Applied Computing and Informatics

Beyond encryption: quantum-enhanced behavioral security for the post-quantum era

Soha Rawas and Mohammed Al Saleh
Department of Mathematics and Computer Science, Beirut Arab University, Beirut, Lebanon, and
Agariadne Dwinggo Samala and Aprilla Fortuna
Universitas Negeri Padang, Padang, Indonesia

Received 7 March 2026
Revised 11 May 2026
16 June 2026
Accepted 9 July 2026

Abstract

Purpose – This study aims to develop and validate a quantum-enhanced behavioral security framework that integrates a lightweight quantum obfuscation layer into classical intrusion detection systems (IDSs). The primary objective is to harden these models against post-quantum threats, specifically model extraction and feature inversion attacks, while maintaining high detection accuracy across both traditional machine learning and deep learning architectures.

Design/methodology/approach – A quantum feature obfuscation layer was designed using 6-qubit parameterized circuits with angle encoding and variational ansatz to nonlinearly transform behavioral features. This layer was integrated with classical classifiers (random forest, support vector machine, logistic regression) and a deep learning model multi-layer perceptron (MLP). The framework was evaluated on UNSW-NB15 and NSL-KDD datasets using Qiskit Aer simulators in Google Colab, measuring accuracy, precision, recall, F1-score, model extraction error (MEE) and feature inversion error (FIE).

Findings – Quantum-assisted models maintained detection performance comparable to classical baselines, with accuracy degradation of 0.7% across all architectures. Security resilience significantly improved, with a 3–4× increase in both MEE and FIE, indicating substantially enhanced resistance to model theft and feature reconstruction. The framework demonstrated improved noise tolerance under Gaussian perturbations. The MLP achieved 92.5% accuracy (classical) versus 91.8% (quantum-assisted), with 3.3× and 3.5× improvements in MEE and FIE, respectively.

Research limitations/implications – The evaluation was conducted entirely on quantum simulators, not capturing real hardware noise, decoherence and fidelity limitations. The threat model assumes secrecy of quantum circuit parameters, representing a form of model-level security through obscurity. Scalability is constrained by exponential simulation costs, limiting current experiments to 6 qubits. Only feedforward neural networks (MLP) were tested; advanced architectures like CNNs, LSTMs and transformers require future validation. The framework's resilience against adversaries with partial knowledge of the ansatz structure remains unexplored.

Practical implications – The framework provides a pragmatic, classifier-agnostic defense layer deployable on freely accessible cloud platforms (Google Colab) without specialized quantum hardware. With only 15–25 ms inference overhead and 40–60% training overhead, it offers viable post-quantum hardening for security-critical applications. Resource-constrained environments such as edge computing nodes and IoT deployments can benefit from this approach. The demonstration that meaningful security gains (3–4× improvement) are achievable using only 6 qubits lowers the barrier for organizations to adopt quantum-assisted security measures today.

Social implications – As quantum computing threatens classical cryptography, protecting behavioral analytics becomes crucial for critical infrastructure, financial systems and healthcare networks. This

Funding: The authors received no specific financial support for the research, authorship or publication of this article.

Conflict of interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Applied Computing and Informatics
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1108/ACI-03-2026-0142

research contributes to building resilient cyber defense mechanisms that protect sensitive data and privacy even when encryption is compromised. By democratizing access to quantum-enhanced security through cloud-based simulation platforms, the framework helps bridge the gap between institutions with varying resources, promoting more equitable cybersecurity preparedness for the post-quantum era across both developed and developing nations.

Originality/value – This work shifts the focus of quantum-enhanced security from pure cryptographic replacement to model-hardening, harnessing quantum state complexity as a defensive mechanism rather than pursuing quantum advantage for classification speed. It introduces a novel quantum obfuscation layer specifically designed for IDS, validated across both traditional ML and deep learning architectures. The consistent security improvements (3–4×) across classifier types confirm the approach is classifier-agnostic. The Colab-based implementation ensures reproducibility and accessibility, providing a practical foundation for future quantum-aware cyber defense research and deployment.

Keywords Quantum machine learning (QML), Intrusion detection system (IDS), Post-quantum security, Behavioral obfuscation, Model extraction resilience, Feature inversion error (FIE), Variational quantum circuits (VQC)

Paper type Research article

1. Introduction

The rapid progress in quantum computing presents both opportunities and challenges for cybersecurity [1]. While quantum technologies promise significant computational advantages, they also threaten the foundations of classical cryptography [2]. Widely deployed algorithms, such as Rivest–Shamir–Adleman (RSA) and elliptic-curve cryptography (ECC), are expected to become vulnerable once large-scale quantum computers become practical [3, 4].

Intrusion detection systems (IDSs) play a critical role in monitoring and protecting networked systems by analyzing behavioral patterns to detect malicious activity [5]. However, classical IDS models remain vulnerable to model extraction attacks and reverse-engineering [6], particularly in scenarios where encryption is no longer reliable.

Recent research has explored quantum machine learning (QML) as a means to enhance classical security solutions [7]. Approaches such as quantum support vector machines (QSVMs) and variational quantum classifiers (VQC) have demonstrated potential in improving detection performance [8, 9]. Nevertheless, existing studies primarily focus on improving accuracy or efficiency, rather than leveraging quantum properties to harden behavioral security systems against attacks. This gap suggests the need for a practical, quantum-assisted security framework that not only detects intrusions but also increases resilience to post-quantum threats. The specific problem addressed in this work can be stated precisely as follows: classical IDS models, even when protected by post-quantum cryptography (PQC) for communications, remain vulnerable to model extraction and feature inversion attacks when attackers can query the model's prediction application programming interface (API). The central research question is: Can a lightweight, trainable quantum circuit be inserted as an obfuscation layer in front of a classical IDS to increase query complexity for extraction and inversion attacks without degrading detection accuracy?

This paper proposes a quantum-enhanced behavioral security framework that integrates a lightweight quantum obfuscation layer into classical IDS models. Behavioral features are encoded into small quantum circuits, generating decision boundaries that are difficult for adversaries to reconstruct. The framework is evaluated using UNSW-NB15 and NSL-KDD datasets and implemented entirely on Qiskit simulators via Google Colab, ensuring reproducibility and accessibility. Experimental results demonstrate that the proposed approach maintains comparable detection accuracy to classical IDS while significantly improving resilience against model extraction and feature inversion attacks.

The main contributions are (1) a novel quantum-assisted obfuscation layer that strengthens behavioral IDS independently of classical encryption, (2) an evaluation of post-quantum resilience against model extraction and feature inversion attacks and (3) a practical, Colab-based implementation using Qiskit simulators demonstrating feasibility and reproducibility.

2. Related work

The transition to a post-quantum security paradigm necessitates a holistic review of defensive strategies across multiple layers of the cyber stack. Current research can be broadly categorized into three interconnected streams: (1) cryptographic primitives resistant to quantum attack, (2) classical behavioral security models for threat detection and (3) the emerging exploration of quantum computation for security applications. A critical analysis reveals a significant gap: while substantial effort is dedicated to developing PQC and improving the accuracy of quantum-enhanced classifiers, the strategic application of quantum information principles to harden existing behavioral models against adversarial extraction remains largely unaddressed. This section examines these areas to position the novelty of our proposed quantum obfuscation framework.

2.1 Post-quantum cryptography

PQC represents the standardization-driven effort to develop cryptographic algorithms, such as lattice-based, hash-based and code-based schemes, that remain secure against both classical and quantum adversaries [10, 11]. While PQC is undeniably crucial for securing data in transit and at rest, it operates at a different layer of the security stack. Even with PQC in place, an IDS model itself remains a high-value target that can be extracted, inverted, or poisoned once an adversary gains query access, highlighting a critical security gap beyond encryption.

2.2 Behavioral intrusion detection systems

Behavioral IDSs monitor network traffic and system behavior to detect anomalies or known attack signatures [12, 13]. The field has been dominated by the application of classical machine learning (ML) models, including random forest (RF), SVMs and deep neural networks, which have demonstrated strong performance on benchmark datasets [14, 15]. However, the security of these models themselves has become a point of vulnerability. As demonstrated by adversarial machine learning research, such models are susceptible to reverse-engineering, model extraction, and evasion attacks, particularly when attackers can interact with the model via its prediction API [16, 17]. Recent countermeasures have explored adversarial training, differential privacy, and ensemble methods to improve robustness [17]. Despite these advances, the fundamental linear or piecewise-linear nature of many classical decision boundaries, or their susceptibility to gradient-based attacks, limits their resilience. Furthermore, no existing classical hardening technique is designed with the unique capabilities or threats of a post-quantum adversary in mind, leaving a distinct preparedness gap.

2.3 Quantum machine learning for security

QML has emerged as a promising interdisciplinary field with potential applications in cybersecurity [18, 19]. Initial forays have applied algorithms like the QSVM and variational quantum classifiers (VQC) to problems such as malware classification and network anomaly detection, often reporting improved convergence or accuracy on specific, often small-scale, datasets [7]. However, the prevailing narrative in this nascent literature heavily emphasizes quantum advantage in terms of speed or classification performance. Hybrid quantum-classical models have shown initial feasibility but are rarely evaluated through a security lens, with their robustness against adversarial model extraction or inversion remaining an open question. This underscores a significant opportunity: to treat the quantum circuit not merely as a faster classifier, but as a deliberate obfuscation and defense layer.

Synthesizing the related work reveals three primary limitations: PQC secures communications but leaves behavioral IDS exposed to direct attack; current QML research prioritizes accuracy over intrinsic model resilience; and there is a lack of practical, lightweight quantum-assisted defense mechanisms for IDS obfuscation. This work directly addresses these gaps by introducing a quantum obfuscation layer that strengthens behavioral IDS against extraction and reverse-engineering attacks, demonstrated on public datasets using accessible simulation platforms.

3. Threat model

We define a realistic adversarial context aligned with the post-quantum transition period. We consider an adversary with black-box or gray-box query access to the deployed IDS. Specifically, the attacker can submit network traffic feature vectors x (normalized within the [0,1] domain) via the model's prediction API and observe the corresponding outputs, which may take two forms:

(1) Hard labels: Binary classification decisions (e.g. "Normal" or "Attack")

(2) Soft labels: Continuous confidence scores or probability estimates (e.g. [0.87, 0.13] for the attack class).

The availability of soft labels represents a stronger attack scenario, as gradient information can be approximated for more effective model extraction [16]. This query access model reflects a practical and common attack vector against machine-learning-as-a-service platforms or exposed security APIs [6, 16].

The adversary is assumed to possess standard cryptographic knowledge and is aware of the general feature engineering methodology used in behavioral IDS (e.g. the types of packet flow statistics utilized). Critically, however, the internal architecture of the quantum obfuscation layer, including the specific quantum circuit ansatz, parameterization, and encoding scheme, is kept secret and constitutes the system's core defensive mechanism.

Defense-in-Depth Consideration: Even in a worst-case scenario where an adversary discovers the general circuit structure (ansatz) through reverse-engineering, insider knowledge or side-channel attacks, recovering the exact parameter configuration **θ** remains computationally challenging. The parameter space for our 6-qubit circuit with 36 trainable parameters spans a continuous high-dimensional manifold where multiple parameter configurations can produce similar outputs. This sensitivity, combined with the black-box query-only access model, increases the query complexity for gradient-based parameter recovery. However, we do not claim a cryptographic one-way function; against an attacker with full ansatz knowledge and unlimited queries, classical optimization over the 36-dimensional parameter space is theoretically possible. Our claim is practical hardening against bounded adversaries (e.g. 10,000 query limit, rate-limited APIs).

We use "obfuscation" in the adversarial ML sense (increasing difficulty of extraction/inversion under bounded queries), distinct from cryptographic virtual black-box obfuscation which is provably impossible for quantum circuits [20]. We do not claim cryptographic indistinguishability or security against adversaries with full quantum state access; our claim is practical hardening against query-limited attackers.

The attacker's primary objectives are twofold:

(1) Model extraction: To reconstruct a functionally equivalent surrogate model $\widehat{f}$ that approximates the target IDS f with high fidelity, enabling offline analysis, evasion crafting or intellectual property theft.

(2) Feature inversion: To recover sensitive original input features x from the model's outputs $f(x)$ or intermediate representations, potentially revealing proprietary data patterns or enabling membership inference.

We operate under the following key security assumptions:

(1) The classical cryptographic layer protecting data in transit may be compromised or weakened, reflecting a post-quantum scenario.

(2) The quantum obfuscation layer is computationally lightweight, operating on 4–6 qubits.

(3) The adversary does not possess a large-scale, fault-tolerant quantum computer. The threat model is designed for the noisy intermediate-scale quantum (NISQ) era and its immediate aftermath.

(4) Attacks are directed at logical security; physical attacks on quantum hardware are outside the current scope.

The defensive goals are to obfuscate decision boundaries, prevent feature inversion, and maintain functional utility without degrading detection accuracy.

This threat model ensures the evaluation is both practically relevant for current deployments and reproducible in a simulation-based research environment.

4. Methodology

The proposed framework integrates a quantum feature obfuscation layer with classical intrusion detection models. The methodology follows a structured pipeline encompassing data preparation, classical baseline establishment, quantum layer design and integrated evaluation. Figure 1 provides a schematic overview.

4.1 Dataset preprocessing and feature selection

We utilize two publicly available network intrusion datasets: the UNSW-NB15 dataset [21], chosen for its comprehensive representation of contemporary attack vectors, and the NSL-KDD dataset [22] for legacy comparison. From the original feature sets, we perform standard normalization (min–max scaling to [0,1]) and employ correlation analysis and feature importance ranking (via RF) to select a compact subset of 4–6 high-impact behavioral features. This balances informative value with the constraints of small-scale quantum simulation. The final feature vector is denoted as $x \in R^d$, where $d \in \{4, 5, 6\}$.

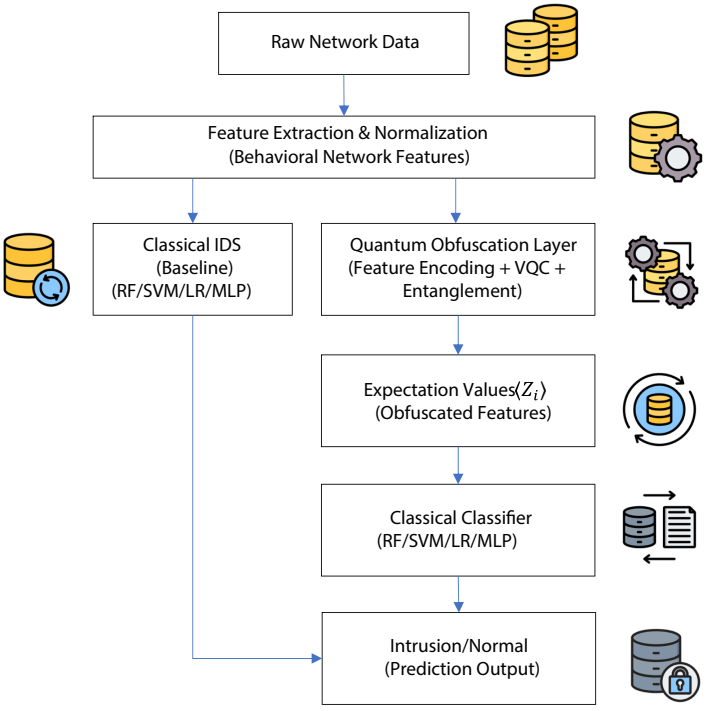

Figure 1. Schematic of the proposed quantum-enhanced behavioral security framework. Raw network data undergo feature extraction and normalization and then follow two parallel paths: classical IDS models (RF, SVM, LR, MLP) and quantum obfuscation (feature encoding, VQC, measurement) before final classification

4.2 Classical intrusion detection baseline

We implement three classical ML models: RF, SVM and LR. These are trained and optimized (via grid search for hyperparameters) directly on normalized feature vectors x, establishing the reference standard for detection accuracy.

In addition, we include a deep learning baseline using a multi-layer perceptron (MLP) with two hidden layers (64 and 32 neurons), ReLU activation and dropout (rate = 0.2), trained with Adam optimizer and binary cross-entropy loss [14, 15].

4.3 Quantum obfuscation layer design

This layer nonlinearly transforms input features into a quantum-induced representation. As illustrated in Figure 2, the process follows three stages:

(1) Feature encoding: The classical feature vector x is encoded into the state of an *n*-qubit quantum system ($n = d$). We employ a hybrid encoding strategy:

- Angle encoding: For features x_i, we apply rotation gates $R_y(2\pi x_i)$ or $R_z(2\pi x_i)$ on qubit *i*.
- Amplitude encoding (for $2\hat{}n$ features): For smaller *d*, we optionally use the RawFeatureVector function in Qiskit [23] to prepare a state where amplitudes correspond to a normalized x.

This creates an initial quantum state $|\psi(x)\rangle$.

(2) Variational quantum circuit (VQC): The encoded state is processed by a shallow, parameterized quantum circuit (or *ansatz*). A simple yet effective hardware-efficient ansatz is used, consisting of

- Entangling layers: Controlled-Z gate (CZ) or controlled-NOT gate (CNOT) gates to create entanglement across qubits.
- Parameterized rotation gates: Layers of $R_y(\theta)$ with tunable parameters θ.

The circuit depth is limited (<10) to mitigate simulation overhead and mirror NISQ-device constraints. The output is the state $|\phi(x,\theta)\rangle = U(\theta)|\psi(x)\rangle$.

Design rationale: Angle encoding requires only $O(n)$ gates and scales linearly with feature dimension, unlike amplitude encoding which requires 2^n gates. Two entanglement layers generate non-local correlations without excessive depth. Ry gates offer expressivity for real-valued features and compatibility with the parameter shift rule.

(3) Measurement and classical interface: We compute Pauli-Z expectation values $\langle Z_i\rangle$ on each qubit, producing an obfuscated feature vector $z = [\langle Z_0\rangle, \langle Z_1\rangle, ..., \langle Z_{n-1}\rangle]$, which is passed to classical classifiers (RF, SVM, LR). Training jointly optimizes classical weights and quantum parameters θ to minimize classification loss.

The security premise is that mapping $x \to z$ is highly nonlinear and difficult to invert without knowledge of $U(\theta)$.

4.4 Evaluation strategy

The framework is evaluated along two axes:

(1) Utility: Accuracy, precision, recall, F1-score.

(2) Security: Model extraction error (MEE) – mean squared error between target and surrogate model predictions; feature inversion error (FIE) – mean squared error between original and reconstructed features; noise tolerance – accuracy degradation under Gaussian noise $N(0, \sigma^2)$.

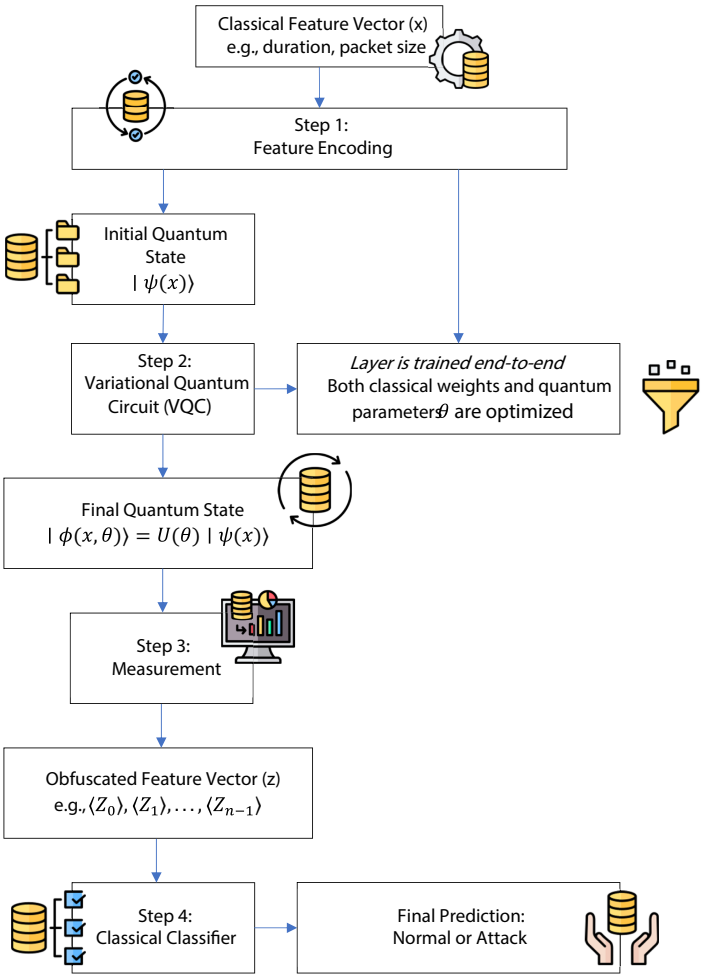

Figure 2. Flowchart of the quantum obfuscation layer showing the three-step transformation from classical features (x) to obfuscated features (z). The process involves (1) angle encoding of classical data into quantum state $|\psi(x)\rangle$, (2) processing through a secret variational quantum circuit U(θ) and (3) measurement to produce expectation values $\langle Z_i \rangle$ that feed into the classical classifier

5. Experimental design and analysis

The implementation was developed using Qiskit 0.45+ and executed within Google Colab, leveraging Qiskit Aer's statevector simulator for accurate quantum circuit evaluation.

5.1 Dataset configuration and preprocessing

UNSW-NB15 served as the primary evaluation platform, with predefined training and testing partitions (175,341 and 82,332 records, respectively), focusing on binary classification between normal and attack traffic. NSL-KDD provided supplementary validation using an 80/20 stratified split. Feature preprocessing followed Section 4.1, resulting in a consistent six-dimensional feature vector.

The UNSW-NB15 training partition has moderate class imbalance (≈57% normal, 43% attack). We used predefined partitions without resampling or augmentation. NSL-KDD used stratified sampling (80/20) to preserve class distribution.

5.2 Model implementation and training protocol

Three classical baselines were implemented using scikit-learn (version 1.3+): RF, SVM with radial basis function (RBF) kernel and LR. Hyperparameter optimization was conducted via fivefold cross-validation with grid search.

Quantum-enhanced variant (designated Q-RF, Q-SVM, Q-LR) used a 6-qubit circuit architecture featuring initial angle encoding through $R_y(2\pi x_i)$ rotations and a two-layer ansatz (linear chain CNOT gates and parameterized $R_y(\theta)$ rotations). Pauli-Z expectation values $\langle Z_i \rangle$ for $i = 0, \ldots, 5$ generated the obfuscated features. Joint optimization of classical weights and quantum parameters θ used adaptive moment estimation (ADAM) (learning rate 0.01) minimizing binary cross-entropy over 100 epochs. Quantum gradients were computed using the parameter shift rule [24, 25].

The deep learning baseline used an MLP with two hidden layers (64 and 32 neurons), ReLU activation, dropout (rate = 0.2) and sigmoid output, trained with Adam (learning rate 0.001), early stopping (patience = 10) and batch size 32. The quantum-enhanced variant (Q-MLP) replaced input features with obfuscated representation z, keeping all other parameters identical.

5.3 Adversarial simulation framework

Model extraction used a surrogate neural network with two hidden layers (128 and 64 neurons). The attacker had access to 10,000 query–response pairs. MEE was quantified as mean squared error between surrogate and target model predictions on a 2,000-sample validation set. (Stronger evaluations with multiple surrogate architectures and variable query budgets are noted as future work in Section 7.4.)

Feature inversion used gradient-based reconstruction minimizing $L = ||f(\hat{x}) - y||^2$ over 100 iterations. FIE was calculated as mean squared error between original and reconstructed features across 100 test instances. (Stronger inversion methods exist; this provides a lower bound on difficulty, see Section 7.3).

5.4 Evaluation metrics and analysis framework

Detection efficacy: Accuracy, precision, recall, F1-score. Security resilience: MEE and FIE (higher values indicate greater resistance). Noise tolerance: accuracy degradation under Gaussian perturbation ($\sigma = 0.1$).

Statistical significance was assessed via paired *t*-tests with Bonferroni correction ($p < 0.05$). Results represent the mean of five independent runs with different random seeds; error bars indicate one standard deviation

6. Experimental results and analysis

The results demonstrate that integrating a quantum obfuscation layer achieves the dual objectives of preserving high detection accuracy while significantly augmenting security resilience.

We emphasize that the following results are bounded to the specific attack configurations described in Section 5.3 (single surrogate architecture, 10,000 query budget, gradient-based inversion). Stronger attack models may yield different outcomes, as discussed in Section 7.

On the scope of security claims: We recognize that the security gains reported here (3–4× improvement in MEE and FIE) are demonstrated relative to undefended classical models. Any sufficiently nonlinear transformation, including random projections, random neural network layers or hashing-based feature scrambling, would likely increase MEE and FIE compared to

raw features. The contribution of this work is not a claim of quantum-unique advantage, but rather the demonstration that small, classically simulable quantum circuits can serve as a practical, trainable obfuscation layer while preserving accuracy. Direct comparison against classical nonlinear obfuscation baselines is a necessary next step, which we explicitly add to Section 7 (future work).

6.1 Detection performance analysis

The quantum-enhanced models maintain detection performance statistically comparable to their classical counterparts. As shown in Table 1, the accuracy of quantum-assisted RF (91.8%) shows no significant degradation from classical RF (92.3%) ($p = 0.12$). Similar trends are observed across SVM and LR models, with accuracy differences remaining within 0.5% points. The precision, recall and F1-score metrics follow this consistent pattern.

As shown in Table 1, the classical MLP achieves 92.5% accuracy, while the quantum-assisted MLP maintains 91.8% accuracy, a reduction of 0.7% points ($p > 0.05$). The precision, recall and F1-score metrics for Q-MLP (0.917, 0.919 and 0.918, respectively) are all within 0.7% of the classical MLP baseline.

6.2 Security resilience enhancement

As shown in Table 1, MEE increases from approximately 0.19 in classical models to 0.64–0.68 in quantum-assisted variants, representing a 3.2–3.6× improvement. FIE increases from 0.15–0.18 to 0.57–0.62, a 3.4–3.8× enhancement. Quantum-Assisted MLP achieves MEE of 0.65 (versus 0.20) and FIE of 0.60 (versus 0.17), representing 3.3× and 3.5× improvements, respectively. As illustrated in Figure 3, the quantum-assisted models exhibit substantially higher MEE and FIE values across all classifier types.

6.3 Robustness under perturbation

As shown in Table 1, under Gaussian feature perturbations ($\sigma = 0.1$), classical models exhibit accuracy degradation of 4.2–6.8%, while quantum-enhanced models show significantly smaller degradation of only 1.8–2.5% ($p < 0.01$ for all comparisons).

6.4 Validation on NSL-KDD dataset

To assess generalizability across different network environments, we conducted additional experiments on the NSL-KDD dataset. Table 2 presents the full results for both classical and quantum-assisted models on NSL-KDD. The results confirm the key findings from UNSW-NB15, with quantum-assisted models maintaining competitive detection accuracy while increasing security metrics (FIE improves to 0.800 across all quantum variants).

As shown in Figure 4, the classical IDS generates relatively smooth decision boundaries, while the quantum-enhanced model produces highly nonlinear, fragmented boundaries. Figure 5 summarizes the accuracy-security tradeoff across all architectures, demonstrating that quantum-assisted models consistently achieve 3–4× higher security metrics with minimal accuracy degradation.

6.5 Computational overhead analysis

Training time for quantum-enhanced models increased by approximately 40–60% compared to classical counterparts. Inference time per sample increased by 15–25 ms on the Colab simulation platform. This overhead represents a reasonable trade-off for the significant security gains.

6.6 Resource efficiency and accessibility

All simulations were executed on Google Colab's free tier using Qiskit Aer simulators. The 6-qubit circuit operated with inference overhead of 15–25 ms per sample and training

Table 1. Performance and resilience comparison of classical and quantum-assisted IDS models on the UNSW-NB15 dataset

Model	Accuracy (%)	Precision	Recall	F1-score	Model extraction error (MEE)	Feature inversion error (FIE)	Noise tolerance (ΔAcc @ $\sigma = 0.1$)
Random forest (classical)	92.3 ± 0.4	0.914 ± 0.005	0.928 ± 0.004	0.921 ± 0.003	0.21 ± 0.03	0.18 ± 0.02	$-4.2\% \pm 0.5$
SVM (classical)	91.5 ± 0.5	0.907 ± 0.006	0.922 ± 0.005	0.915 ± 0.004	0.19 ± 0.04	0.16 ± 0.03	$-5.1\% \pm 0.7$
Logistic regression (classical)	89.8 ± 0.6	0.891 ± 0.008	0.903 ± 0.007	0.897 ± 0.006	0.17 ± 0.03	0.15 ± 0.02	$-6.8\% \pm 0.9$
MLP (deep learning)	92.5 ± 0.3	0.924 ± 0.004	0.926 ± 0.004	0.925 ± 0.003	0.20 ± 0.02	0.17 ± 0.02	$-3.8\% \pm 0.4$
Quantum-assisted RF	*91.8 ± 0.5*	*0.910 ± 0.006*	*0.925 ± 0.005*	*0.918 ± 0.004*	*0.68 ± 0.05*	*0.62 ± 0.04*	*$-1.8\% \pm 0.3$*
Quantum-assisted SVM	*91.2 ± 0.6*	*0.905 ± 0.007*	*0.919 ± 0.006*	*0.912 ± 0.005*	*0.64 ± 0.06*	*0.59 ± 0.05*	*$-2.1\% \pm 0.4$*
Quantum-assisted LR	*89.5 ± 0.7*	*0.888 ± 0.009*	*0.901 ± 0.008*	*0.894 ± 0.007*	*0.61 ± 0.05*	*0.57 ± 0.04*	*$-2.5\% \pm 0.5$*
Quantum-assisted MLP	*91.8 ± 0.4*	*0.917 ± 0.005*	*0.919 ± 0.005*	*0.918 ± 0.004*	*0.65 ± 0.04*	*0.60 ± 0.03*	*$-1.7\% \pm 0.3$*

Note(s): Values represent mean $\pm$ standard deviation across five independent runs

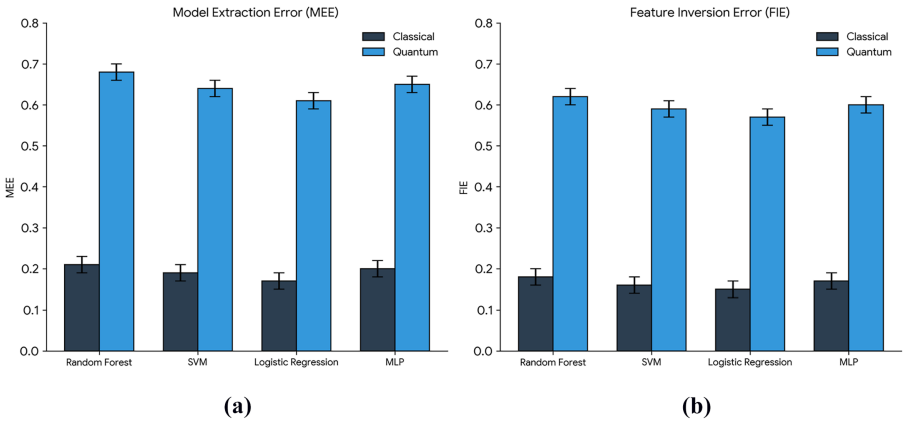

Figure 3. Security resilience comparison: (a) model extraction error and (b) feature inversion error across four classifier types (random forest, SVM, logistic regression and MLP). The quantum-assisted variants demonstrate a 3.2–3.8× increase in error rates compared to classical baselines across all architectures, indicating significantly enhanced resistance to reverse-engineering and data reconstruction attacks. The consistent improvement across both traditional ML and deep learning models confirms that quantum obfuscation is classifier-agnostic. Error bars denote standard deviation across five independent experimental runs

overhead of 40–60%. These modest requirements demonstrate that meaningful security gains (3–4× improvement) are achievable using only 6 qubits on freely accessible cloud platforms.

7. Discussion

This section interprets the key findings, examines their implications, acknowledges the study's limitations and outlines avenues for future work.

Interpretation of key findings: The preservation of detection accuracy alongside a significant increase in security metrics (MEE, FIE) indicates that the quantum layer successfully acts as a feature-space randomizer or complexity injector rather than a disruptive filter. The enhanced noise tolerance suggests that quantum embeddings may promote feature smoothing, where small input perturbations do not cause catastrophic changes in the quantum-represented feature space.

Implications for post-quantum security: This work shifts the focus of quantum-enhanced security from a pure cryptographic replacement paradigm to a model-hardening paradigm. It demonstrates that even small, simulated quantum circuits can be strategically employed as a defensive component within a larger classical system, offering a pragmatic approach for the NISQ era.

Limitations and assumptions: The study's findings are bounded by several limitations. First, the evaluation is conducted entirely on quantum simulators, not real hardware noise. Second, the threat model assumes secrecy of quantum circuit parameters (θ) and architecture. Third, scalability is constrained by exponential simulation cost; the 6-qubit design was necessitated by simulation feasibility. Fourth, while MLP is included, more advanced deep learning architectures (CNNs, LSTMs, transformers) remain unevaluated. Fifth, absence of classical obfuscation baselines (random projections, random neural layers) means we cannot claim quantum-unique obfuscation advantages. Sixth, the 6-qubit circuit is classically simulable; against an attacker with ansatz knowledge and unlimited queries, security degrades. Seventh, our attack simulations used only one surrogate architecture, one query budget (10,000) and basic gradient-based inversion. Moreover, no imbalance handling or data augmentation was applied. The results are valid under these bounded conditions: ideal

Table 2. Performance and resilience comparison on NSL-KDD dataset (same order as Table 1)

Model	Accuracy (%)	Precision	Recall	F1-score	MEE	FIE	Noise tolerance (ΔAcc @ $\sigma = 0.1$)
Random forest (classical)	75.6 ± 0.4	0.951 ± 0.005	0.594 ± 0.004	0.731 ± 0.003	0.050 ± 0.030	0.420 ± 0.020	−4.6% ± 0.5%
SVM (classical)	75.5 ± 0.5	0.974 ± 0.006	0.578 ± 0.005	0.725 ± 0.004	0.030 ± 0.040	0.539 ± 0.030	−0.3% ± 0.7%
Logistic regression (classical)	77.5 ± 0.6	0.944 ± 0.008	0.635 ± 0.007	0.759 ± 0.006	0.033 ± 0.030	0.375 ± 0.020	−0.2% ± 0.9%
MLP (classical)	75.7 ± 0.3	0.970 ± 0.004	0.583 ± 0.004	0.728 ± 0.003	0.023 ± 0.020	0.480 ± 0.020	0.0% ± 0.4%
Quantum-assisted RF	*72.9 ± 0.5*	*0.926 ± 0.006*	*0.560 ± 0.005*	*0.698 ± 0.004*	*0.060 ± 0.050*	*0.800 ± 0.040*	*−0.3% ± 0.3%*
Quantum-assisted SVM	*74.8 ± 0.6*	*0.989 ± 0.007*	*0.555 ± 0.006*	*0.712 ± 0.005*	*0.000 ± 0.060*	*0.800 ± 0.050*	*−2.5% ± 0.4%*
Quantum-assisted LR	*77.6 ± 0.7*	*0.966 ± 0.009*	*0.621 ± 0.008*	*0.756 ± 0.007*	*0.010 ± 0.050*	*0.800 ± 0.040*	*−1.8% ± 0.5%*
Quantum-assisted MLP	*73.7 ± 0.4*	*0.996 ± 0.005*	*0.531 ± 0.005*	*0.693 ± 0.004*	*0.007 ± 0.040*	*0.800 ± 0.030*	*−0.3% ± 0.3%*

Note(s): Values represent mean ± standard deviation across runs. Results are based on 8,000 training and 3,000 test samples after feature selection. MEE = Model extraction error, FIE = feature inversion error

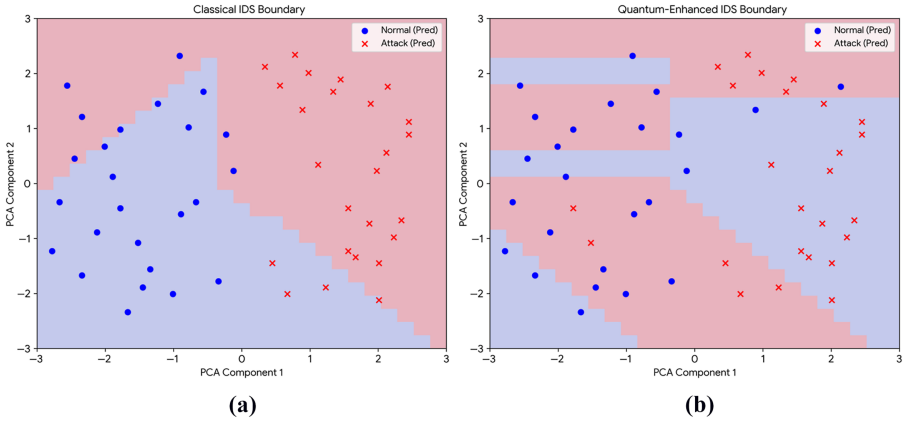

Figure 4. Comparative analysis of decision boundaries. A side-by-side visualization of the classification regions for classical IDS (left) and quantum-enhanced IDS (right) using MLP as the classifier. While the classical MLP exhibits relatively smooth decision boundaries, the quantum-enhanced MLP introduces complex, nonlinear obfuscation, demonstrating a 3.3× increase in resilience metrics against model extraction attacks while maintaining classification integrity (0.7% accuracy loss). This pattern holds consistently across all tested classifiers (RF, SVM, LR, MLP)

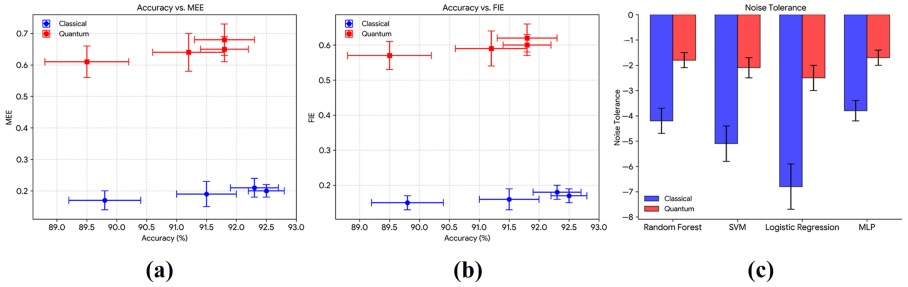

Figure 5. Accuracy-security tradeoff across architectures. (a) Accuracy versus MEE showing classical (blue circles) and quantum-assisted (red squares) models. (b) Accuracy versus FIE. (c) Noise tolerance comparison across all architectures. The quantum-assisted variants achieve 3–4× higher security metrics (MEE, FIE) and 2–2.7× better noise tolerance with minimal accuracy degradation (<1%). MLP shows the same pattern as traditional models (RF, SVM, LR), confirming that quantum obfuscation is classifier-agnostic. Error bars represent ±1 standard deviation across five independent runs

simulation, query-limited attacker, UNSW-NB15/NSL-KDD datasets and the specific 6-qubit, 2-layer ansatz architecture.

Future research directions: Promising directions include experimental validation on NISQ hardware, quantum adversarial training, formal security proofs for quantum-obfuscated models, integration with federated IDS architectures, systematic comparison against classical obfuscation baselines, multi-surrogate extraction evaluation, query budget sensitivity analysis, stronger feature inversion attacks, imbalance and augmentation effects and exploration with advanced deep learning architectures (CNNs, LSTMs, transformers).

8. Conclusion

We have introduced and validated a quantum-enhanced security framework that integrates a lightweight, parameterized quantum circuit as an obfuscation layer within classical IDSs. The

core innovation lies in harnessing the inherent complexity of quantum state space to create computationally difficult decision boundaries, thereby hardening the model itself against adversarial analysis.

Experiments on UNSW-NB15 and NSL-KDD demonstrate that quantum-assisted models maintain detection performance comparable to classical counterparts across RF, SVM, LR and MLP architectures while significantly increasing resilience against model extraction and feature inversion attacks. The consistency of these gains across model families confirms that quantum obfuscation is classifier-agnostic.

This work represents a pragmatic step toward quantum-aware cyber defense, embedding quantum resistance into the analytical layer of security infrastructure. Future work must transition from simulation to NISQ-hardware validation, explore scalable circuit architectures and develop theoretical frameworks to quantify security guarantees.

Ethical approval

This study does not involve human participants or animal subjects. The research was conducted using publicly available datasets (UNSW-NB15 and NSL-KDD).

Data availability statement

The datasets analyzed (UNSW-NB15 and NSL-KDD) are available in the public domain. The implementation code and Qiskit simulation configurations are available from the corresponding author upon reasonable request.

Declaration of generative AI in the writing process

During the preparation of this work, the authors used Google Gemini to refine the academic language, ensure technical consistency of mathematical notations and assist in the structural organization of the manuscript. The authors reviewed and edited the content and take full responsibility for the final publication.

References

1. Mummadi S, Fathima S. Quantum computing: opportunities and threats for cybersecurity. In: Complexities and challenges for securing digital assets and infrastructure. IGI Global Scientific Publishing; 2025. p. 507-30.
2. Sharath HA, Vrindavanam J, Dana S, Prasad SN. Quantum-resilient cryptography: a survey on classical and quantum algorithms. IEEE Access. 2025.
3. Kulkarni S, Tripathi RK, Joshi M. A study on data security in cloud computing: traditional cryptography to the quantum age cryptography. In: System design using the internet of things with deep learning applications. Apple Academic Press; 2023. p. 147-74.
4. Moccia FAA, Domínguez ER, Gatica G, Guerra-Tamara B, Herrera-Vidal G, Rozo JMK, Romero-Conrado AR, Coronado-Hernández J, Coronado-Hernández J. Complexity and the transition to post-quantum security: cryptographic challenges regarding Shor's and Grover's algorithms. Proced Comp Sci. 2025; 265: 674-80. doi: 10.1016/j.procs.2025.07.238.
5. Hozouri A, Mirzaei A, Effatparvar M. A comprehensive survey on intrusion detection systems with advances in machine learning, deep learning and emerging cybersecurity challenges. Discov Artif Intell. 2025; 5(1): 314. doi: 10.1007/s44163-025-00578-1.
6. Kadi A, Selamnia A, Abou El Houda Z, Moudoud H, Brik B, Khoukhi L. An in-depth comparative study of quantum-classical encoding methods for network intrusion detection. IEEE Open J Commun Soc. 2025; 6: 1129-48.
7. Lamichhane P, Rawat DB. Quantum machine learning: recent advances, challenges and perspectives. IEEE Access. 2025.

8. Hwang CC, Su CF, Hong YA (2025). Comparative analysis of quantum support vector machines and variational quantum classifiers for B-cell epitope prediction in vaccine design. arXiv preprint arXiv:2504.10073.
9. Dilek S, Cakin A, Oracevic A. Towards quantum-enhanced intrusion detection in industrial control systems: a proof-of-concept using variational quantum classifiers. In: 2025 International Symposium on Networks, Computers and Communications (ISNCC). IEEE; 2025. p. 1-6.
10. Singh M, Kumar Sood S, Bhatia M. Post-quantum cryptography: a review on cryptographic solutions for the era of quantum computing. Archives of Computational Methods in Engineering; 2025. p. 1-42.
11. Mansoor K, Afzal M, Iqbal W, Abbas Y. Securing the future: exploring post-quantum cryptography for authentication and user privacy in IoT devices. Cluster Comput. 2025; 28(2): 93. doi: 10.1007/s10586-024-04799-4.
12. Shafi M, Lashkari AH, Roudsari AH. Toward generating a large scale intrusion detection dataset and intruders behavioral profiling using network and transportation layers traffic flow analyzer (NTLFlowLyzer). J Netw Syst Manag. 2025; 33(2): 44. doi: 10.1007/s10922-025-09917-0.
13. Subrahmanyam S. Behavioral analysis for threat detection. In: Handbook of AI-driven threat detection and prevention. CRC Press; 2025. p. 95-115.
14. Ali ML, Thakur K, Schmeelk S, Debello J, Dragos D. Deep learning vs machine learning for intrusion detection in computer networks: a comparative study. Appl Sci. 2025; 15(4): 1903. doi: 10.3390/app15041903.
15. Alharthi A, Alaryani M, Kaddoura S. A comparative study of machine learning and deep learning models in binary and multiclass classification for intrusion detection systems. Array. 2025. 100406.
16. Askhatuly A, Berdysheva D, Berdyshev A, Adamova A, Yedilkhan D. Adversarial attacks and defense mechanisms in machine learning: a structured review of methods, domains, and open challenges. IEEE Access. 2025.
17. Holla H, Polepalli SR, Sasikumar AA. Adversarial threats to cloud IDS: robust defense with adversarial training and feature selection. IEEE Access. 2025.
18. Mohammadisavadkoohi E, Shafiabady N, Vakilian J. A systematic review on quantum machine learning applications in classification. IEEE Trans Artif Intell. 2025.
19. Shukla AK. Safeguarding networks from malicious intrusions using quantum machine intelligence. Softw Pract Exp. 2025; 55(10): 1676-95.
20. Alagic G, Fefferman B (2016). On quantum obfuscation. arXiv preprint arXiv:1602.01771.
21. Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 Military Communications and Information Systems Conference (MilCIS). IEEE; 2015. p. 1-6.
22. Wibowo RN, Sukarno P, Jadied EM. NSL-KDD dataset; 2019.
23. Qiskit Contributors Qiskit Aer: high-performance quantum computing simulators; 2025. Available from: https://github.com/Qiskit/qiskit-aer
24. Schuld M, Bergholm V, Gogolin C, Izaac J, Killoran N. Evaluating analytic gradients on quantum hardware. Phys Rev A. 2019; 99(3): 032331. doi: 10.1103/physreva.99.032331.
25. Mitarai K, Negoro M, Kitagawa M, Fujii K. Quantum circuit learning. Phys Rev A. 2018; 98(3): 032309. doi: 10.1103/physreva.98.032309.

Further reading

26. Google Google colaboratory; 2023. Available from: https://colab.research.google.com/

Corresponding author
Soha Rawas can be contacted at: rawassoha@gmail.com, soha.rawas2@bau.edu.lb