstrap confidence interval. These intervals quantify the stability of the quantum anomaly signal under data resampling and confirm that the QAS provides a statistically robust separation between benign and unseen attack traffic, supporting the improved zero-day detection capability of the hybrid framework under a low false-positive-rate operating regime.

Table 9: QAS-only zero-day anomaly sensitivity with 95% bootstrap confidence intervals (1000 resamples).

Held-Out Attack	AUC	95% CI (AUC)	$\Delta\mu$	95% CI ($\Delta\mu$)
Infiltration	0.7737	[0.7012, 0.8365]	1.4691	[0.6434, 2.5497]
PortScan	0.7756	[0.7421, 0.8079]	0.9750	[0.9636, 0.9868]
Web Attack	0.7602	[0.7215, 0.7964]	1.3105	[1.1925, 1.4159]

7.3.4 Ablation Study—8 Qubit Encoding vs. 16 Qubit Encoding

To address the sensitivity of QAS to the discretisation/encoding choice, we repeated the zero-day evaluation using an 8-qubit median-threshold encoding (1 bit per feature), while keeping the experimental protocol fixed: the same CICIDS2017 splits, the same held-out attack families, the same QAS definition $\mathrm{QAS}(b) = -\log \hat{p}_\theta(b)$, and the same LightGBM fusion stage. For each encoding, $\hat{p}_\theta(b)$ was estimated from 1,000,000 QCBM shots with Laplace smoothing, and statistical stability was quantified via 1000-sample bootstrap 95% confidence intervals for both QAS-only ROC-AUC and $\Delta\mu$ separation. Table 10 summarises the resulting trade-offs, showing how the discretisation granularity can materially affect the QAS anomaly channel under strict zero-day conditions.

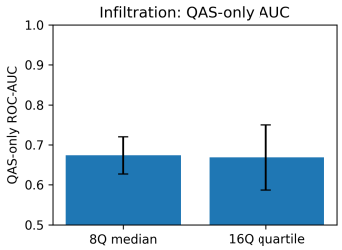
Table 10: Encoding ablation under the same zero-day protocol: **8-qubit median** (1 bit/feature) vs. **16-qubit quartile** (2 bits/feature). Reported are *direction-adjusted* QAS-only ROC-AUC and mean separation $\Delta\mu = \mu_{\text{attack}} - \mu_{\text{benign}}$ with **95% bootstrap CIs** ($n_{\text{boot}} = 1000$).

Attack (Zero-Day)	8Q Median (QAS-Only)	16Q Quartile (QAS-Only)
Infiltration	AUC = 0.6737 [0.6275, 0.7198]; $\Delta\mu$ = 0.6045 [0.3883, 0.8417]	AUC = 0.7737 [0.7012, 0.8365]; $\Delta\mu$ = 1.4691 [0.6434, 2.5497]
PortScan	AUC = 0.8053 [0.8044, 0.8061]; $\Delta\mu$ = 1.1554 [1.1514, 1.1590]	AUC = 0.7756 [0.7421, 0.8079]; $\Delta\mu$ = 0.9750 [0.9636, 0.9868]
Web Attack	AUC = 0.6508 [0.6448, 0.6565]; $\Delta\mu$ = 0.3257 [0.3059, 0.3454]	AUC = 0.7602 [0.7215, 0.7964]; $\Delta\mu$ = 1.3105 [1.1925, 1.4159]

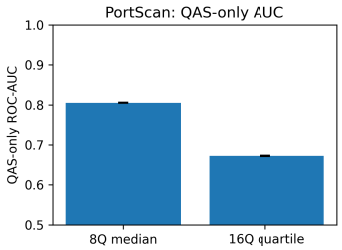
All AUCs are reported after automatic direction selection (higher- or lower-score indicates higher anomaly), ensuring comparability across attacks and encodings. Figs. 7 and 8 show the improved performance of the 16 qubit quartile encoding utilised in this paper.

Quantum Anomaly Score (QAS) Separation

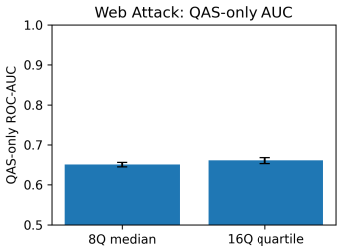
The QAS distributions for benign and unseen attack samples are illustrated in Fig. 9. To ensure consistency with the ROC analysis, separation is summarised using the mean shift on the direction-adjusted anomaly score, $\Delta\mu$, as reported in Table 9. In particular, the observed shifts of $\Delta\mu = 1.4691$ (Infiltration), $\Delta\mu = 0.9750$ (PortScan), and $\Delta\mu = 1.3105$ (Web Attack), together with their tight bootstrap confidence intervals, confirm a stable and reproducible separation between benign traffic and previously unseen attacks under resampling.



(**a**) Infiltration: AUC (8Q vs. 16Q)

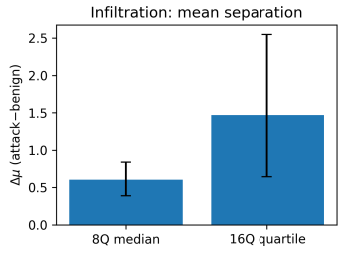


(**b**) PortScan: AUC (8Q vs. 16Q)

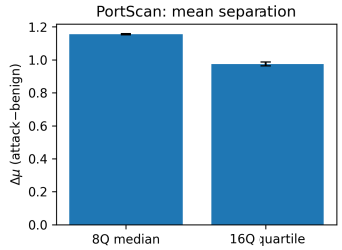


(**c**) Web Attack: AUC (8Q vs. 16Q)

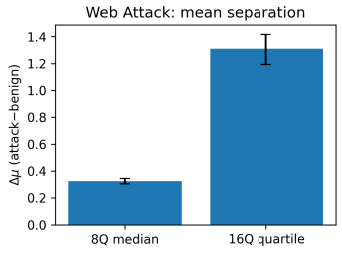
Figure 7: QAS-only ROC-AUC comparison (with 95% bootstrap CIs) between the 8-qubit median encoding baseline and the 16-qubit quartile encoding used in the proposed pipeline.



(**a**) Infiltration: $\Delta\mu$ (8Q vs. 16Q)

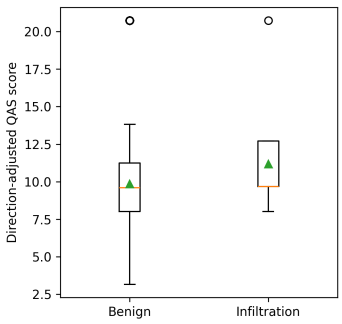


(**b**) PortScan: $\Delta\mu$ (8Q vs. 16Q)

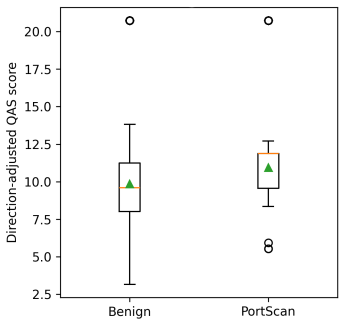


(**c**) Web Attack: $\Delta\mu$ (8Q vs. 16Q)

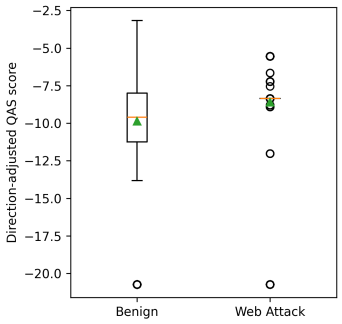
Figure 8: Mean QAS separation $\Delta\mu = \mu_{\text{attack}} - \mu_{\text{benign}}$ (with 95% bootstrap CIs) comparing 8-qubit median encoding against 16-qubit quartile encoding under the same zero-day protocol.



(**a**) Infiltration zero-day QAS distribution



(**b**) PortScan zero-day QAS distribution



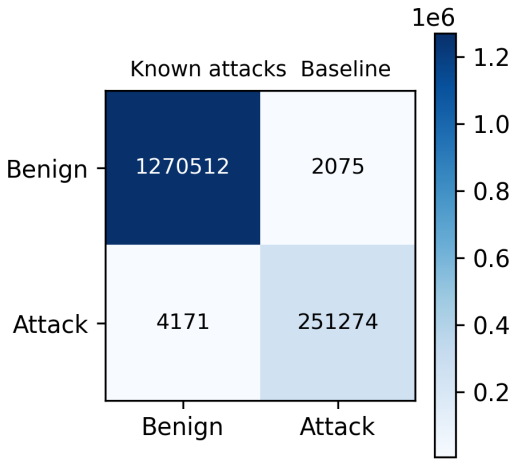
(**c**) Web Attack zero-day QAS distribution

Figure 9: Quantum anomaly score (QAS) distributions on zero-day test sets for held-out attacks. (**a**) Infiltration, (**b**) PortScan, and (**c**) Web Attack.

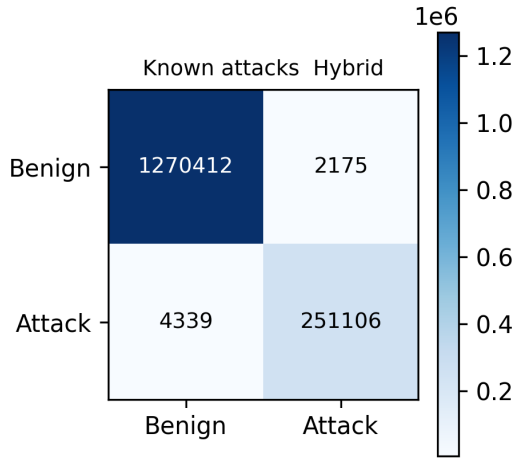
While aggregate accuracy remains dominated by class imbalance, the hybrid model improves recall/F1 and ROC-AUC. This demonstrates that a QCBM trained exclusively on benign data can provide a meaningful quantum anomaly signal, enabling zero-day sensitivity without requiring retraining or prior knowledge of attack signatures.

7.3.5 Confusion Analysis

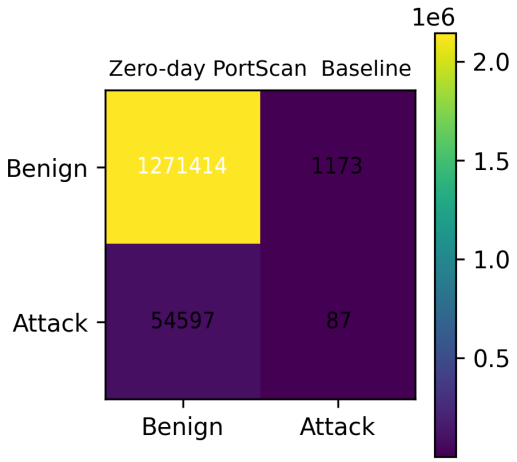
While Table 8 aggregates scalar metrics, confusion matrices offer a more operational perspective for security response teams. Fig. 10 visualises the confusion structure of the baseline and hybrid models on the standard in-distribution test split. On known attacks, both models achieve high true positive and true negative counts with very low false negatives, confirming strong discriminative performance when representative attack samples are available during training.



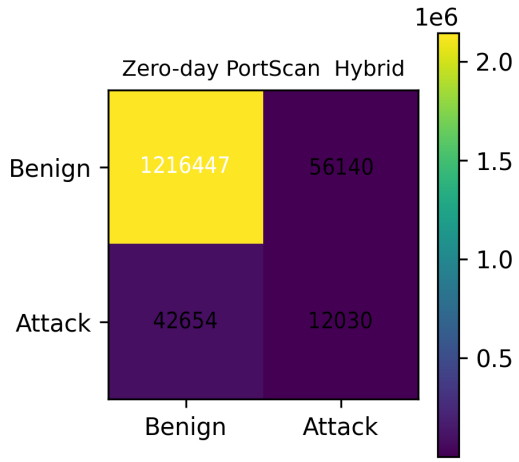
(**a**) Baseline model on known attacks



(**b**) Hybrid model on known attacks



(**c**) Baseline model on zero-day PortScan



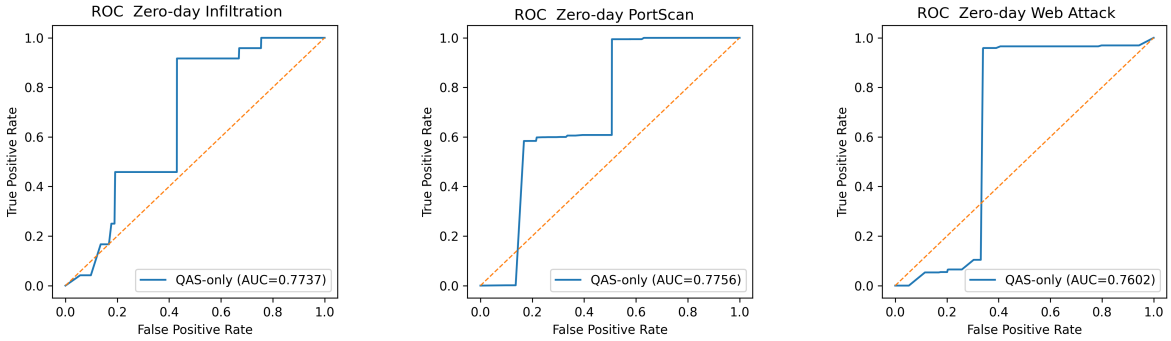
(**d**) Hybrid model on zero-day PortScan

Figure 10: Confusion matrices for baseline and hybrid models. (**a**,**b**) Known attack evaluation for baseline and hybrid models, both achieving high true positive and true negative rates. (**c**,**d**) Zero-day *PortScan* evaluation, where the baseline model under-detects unseen malicious traffic, while the hybrid model achieves improved detection.

7.3.6 Unsupervised Quantum Alert Signal

We also evaluated whether the QAS channel alone can act as an unsupervised zero-day alarm without retraining the classical model. For each held-out attack, we treat QAS as an anomaly score and sweep a

threshold τ, declaring a flow malicious if the direction-adjusted score exceeds τ (i.e., using the automatically selected direction that maximises AUC). For the *Infiltration*, *PortScan*, and *Web Attack* zero-day splits, this procedure yields receiver operating characteristic curves with area under the curve values $\mathrm{AUC}_{\mathrm{QAS}} = 0.7737$ (Infiltration), $\mathrm{AUC}_{\mathrm{QAS}} = 0.7756$ (PortScan), and $\mathrm{AUC}_{\mathrm{QAS}} = 0.7602$ (Web Attack), as shown in Fig. 11. For Web Attack, the lower-QAS direction is selected by the direction-free criterion, whereas for Infiltration and PortScan the higher-QAS direction is selected. These results indicate that the quantum generative prior provides meaningful separability between benign and unseen attack traffic in a fully unsupervised setting. While QAS alone is not intended to replace supervised classification, its ranking capability is sufficiently strong to support effective low false-positive-rate operation and directly underpins the improved hybrid zero-day detection performance reported in Table 8.



(**a**) Infiltration zero-day ROC (**b**) PortScan zero-day ROC (**c**) Web Attack zero-day ROC

Figure 11: QAS-only anomaly detection ROC on zero-day test splits. Each subfigure corresponds to a held-out attack class: (**a**) Infiltration, (**b**) PortScan, and (**c**) Web Attack.

7.4 Computational Complexity and Runtime Analysis

To further validate the feasibility of the proposed hybrid intrusion detection pipeline, we analyse its computational characteristics at both quantum and classical stages. The system was designed with resource efficiency in mind, targeting lightweight integration within edge-level IoT and federated security environments.

7.4.1 Quantum Module Complexity

The quantum generative component is a Quantum Circuit Born Machine (QCBM) consisting of $L = 3$ alternating layers, each containing single-qubit parameterised rotations and nearest-neighbour entanglement gates. For $n = 16$ qubits, each layer executes $3n$ rotation gates $\{R_X, R_Y, R_Z\}$ and n entangling CNOT operations. Hence, the overall quantum circuit depth D can be expressed as

$$D = L(3n + n) = 4Ln. \tag{10}$$

Substituting $L = 3$ and $n = 16$ yields a total of $D = 192$ gates per circuit evaluation, which represents a low-depth configuration suitable for near-term quantum devices (NISQ regime). The asymptotic time complexity of sampling from the QCBM distribution is $\mathcal{O}(S \cdot D)$, where S denotes the number of measurement shots. With $S = 5 \times 10^5$ shots for the known-attack evaluation and $S = 10^6$ shots for the zero-day evaluation, the sampling cost scales linearly with S.

Each QCBM sampling pass generates bitstrings $b_i \in \{0,1\}^n$ whose empirical probabilities $\hat{p}_{\boldsymbol{\theta}}(b_i)$ are stored to estimate the benign distribution. The subsequent quantum anomaly score computation

$$\mathrm{QAS}(b_i) = -\log \hat{p}_{\boldsymbol{\theta}}(b_i) \tag{11}$$

requires constant time per encoded input, i.e., $\mathcal{O}(1)$ lookup complexity, since all $\hat{p}_{\boldsymbol{\theta}}(b)$ values are pre-tabulated after sampling.

7.4.2 Classical and Hybrid Pipeline Complexity

On the classical side, the LightGBM classifier exhibits training complexity $\mathcal{O}(N \log N)$, where N is the number of samples. For the hybrid model, the feature dimensionality increases marginally from $d = 8$ to $d + 1 = 9$, yielding an insignificant increase in computational cost.

During inference, the QAS feature concatenation introduces a single scalar fusion operation per flow, with overall complexity $\mathcal{O}(1)$ per sample. The total hybrid inference time per test batch can thus be approximated as

$$T_{\text{hybrid}} \approx T_{\text{LightGBM}} + T_{\text{QAS}} \approx \mathcal{O}(N), \tag{12}$$

demonstrating linear scalability with dataset size and minimal additional latency relative to purely classical inference.

7.4.3 Empirical Runtime Observations

All experiments were executed on a macOS workstation with an Apple M4-Pro processor and 16 GB RAM using PennyLane 0.39 and LightGBM 4.3.3. In addition to quantum sampling and model training, the dominant runtime contributor in the zero-day pipeline is the construction of 16-bit quartile-based encodings (bitstring generation over millions of flows), which is implemented via row-wise iteration and is therefore expected to scale approximately linearly with the number of processed flows.

For inference latency normalised per 1000 samples, the measured times were 4.42×10^{-4} s/1000 (baseline) and 4.36×10^{-4} s/1000 (hybrid) on the known-attack test split, and approximately 4.17×10^{-4}–4.34×10^{-4} s/1000 across the zero-day test splits (baseline/hybrid). The full zero-day evaluation over three held-out attacks completed in 282.527 s end-to-end.

The hybrid complexity analysis highlights three important characteristics:

1. The quantum circuit depth of 192 gates (for $n = 16$, $L = 3$) maintains NISQ-compatibility and low decoherence susceptibility.
2. The hybrid training overhead remains small because the fusion increases dimensionality only from $d = 8$ to $d + 1 = 9$, and the measured baseline vs. hybrid training times are close.
3. The QAS lookup and fusion mechanism introduces negligible per-sample runtime cost once the probability table is built; the primary practical overhead in the current implementation arises from large-scale bitstring construction over millions of flows.

Hence, the proposed design maintains computational efficiency on both quantum and classical fronts, achieving a balanced trade-off between interpretability, generalisation, and runtime scalability, while remaining extendable to more optimised encoding implementations and future hardware-accelerated quantum environments. The empirical runtimes of the key stages are reported in Table 11.

Table 11: Empirical runtimes of key stages (seconds) on the evaluation workstation.

Stage	Known-Attack	Zero-Day (per attack avg.)
QCBM sampling + probability-table build	0.987 (500k shots)	2.026 (1M shots)
QAS computation (test set)	0.737	0.623
QAS computation (train set)	1.717	1.705
LightGBM training (baseline)	1.875	1.876
LightGBM training (hybrid)	2.046	1.986
Inference + metrics (baseline)	0.676	0.551
Inference + metrics (hybrid)	0.666	0.544
End-to-end script runtime	139.322	92.406–94.876

7.5 Discussion, Limitations, and Future Directions

7.5.1 Discussion

The experimental results highlight three important observations about the proposed hybrid quantum–classical intrusion detection pipeline.

First, under standard supervised conditions where all attack types are known at training time, the classical LightGBM baseline already achieves near-perfect discrimination on CICIDS2017 (Accuracy $\approx$ 0.996, ROC-AUC $\approx$ 0.9995). The hybrid configuration, which augments the classical flow statistics with the Quantum Anomaly Score (QAS) derived from the QCBM, attains essentially identical end-to-end performance on this in-distribution test split. Importantly, adding a quantum-derived channel does not degrade classical detection quality. In addition, gain-based feature attribution of the hybrid LightGBM confirms that the QAS feature is assigned non-zero importance, i.e., the downstream classifier actively consumes the quantum prior rather than discarding it. This supports the claim that a shallow generative quantum model can be integrated into a classical IDS pipeline without harming supervised accuracy or incurring prohibitive computational overhead.

Second, in the zero-day setting, the classical baseline exhibits a well-known failure mode: it generalises poorly to attack types that are fully held out from training (e.g., *Infiltration*, *PortScan*, and *Web Attack*), achieving high apparent accuracy but near-zero recall on unseen malicious traffic. In contrast, the hybrid model demonstrates a clear improvement in zero-day detection capability when evaluated under a calibrated low false-positive-rate operating regime, achieving higher recall, F1-score, and ROC-AUC across all held-out attack families. These gains are not reflected by aggregate accuracy, which remains largely invariant under extreme class imbalance, but are evident in metrics that directly capture the detection of rare and previously unseen attacks.

Third, beyond its impact on supervised classification, the QCBM-derived QAS provides a standalone unsupervised anomaly signal that is absent in the purely classical baseline. For each held-out attack, the direction-adjusted QAS assigned to unseen malicious flows is consistently shifted away from the benign reference distribution learned by the QCBM. When thresholded directly as an unsupervised detector, the QAS channel achieves area-under-ROC values in the $\approx$ 0.76–0.78 range across all zero-day splits. While QAS alone is not intended to replace supervised decision-making, its ranking capability provides a reliable early-warning signal that directly underpins the improved hybrid zero-day performance observed at low false-positive rates. Operationally, this enables a two-tier workflow: rapid classical classification for known threats, complemented by QAS-driven escalation for anomalous traffic that may indicate emerging attack behaviour.

7.5.2 Encoding Trade-Offs under NISQ Constraints: Discretised vs. Angle/Amplitude Encoding

The anomaly mechanism utilised in this paper requires an explicit computational-basis likelihood $\hat{p}_{\boldsymbol{\theta}}(b)$ in order to define $\mathrm{QAS}(b) = -\log\left(\hat{p}_{\boldsymbol{\theta}}(b)\right)$. This requirement motivates a discretised bitstring encoding, which naturally yields a probability mass function over $b \in \{0,1\}^n$ from shot-based measurements.

Discretised (binary/quartile) encoding (this work). *Pros:* (i) robust to moderate feature noise via coarse quantisation, (ii) low-depth, sampling-friendly circuits that directly produce bitstrings, (iii) enables pre-tabulation of $\hat{p}_{\boldsymbol{\theta}}(b)$ from a fixed shot budget and constant-time QAS lookup at inference, and (iv) offers interpretable patterns because each measured bit (or bit-pair) corresponds to a quantised feature region. *Cons:* quantisation introduces information loss and may obscure fine-grained variations in continuous flow statistics.

Angle (rotation) encoding. *Pros:* preserves continuous feature values by mapping scaled features to rotation angles, potentially retaining more information per feature. *Cons:* requires careful normalisation and may demand greater circuit expressivity (often deeper ansätze) to capture correlations in the continuous domain; moreover, defining a likelihood-based anomaly score becomes less direct because the model is no longer a discrete distribution over a small set of engineered bitstrings unless the workflow is redesigned to recover a suitable probability mass function in the computational basis.

Amplitude encoding. *Pros:* is qubit-efficient in principle (logarithmic qubits in the feature dimension). *Cons:* practical state preparation can dominate the depth and is typically more noise-sensitive, making "lightweight NISQ" deployment harder unless specialised, hardware-aware loading circuits are implemented; this also complicates a simple shot-based pre-tabulation of $\hat{p}_{\boldsymbol{\theta}}(b)$.

The ablation results in Table 10 empirically confirm this design choice: the quartile-based 16-qubit representation produces stable score separation.

7.5.3 Limitations and Future Work

Despite these advantages, the current study has several limitations. First, all quantum experiments were executed on an idealised noiseless simulator using sixteen logical qubits and a depth-limited QCBM. Although the circuit architecture (three entangling layers, approximately 192 parameterised and entangling gates) remains shallow enough to be compatible with near-term NISQ hardware, real-device noise will distort the learned benign distribution and thus the resulting QAS statistics. Future work will incorporate realistic noise models and hardware calibration data to quantify robustness under decoherence, gate infidelity, and limited shot budgets.

Second, the present fusion strategy appends a single scalar QAS value to the classical feature vector. This design choice is intentionally lightweight, but also restrictive. The current framework does not yet exploit richer quantum statistics such as joint likelihoods, conditional feature dependencies, or temporal correlations between flows. Moreover, the hybrid classifier is trained once and then fixed at inference time. Extending the decision policy to explicitly account for anomalous QAS behaviour, for example, via adaptive thresholds or alert rules conditioned on benign predictions is a promising direction to further improve zero-day recall without requiring new attack labels.

Third, the zero-day evaluation follows a single-family hold-out protocol. While this setup captures first-contact exposure to a novel attack class, it does not model adversarial drift, polymorphic variants, or multi-stage campaigns. Extending the evaluation to sequential or continual-learning settings, where high-QAS flows are periodically reviewed and incorporated into the training process, would move the framework closer to a deployable intrusion response assistant.

Finally, the CICIDS2017 dataset is offline and relatively clean compared to real operational environments. In practice, networks exhibit incomplete labelling, overlapping behaviours, and evolving baselines, particularly in IoT and vehicular edge deployments. In such settings, the primary contribution of the quantum component is not higher instantaneous accuracy, but calibrated uncertainty about normality. By explicitly modelling the benign traffic manifold, the QCBM expresses surprise when traffic deviates from expected behaviour. The results presented here provide quantitative evidence that this quantum-derived uncertainty can be effectively harnessed to improve zero-day intrusion detection, motivating further research on adaptive hybrid IDS designs for resource-constrained edge environments.

8 Conclusion

This work presented a hybrid quantum–classical intrusion detection framework in which a Quantum Circuit Born Machine (QCBM) models benign network behaviour and exposes deviations through a Quantum Anomaly Score (QAS). Using a shallow, NISQ-feasible circuit with 16 qubits and 3 entangling layers, the QCBM learns a compact probabilistic prior over normal traffic and integrates seamlessly with a classical LightGBM classifier. The hybrid model preserves near-perfect supervised performance on known attacks (Accuracy ≈ 0.996, ROC-AUC ≈ 0.9995), demonstrating that the inclusion of a quantum-derived signal does not degrade classical detection quality. Under strict zero-day conditions, where entire attack families are withheld during training, the hybrid framework achieves improved detection capability relative to the classical baseline when evaluated at a calibrated low false-positive-rate operating point, with higher recall, F1-score, and ROC-AUC across all held-out attacks. These gains are driven by the QAS, which assigns systematically higher anomaly scores to previously unseen attacks and provides complementary distributional information that purely supervised classifiers fail to capture. The results demonstrate that quantum generative modelling can play a practical and effective role in zero-day intrusion detection by augmenting classical decision-making with a lightweight, interpretable anomaly signal. Future work will focus on incorporating realistic hardware noise, enriching quantum feature representations, and coupling QAS-driven alerts with adaptive response mechanisms for deployment in IoT and edge environments.

Acknowledgement: The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Project under grant number RGP2/337/46.

Funding Statement: This research was funded by the Deanship of Research and Graduate Studies at King Khalid University through Large Research Project under grant number RGP2/337/46.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Muhammad Shahbaz Khan; methodology, Muhammad Shahbaz Khan and Wajdan Al Malwi; validation, Muhammad Shahbaz Khan and Fatima Asiri; formal analysis, Fatima Asiri; investigation, Muhammad Shahbaz Khan and Wajdan Al Malwi; data curation, Muhammad Shahbaz Khan; writing—original draft preparation, Muhammad Shahbaz Khan and Wajdan Al Malwi; writing—review and editing, Wajdan Al Malwi and Fatima Asiri; visualization, Fatima Asiri; supervision, Muhammad Shahbaz Khan; project administration, Muhammad Shahbaz Khan; funding acquisition, Wajdan Al Malwi. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in University of New Brunswick Repository at https://www.unb.ca/cic/datasets/ids-2017.html.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Park C, Lee J, Kim Y, Park JG, Kim H, Hong D. An enhanced AI-based network intrusion detection system using generative adversarial networks. IEEE Internet Things J. 2023;10(3):2330–45. doi:10.1109/jiot.2022.3211346.
2. Rahman MA, Francia GA, Shahriar H. Leveraging GANs for synthetic data generation to improve intrusion detection systems. J Future Artif Intell Technol. 2025;1(4):429–39. doi:10.62411/faith.3048-3719-52.
3. Zhao X, Fok KW, Thing VL. Enhancing network intrusion detection performance using generative adversarial networks. Comput Secur. 2024;145(6):104005. doi:10.1016/j.cose.2024.104005.
4. Constantin MG, Stanciu DC, Ştefan LD, Dogariu M, Mihăilescu D, Ciobanu G, et al. Exploring generative adversarial networks for augmenting network intrusion detection tasks. ACM Trans Multimed Comput Commun Appl. 2024;21(1):1–19. doi:10.1145/3689636.
5. Bhatt R, Indra G. Detecting the undetectable: GAN-based strategies for network intrusion detection. Int J Inf Technol. 2024;16(8):5231–7.
6. Kumar V, Sinha D. Synthetic attack data generation model applying generative adversarial network for intrusion detection. Comput Secur. 2023;125(9):103054. doi:10.1016/j.cose.2022.103054.
7. Djenouri Y, Nabil Belbachir A, Belhadi A, Michalak T, Srivastava G. Next-Gen metaverse security through intrusion detection enhanced by transformers and GANs. IEEE Internet Things J. 2025;12(12):20640–51. doi:10.1109/jiot.2025.3545803.
8. Mahmoudi I, Boubiche DE, Athmani S, Toral-Cruz H, Chan-Puc FI. Toward generative AI-based intrusion detection systems for the Internet of Vehicles (IoV). Future Internet. 2025;17(7):310. doi:10.3390/fi17070310.
9. Lamichhane P, Rawat DB. Quantum machine learning: recent advances, challenges, and perspectives. IEEE Access. 2025;13(6):94057–105. doi:10.1109/access.2025.3573244.
10. Rahman MA, Shahriar H, Clincy V, Hossain MF, Rahman M. A quantum generative adversarial network-based intrusion detection system. In: Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC); 2023 Jun 26–30; Torino, Italy. p. 1810–5.
11. Abreu D, Rothenberg CE, Abelém A. QML-IDS: quantum machine learning intrusion detection system. In: Proceedings of the 2024 IEEE Symposium on Computers and Communications (ISCC); 2024 Jun 26–29; Paris, France. p. 1–6.
12. Cirillo F, Esposito C. Intrusion detection using quantum generative adversarial networks: a federated approach with noisy simulators. In: Proceedings of the IET Space and Communications Conference 2025; 2025 Jun 17–18; London, UK. p. 31–5.
13. Naaman R, de Magalhaes FG, Ouattara JY, Nicolescu G. Quantum enhanced anomaly detection for ADS-B data using hybrid deep learning. arXiv:250915991. 2025.
14. Shahbaz Khan M, Ahmad J, Al-Dubai A, Pitropakis N, Ghaleb B, Ullah A, et al. Chaotic quantum encryption to secure image data in post quantum consumer technology. IEEE Trans Consum Electron. 2024;70(4):7087–101. doi:10.1109/tce.2024.3415411.
15. Nicesio OK, Leal AG, Gava VL. Quantum machine learning for network intrusion detection systems, a systematic literature review. In: Proceedings of the 2023 IEEE 2nd International Conference on AI in Cybersecurity (ICAIC); 2023 Feb 7–9; Houston, TX, USA. p. 1–6.
16. Islam M, Turkeli S, Ozaydin F. A survey of quantum generative adversarial networks: architectures, use cases, and real-world implementations. arXiv:250618002. 2025.
17. Iliyasu AS, Deng H. N-GAN: a novel anomaly-based network intrusion detection with generative adversarial networks. Int J Inf Technol. 2022;14(7):3365–75.
18. Balaji S, Narayanan SS. Dynamic distributed generative adversarial network for intrusion detection system over internet of things. Wirel Netw. 2023;29(5):1949–67. doi:10.1007/s11276-022-03182-8.
19. Vu L, Nguyen QU, Nguyen DN, Hoang DT, Dutkiewicz E. Deep generative learning models for cloud intrusion detection systems. IEEE Trans Cybern. 2023;53(1):565–77. doi:10.1109/tcyb.2022.3163811.
20. Sharafaldin I, Lashkari AH, Ghorbani AA, et al. Toward generating a new intrusion detection dataset and intrusion traffic characterization. Int Conf Inf Syst Secur Priv. 2018;1(2018):108–16. doi:10.5220/0006639801080116.

21. Sharafaldin I, Habibi Lashkari A, Ghorbani AA. A detailed analysis of the CICIDS2017 data set. In: International Conference on Information Systems Security and Privacy. Berlin/Heidelberg, Germany: Springer; 2018. p. 172–88.
22. Hindy H, Atkinson R, Tachtatzis C, Colin JN, Bayne E, Bellekens X. Utilising deep learning techniques for effective zero-day attack detection. Electronics. 2020;9(10):1684. doi:10.3390/electronics9101684.