

Adversarial Robustness of Quantum Supervised Learning for Network Intrusion Detection: White-Box Attacks, PGD-Aware Training, and Classical Benchmarking

Rahul Bhoge

bhoge.rahul@gmail.com

HVPM's College of Engineering and Technology, India

Ranjit Keole

HVPM's College of Engineering and Technology, India

Pravin Karde

Government Polytechnic , Amravati, India

Research Article

Keywords: adversarial robustness, quantum supervised learning, quantum neural network, intrusion detection, FGSM, PGD, UNSW-NB15, adversarial training

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10905898/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Quantum supervised learning (QSL) is increasingly investigated for intrusion detection, yet its behavior under adversarial manipulation remains insufficiently characterized in network-security settings. This study presents a controlled white-box evaluation of a four-qubit variational QSL intrusion detector and a matched classical multilayer perceptron (MLP) on the UNSW-NB15 benchmark. Categorical encoding, standardization, principal component analysis to four dimensions, and scaling to $[0, \pi]$ were applied consistently. A clean QSL model and a PGD-aware QSL model using 50% clean, 25% FGSM, and 25% PGD samples were evaluated against FGSM and seven-step PGD attacks at $\epsilon = 0.02\pi$, 0.05π , and 0.10π over ten matched random seeds. Clean and adversarially trained MLP controls used the same representation and attack protocol. PGD-aware QSL improved accuracy over clean QSL from 9.10% to 26.90% at PGD 0.05π and from 2.60% to 20.73% at PGD 0.10π , with large paired effect sizes. The improvement was accompanied by a clean-accuracy reduction from 70.73% to 61.57%. Robust MLP retained higher absolute adversarial accuracy, reaching 64.27% and 50.50% at PGD 0.05π and 0.10π , respectively. The results demonstrate that adversarial training materially strengthens the tested QSL model but do not support a quantum-advantage claim. The study provides a reproducible robustness benchmark emphasizing attack validation, repeated seeds, effect sizes, and classical controls. The implementation notebook and supporting experimental outputs used for reproducibility are publicly archived on Zenodo (Bhoge et al. 2026).

1. Introduction

Network intrusion detection systems increasingly rely on machine learning to identify malicious traffic, but predictive accuracy alone is not sufficient for security-critical deployment. Carefully constructed adversarial perturbations can cause trained classifiers to change their decisions while remaining bounded in the input space. FGSM provides a computationally inexpensive single-step attack, whereas PGD iteratively maximizes loss within a constrained perturbation set (Goodfellow et al., 2015; Madry et al., 2018).

Quantum machine learning (QML) extends supervised learning with parameterized quantum circuits and quantum feature maps. However, quantum classifiers are not inherently immune to adversarial manipulation. Lu et al. established the vulnerability of quantum learning systems to crafted adversarial examples (Lu et al. 2020) and later experimental work on superconducting qubits showed that adversarial training can materially improve quantum-classifier robustness (Ren et al. 2022). More recent studies continue to report vulnerabilities in quantum neural networks and quantum kernel classifiers and emphasize the need for controlled threat models and explicit defense evaluation (Winderl et al., 2024; Montalbano & Banchi, 2025).

The intrusion-detection literature is also moving toward QML, but a 2026 survey found that the field remains dominated by simulated environments, heterogeneous benchmarks, inconsistent evaluation practices, and limited statistical rigor (Kaissar et al. 2026). The same survey identifies UNSW-NB15

among the commonly used IDS benchmarks and calls for standardized evaluation and practical validation. UNSW-NB15 provides modern normal and attack traffic and includes established training and testing partitions (Moustafa & Slay 2015).

This work addresses a narrower question than whether quantum learning is generally superior to classical learning: can white-box adversarial training improve the robustness of a compact QSL intrusion detector, and how does the resulting robustness-accuracy trade-off compare with a matched classical neural baseline? The contribution is therefore an empirical robustness benchmark rather than a quantum-advantage claim.

The main contributions are:

- i. a four-qubit EstimatorQNN intrusion detector with direct input-gradient attacks;
- ii. a PGD-aware mixed adversarial-training protocol;
- iii. ten matched-seed evaluation under FGSM and PGD at three perturbation budgets;
- iv. attack-bound, domain, clipping, and attack-success-rate audits;
- v. paired statistical testing with effect sizes and multiplicity correction; and
- vi. a matched clean/robust MLP control that prevents quantum-specific overinterpretation.

Research context and motivation. Network intrusion detection has evolved from signature matching toward statistical learning and deep representation learning because contemporary traffic contains high-dimensional, heterogeneous, and rapidly changing behavior. Yet the security value of a detector cannot be inferred from clean-test accuracy alone. An attacker who understands the model or can approximate its decision surface may deliberately perturb admissible input features to alter the predicted class. This distinction is particularly important for learning-based intrusion detection, where a small reduction in recall on malicious traffic can translate into a large operational exposure. The adversarial-machine-learning literature therefore treats robustness as a separate property from ordinary generalization and recommends evaluation under explicit threat models rather than assuming that a well-performing clean classifier is secure (Goodfellow et al., 2015; Madry et al., 2018; Carlini & Wagner, 2017; Tramèr et al., 2020).

Why quantum supervised learning requires robustness evaluation. Quantum machine learning offers a different representation and optimization paradigm in which classical features are embedded into quantum states and processed by parameterized quantum circuits. Quantum feature maps can induce nonlinear decision boundaries in Hilbert space, while variational circuits provide trainable hybrid quantum-classical models (Havlíček et al., 2019; Schuld et al., 2020; Cerezo et al., 2021). These properties motivate investigation of QML for cybersecurity, but they do not imply either predictive superiority or adversarial robustness. Quantum classifiers have themselves been shown to admit adversarial examples, and both empirical and theoretical studies now examine their vulnerability, robustness bounds, noise effects, randomized encodings, and adversarial training (Lu et al., 2020; Weber et al., 2021; Ren et al., 2022; West et al., 2023; Winderl et al., 2024; Gong et al., 2024).

Problem statement. Existing QML-based intrusion-detection studies commonly emphasize clean classification metrics, small proof-of-concept subsets, or simulator-based feasibility. Fewer studies combine a direct white-box attack against the quantum model, iterative PGD evaluation, adversarial training, repeated random seeds, attack-constraint auditing, corrected statistical tests, and a matched adversarially trained classical neural baseline. Consequently, a reported improvement in a quantum detector may reflect data reduction, optimization, or evaluation choices rather than a robust property of the quantum model. A controlled experiment is therefore required to separate three questions: whether the clean QSL model is vulnerable; whether adversarial training improves the same QSL architecture; and whether any robustness gain remains competitive when the identical reduced feature representation is given to a classical neural control.

Research questions. This study is organized around four research questions.

RQ1 asks how strongly a four-qubit QSL intrusion detector degrades under bounded white-box FGSM and PGD perturbations.

RQ2 asks whether PGD-aware mixed adversarial training improves robustness while preserving acceptable clean performance.

RQ3 asks whether the observed improvements are stable across ten matched random seeds and remain supported by paired statistical tests, effect sizes, confidence intervals, and Holm correction.

RQ4 asks how the robust QSL detector compares with clean and adversarially trained MLP baselines under the same four-dimensional $[0, \pi]$ representation and attack budgets.

Contributions. The principal contributions are:

- (1) a reproducible four-qubit EstimatorQNN intrusion-detection pipeline using ZZFeatureMap and RealAmplitudes;
- (2) genuine white-box FGSM and iterative PGD attacks generated through input gradients of the quantum model;
- (3) PGD-aware adversarial training with a 50% clean, 25% FGSM, and 25% PGD mixture;
- (4) a ten-seed matched evaluation with perturbation audits, attack-success analysis, 95% confidence intervals, paired t-tests, Wilcoxon signed-rank tests, Cohen's d_z , and Holm-adjusted p-values; and
- (5) a rigorous classical control consisting of clean and adversarially trained MLPs evaluated on the same reduced representation. The implementation notebook and supporting experimental outputs are archived for reproducibility (Bhoge et al. 2026).

2. Related Work

The expanded literature base used in this revision covers four complementary areas:

- i. Foundational quantum machine learning and variational circuits (Biamonte et al., 2017; Schuld et al., 2020; Preskill, 2018; Havlíček et al., 2019; Schuld & Petruccione, 2018; Cerezo et al., 2021; Benedetti et al., 2019; McClean et al., 2018; Sim et al., 2019; Pérez-Salinas et al., 2020; Abbas et al., 2021; Larocca et al., 2023),
- ii. Quantum learning for intrusion detection and cybersecurity (Kaissar et al., 2026; Payares & Martinez-Santos, 2021; Abreu et al., 2024; Cirillo et al., 2026; Kim & Madhavi, 2024; Küçükkara et al., 2024; Aparcana-Tasayco et al., 2025; Sai et al., 2025; Abreu et al., 2025; S.S., 2025; Cirillo et al., 2026),
- iii. Adversarial machine learning and robust optimization (Goodfellow et al., 2015; Madry et al., 2018; Szegedy et al., 2014; Kurakin et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Tramèr et al., 2020; Croce & Hein, 2020), and
- iv. Adversarial robustness of quantum classifiers, including empirical, certified, noise-aware, kernel-based, and generalization-oriented studies (Lu et al., 2020; Ren et al., 2022; Winderl et al., 2024; Montalbano & Banchi, 2025; Kustiawan & Ghauth, 2026; Weber et al., 2021; Anil et al., 2024; West et al., 2023; Gong et al., 2024; Berberich et al., 2024; Huang et al., 2022; Georgiou et al., 2025; Dowling et al., 2026).

This broader positioning is important because high clean accuracy alone does not establish security: a detector intended for adversarial environments must also be evaluated under explicit threat models and matched classical controls.

2.1 Quantum machine learning for intrusion detection

QML-based intrusion detection has been explored using quantum support vector machines, variational classifiers, quantum neural networks, and hybrid architectures. Payares and Martinez-Santos compared several quantum models for DDoS detection (Payares & Martinez-Santos 2021). More recently, Kukliansky et al. developed QNN architectures for network anomaly detection and reported a small hardware implementation on IonQ Aria-1 using NF-UNSW-NB15 (Kukliansky et al. 2024). A comprehensive 2026 review of 25 QML-IDS studies concludes that hybrid quantum-classical methods dominate current work, while simulator dependence, hardware constraints, and inconsistent evaluation remain major limitations (Kaissar et al. 2026).

2.2 Adversarial robustness in quantum learning

Adversarial examples are a well-established problem in classical learning (Goodfellow et al., 2015; Madry et al., 2018). Quantum classifiers exhibit analogous vulnerabilities. Lu et al. (2020). Ren et al. experimentally demonstrated adversarial learning on programmable superconducting qubits and showed that adversarial training can substantially improve robustness (Ren et al., 2022). Winderl et al.

evaluated white-box attacks and defenses for QNNs under depolarization noise and reported that noise does not universally confer robustness (Winderl et al., 2024). Montalbano and Banchi extended adversarial analysis to quantum kernel methods and demonstrated both vulnerability and defense through adversarial data augmentation (Montalbano & Banchi, 2025). Recent reviews show that QML robustness research is still concentrated on input-level evasion attacks and empirical robust accuracy, with limited standardized benchmarking (Kustiawan & Ghauth, 2026).

2.2.1 Variational quantum classifiers and feature-space learning

Variational quantum classifiers combine a data-encoding circuit with trainable parameterized gates and a classical optimization loop. Circuit-centric classifiers and related variational models are attractive in near-term research because they can be implemented with shallow circuits and hybrid optimization (Schuld et al., 2020; Cerezo et al., 2021). However, trainability is not guaranteed. Barren plateaus can suppress gradients as circuit size and expressibility increase, and circuit architecture therefore affects both optimization and generalization (McClean et al., 2018; Sim et al., 2019; Larocca et al., 2023). The present work deliberately restricts the quantum model to four input dimensions and a compact ansatz, prioritizing controlled adversarial analysis over claims of large-scale quantum advantage.

Quantum feature maps are another important component because the security behavior of a classifier is partly determined by how classical perturbations propagate through the encoding (Havlíček et al. 2019). How quantum-enhanced feature spaces can be used for supervised learning, while data re-uploading provides an alternative mechanism for increasing representational capacity (Pérez-Salinas et al., 2020). In this study, the ZZFeatureMap is retained across clean and robust quantum experiments so that robustness differences can be attributed to training rather than a change of encoding.

2.2.2 QML for intrusion detection and cybersecurity

Research applying QML to intrusion detection has grown from early comparative demonstrations toward more systematic benchmarking. Payares and Martinez-Santos explored quantum machine learning for DDoS intrusion detection (Payares and Martinez-Santos 2021), while Abreu et al. proposed QML-IDS as a quantum-machine-learning intrusion-detection framework (Abreu et al. 2024). More recent work has examined quantum and hybrid network-security models, platform-independent DDoS classification, IoT anomaly detection, and noisy-device benchmarking (Kim & Madhavi, 2024; Küçükkara et al., 2024; Aparcana-Tasayco et al., 2025; Abreu et al., 2025; Cirillo et al., 2026). Reviews of QML for cybersecurity similarly emphasize the need to distinguish promising proof-of-concept accuracy from deployable security performance (Kaissar et al., 2026; Sai et al., 2025).

A recurring limitation is that many experiments compress high-dimensional network data to a small number of quantum features and execute on ideal simulators. This is a practical necessity for current research but changes the interpretation of the result: the experiment evaluates a hybrid reduced-

dimensional model, not a full-scale replacement for classical network monitoring. Preskill's NISQ perspective and later variational-algorithm literature underline the importance of circuit depth, noise, trainability, and hardware constraints (Preskill, 2018; Cerezo et al., 2021). The current study therefore reports the simulator setting explicitly and treats hardware execution and larger feature spaces as future validation rather than evidence already established.

2.2.3 Adversarial machine learning and robust optimization

Adversarial examples expose the sensitivity of learned decision functions to deliberately constructed perturbations. FGSM uses the sign of the input gradient to produce a one-step perturbation, whereas PGD iteratively updates the input and projects it back into the permitted perturbation set (Goodfellow et al., 2015; Madry et al., 2018). Strong evaluation requires more than one attack because weak or poorly configured attacks can create a false impression of robustness (Carlini and Wagner, 2017; Tramèr et al., 2020; Croce and Hein, 2020) therefore motivate careful threat-model specification, iterative attacks, adaptive evaluation, and attack auditing.

Adversarial training remains one of the most established empirical defenses: the training distribution is augmented with adversarially perturbed samples so that the model optimizes performance over a neighborhood of the observed data. The defense often introduces a clean-accuracy/robustness trade-off, making it essential to report both clean and attacked metrics. The present experiment adopts this principle but does not assume that adversarial training makes a model universally robust; its conclusions are restricted to the specified L-infinity budgets, attacks, feature representation, and binary UNSW-NB15 task.

2.2.4 Adversarial robustness of quantum classifiers

Quantum adversarial machine learning has developed into a distinct research area. Lu et al. demonstrated that quantum classifiers can be vulnerable to adversarial examples (Lu et al. 2020), while Ren et al. experimentally studied quantum adversarial learning on programmable superconducting qubits (Ren et al. 2022). Theoretical work has examined provable robustness through quantum hypothesis testing (Weber et al., 2021), and later studies have investigated quantum-enhanced robustness, randomized encodings, robust training, universal perturbations, kernel methods, and the interaction between noise and attacks (West et al., 2023; Anil et al., 2024; Gong et al., 2024; Berberich et al., 2024; Winderl et al., 2024; Montalbano & Banchi, 2025).

These studies show that robustness is architecture- and setting-dependent. Noise may sometimes alter sensitivity, but it is not a substitute for a defense; randomized encodings can modify the attack surface; and adversarial training can improve empirical robustness without establishing a universal guarantee. Recent reviews therefore call for stronger baselines, clearer threat models, repeated trials, and reproducible attack settings (Kustiawan & Ghauth, 2026; Razavi-Far et al., 2026). The current work

follows that direction by comparing the robust QSL model against an adversarially trained classical MLP rather than evaluating the quantum model in isolation.

The literature synthesis and the specific positioning of the present study are summarized in Table 1.

Table 1. Literature analysis and relevance of the present study

Study	Primary focus	Adversarial evaluation	Classical control	Relevance / remaining gap
Lu et al. (2020)	Quantum adversarial ML	Yes	Not IDS-specific	Establishes vulnerability of quantum classifiers
Ren et al. (2022)	Experimental quantum adversarial learning	Yes	Task-specific	Hardware evidence; not network IDS
West et al. (2023)	Quantum adversarial robustness	Yes	Comparative	Motivates robustness analysis beyond clean accuracy
Winderl et al. (2024)	QNN attacks/defenses under noise	White-box	Limited	Highlights noise/attack interaction
Abreu et al. (2024)	QML intrusion detection	No central adversarial study	Yes	QML-IDS feasibility; robustness remains open
Montalbano & Banchi (2025)	Quantum kernel adversarial learning	Yes	Kernel comparison	Different QML family from EstimatorQNN
Cirillo et al. (2026)	QML IDS benchmarking	Noise-oriented	Yes	Supports realistic benchmarking and controls
Kustiawan & Ghauth (2026)	QML adversarial robustness review	Review	Review	Calls for standardized robust evaluation
Present study	4-qubit QSL IDS + robust MLP	FGSM + iterative PGD	Clean/robust MLP	10 seeds, audits, CIs, effect sizes, Holm correction

Evaluation design in QML-IDS research. A central methodological issue in QML-IDS is the separation between feasibility demonstrations and security evaluations. A feasibility experiment asks whether a quantum or hybrid learner can classify a reduced intrusion dataset under a chosen simulator or device. A security evaluation asks a stricter question: whether the learned decision rule remains reliable when inputs are deliberately manipulated under an explicit adversary model. The second question requires attack parameters, constraints, success criteria, and baselines to be specified with the same care as the

classifier itself. This distinction is well established in classical adversarial learning, where apparent robustness can disappear when the attack is strengthened or adapted (Carlini & Wagner, 2017; Tramèr et al., 2020; Croce & Hein, 2020). For QML-IDS, the distinction is especially important because computational constraints often encourage small samples, dimensionality reduction, shallow circuits, and limited repetitions. Each choice may be reasonable, but the resulting evidence must be interpreted within those boundaries rather than generalized to unrestricted network traffic.

Role of repeated trials and statistical inference. Quantum and hybrid learning experiments can exhibit variability from data sampling, parameter initialization, stochastic optimization, and finite-shot or hardware effects. Even in deterministic statevector-style simulation, the classical optimization path and sampled training subset can materially affect the final classifier. Reporting a single run therefore risks confusing an initialization-specific outcome with a reproducible model property. Repeated-seed evaluation is a practical way to expose this variability. The present study uses ten matched seeds and retains the pairing when comparing clean and robust conditions. Confidence intervals summarize uncertainty around mean performance, paired tests assess whether within-seed differences are systematically directional, and effect sizes quantify practical magnitude. Holm correction is used because multiple hypotheses are examined. This statistical structure responds to broader calls for stronger empirical methodology in QML benchmarking and cybersecurity evaluation (Kaissar et al., 2026; Cirillo et al., 2026).

Classical controls and the interpretation of quantum results. A classical control is indispensable when a quantum model is trained on aggressively compressed data. Without such a control, an observed result cannot distinguish the contribution of the quantum circuit from the contribution of preprocessing, feature selection, or the intrinsic separability of the reduced dataset. The appropriate baseline should therefore receive the same transformed information and be evaluated under the same train/test partitions and attacks. This study uses a compact MLP rather than an oversized deep network so that the comparison remains focused on learning behavior in the common four-dimensional representation. Parameter-count equivalence is not claimed, because equal parameter counts do not imply equal functional capacity across model families. Instead, the baseline addresses the operational question of whether a conventional neural learner can obtain stronger clean or adversarial performance from the same information. This is consistent with cautious QML benchmarking practice, which treats classical comparison as a prerequisite for interpreting quantum utility (Biamonte et al., 2017; Cerezo et al., 2021; Cirillo et al., 2026).

Robustness, noise, and hardware realism. Noise and adversarial robustness are related but distinct. Device noise can alter gradients and decision surfaces, and some theoretical or empirical studies investigate whether stochasticity can affect adversarial sensitivity. Nevertheless, uncontrolled noise cannot be assumed to provide security because it may simultaneously degrade clean accuracy, create unstable behavior, or be incorporated into an adaptive attacker model. Winderl et al. explicitly examine white-box attacks and defenses under depolarization noise (Winderl et al. 2024), while later work considers formal robustness perspectives (Weber et al. 2021). The current study intentionally isolates

the adversarial-training effect in a simulator setting before adding hardware noise as another variable. This sequencing makes the result easier to interpret: any improvement from clean QSL to PGD-aware QSL occurs without changing the physical noise model. No claim is made that the simulator result predicts identical behavior on a real QPU.

Position of the present work. Taken together, the literature suggests that a useful adversarial-QML IDS study should combine a clearly bounded input space, a direct attack on the evaluated model, iterative attack evaluation, a defense that is trained rather than postulated, repeated trials, and a meaningful classical control. The present experiment is designed around these requirements. It does not attempt to maximize qubit count or claim that a four-qubit classifier represents production-scale network traffic. Instead, it uses the small circuit as a controlled laboratory for testing whether adversarial training can improve a quantum intrusion detector and whether that improvement survives comparison with a classical learner. This narrower claim is more verifiable and more useful for subsequent scaling studies than a broad assertion of quantum advantage.

2.3 Research gap

Existing QML-IDS studies primarily emphasize clean predictive performance, whereas adversarial-QML studies often use image or synthetic classification tasks rather than network traffic. This leaves a practical gap at the intersection of intrusion detection, direct white-box quantum gradients, iterative adversarial attacks, repeated-seed statistics, attack auditing, and matched classical controls. The present study is designed specifically around this gap.

3. Materials and Methods

3.1 Dataset and task

The binary intrusion-detection task uses the standard UNSW-NB15 training and testing CSV partitions. The source dataset contains contemporary normal activity and nine attack families; the official partition contains 175,341 training and 82,332 testing records (Moustafa & Slay, 2015). To keep repeated quantum simulation computationally tractable while preserving matched comparisons, each experimental seed uses a deterministic stratified subset of 800 training, 200 validation, and 300 test records. The same seed-specific subsets are reused for QSL and MLP experiments.

3.2 Preprocessing

The identifier and attack-category fields are removed for binary classification. Numerical attributes are standardized, categorical attributes are one-hot encoded, and transformed values are reduced with principal component analysis to four components. A MinMax transformation then maps the four components to $[0, \pi]$, matching the input domain used by the quantum feature map. Transformation

objects are fitted only on the training split and applied to validation and test data. The complete matched preprocessing, model-training, attack, and evaluation workflow is shown in Fig. 1.

3.2.1 Rationale for dimensionality reduction and angle scaling

The original UNSW-NB15 feature space is substantially larger than the number of qubits used in this experiment. Directly assigning one feature to one qubit would therefore exceed the intended compact circuit and make repeated gradient-based adversarial evaluation computationally expensive. PCA is used as a controlled dimensionality-reduction step rather than as evidence of quantum feature selection. The transformation is fitted only on the training split and then applied unchanged to validation and test samples. This preserves the experimental separation between model development and final evaluation. Four principal components are retained because the quantum branch uses four qubits; the same four values are supplied to the MLP control so that both model families operate on identical information.

After PCA, MinMax scaling maps each retained component into $[0, \pi]$. The interval is chosen to match the angular parameterization used by the quantum feature map and to create a finite domain for adversarial constraints. This scaling also makes the attack budgets interpretable as fixed fractions of π . An epsilon of 0.02π is a relatively small displacement in the transformed domain, 0.05π represents a moderate perturbation, and 0.10π is deliberately stronger. These values should not be interpreted as direct physical changes to raw packet fields because the attack is performed after PCA and scaling. The experiment therefore measures robustness in the learned four-dimensional representation, not semantic feasibility of every perturbation in the original network-feature space.

This distinction matters for external validity. Feature-space attacks can reveal model sensitivity and provide a controlled comparison of defenses, but a deployment attacker may be constrained by protocol semantics, inter-feature dependencies, and the requirement that manipulated traffic remain executable. Future work should therefore connect transformed-space robustness with problem-space attacks that preserve valid network behavior. The present study treats transformed-space FGSM/PGD as a rigorous first stage because the perturbation set is explicit, differentiable, reproducible, and shared by the quantum and classical models.

3.3 Four-qubit QSL model

The quantum classifier is implemented as an EstimatorQNN. A four-qubit ZZFeatureMap with one repetition and linear entanglement encodes the four PCA components. It is composed with a RealAmplitudes variational ansatz using one repetition and linear entanglement. Input gradients are enabled so that adversarial examples are generated directly against the quantum model rather than a surrogate classifier. Qiskit Machine Learning explicitly supports backward computation of EstimatorQNN input gradients when `input_gradients = True` (Qiskit Machine Learning Contributors, 2026). The four-qubit feature-map and variational-circuit structure used by the QSL model is shown in Fig. 2.

3.3.1 Quantum circuit design choices and trainability

The QSL circuit separates data encoding from trainable processing. ZZFeatureMap introduces feature-dependent single-qubit rotations and pairwise interactions, creating a nonlinear embedding of the four PCA components. RealAmplitudes then supplies the variational layer whose parameters are optimized by the classical training loop. This separation is useful experimentally because the same encoding can be retained while comparing clean and adversarial training. A deeper or more expressive circuit might increase representational capacity, but it would also increase optimization cost and potentially aggravate barren-plateau behavior (McClean et al., 2018; Sim et al., 2019). The compact ansatz is therefore a deliberate compromise that makes ten-seed white-box evaluation feasible.

EstimatorQNN is used because it exposes a differentiable quantum-neural-network interface compatible with input gradients. This capability is essential to the threat model: FGSM and PGD must differentiate the loss with respect to the actual quantum model input rather than attack a substitute classical classifier. The resulting adversarial examples are therefore white-box examples for the evaluated QSL decision function. This is methodologically stronger than claiming quantum robustness from transfer attacks alone, although it remains specific to the simulator, circuit, preprocessing, and loss used in the experiment.

Trainability remains a key limitation of variational quantum learning. Parameterized quantum circuits may exhibit flat optimization landscapes, and increasing qubit count does not automatically improve useful capacity (McClean et al., 2018; Larocca et al., 2023). For this reason, the study does not extrapolate the four-qubit result to larger circuits. The contribution is instead to establish an adversarial evaluation protocol that can later be reused when scaling the representation, changing the ansatz, or moving to noisy hardware.

3.4 Classical MLP control

The matched classical control is a compact multilayer perceptron with four inputs, hidden layers of 16 and 8 ReLU units, and one binary logit output. Clean_MLP is trained on clean data only. Robust_MLP starts from the clean model and is adversarially trained with the same clean/FGSM/PGD proportions and perturbation domain used for QSL. This baseline is a methodological control; parameter-count equivalence with the QNN is not claimed.

3.5 Threat model and attacks

The adversary is white-box and untargeted: model structure, parameters, loss, and input gradients are assumed known. For input x , label y , model parameters θ , and loss L , FGSM is defined as

$$x_{adv} = \Pi_{[0, \pi]} \{ x + \epsilon \operatorname{sign}(\nabla_x L(\theta, x, y)) \}$$

PGD iterates the same loss-maximizing direction with step size α while projecting the perturbed sample into both the $L_\infty \epsilon$ -ball around the original input and the valid $[0, \pi]$ domain:

$$x^{(k+1)} = \Pi_{[0, \pi]} \cap B_\infty(x, \epsilon) \{ x^{(k)} + \alpha \text{sign}(\nabla_x L(\theta, x^{(k)}, y)) \}$$

2

The evaluation budgets are $\epsilon = 0.02\pi$, 0.05π , and 0.10π . PGD uses three steps during robust training and seven steps during final evaluation. Attack Success Rate (ASR) is computed only over inputs classified correctly before attack:

$$ASR = N(\text{correct} \rightarrow \text{wrong}) / N(\text{clean correct})$$

3

3.6 PGD-aware adversarial training

Each robust-training mini-batch is partitioned into approximately 50% clean, 25% FGSM, and 25% PGD samples. Adversarial training uses $\epsilon = 0.05\pi$. For the QSL model, the best checkpoint is selected by the mean validation F1 across clean, FGSM, and PGD conditions, with worst-case F1 as a tie-break. This criterion is intended to avoid selecting a model that preserves clean performance while collapsing under PGD.

3.7 Repeated-seed protocol and statistics

The primary experiment uses ten matched seeds: 11, 23, 37, 41, 53, 67, 71, 83, 97, and 101. For each model-condition pair, accuracy, precision, recall, F1, and ROC-AUC are reported as mean and standard deviation; 95% t-based confidence intervals are also computed. Matched model comparisons use paired t-tests, Wilcoxon signed-rank tests, and Cohen's d_z . Holm adjustment is applied to families of quantum-versus-classical comparisons. Because $n = 10$ remains modest, effect sizes and confidence intervals are interpreted alongside p-values rather than treating statistical significance as the sole criterion.

3.8 Implementation environment

The completed QSL experiment used Qiskit 2.5.2, Qiskit Aer 0.17.2, and Qiskit Machine Learning 0.9.1 in PAPER mode. The reported run used CPU statevector simulation; therefore, the experiment evaluates simulated quantum circuits and does not constitute execution on a physical QPU. This distinction is maintained throughout the paper for better interpretation.

3.9 Formal problem definition and optimization objectives

Let $x \in [0, \pi]^4$ denote the four-dimensional PCA-transformed and angle-scaled feature vector and $y \in \{0, 1\}$ the binary label. The quantum classifier implements a parameterized score $f_\theta(x)$ obtained from the expectation value produced by the EstimatorQNN. A probability-like output is used with a binary classification loss $L(f_\theta(x), y)$. Clean training seeks parameters θ that minimize the empirical risk over the

sampled training set. The same preprocessing transformation fitted on the training data is applied to validation and test data to prevent information leakage.

For white-box FGSM, the adversarial sample is formed as

$$x_{adv} = clip_{[0, \pi]}(x + \epsilon \text{ sign}(\nabla_x L(f_\theta(x), y)))$$

4

PGD extends this idea through repeated gradient-sign updates with projection onto the L_∞ ball B_ε(x) and the valid [0, π] feature domain. The evaluation PGD attack uses seven steps, whereas the adversarial-training PGD component uses three steps to control computational cost. The perturbation audit independently verifies that $\max|x_{adv} - x| \leq \epsilon$ and that every generated feature remains within the legal transformed domain.

PGD-aware training optimizes a mixed empirical objective constructed from 50% clean examples, 25% FGSM examples, and 25% PGD examples. This mixture was chosen to expose the model to both one-step and iterative perturbations without removing clean samples from the training distribution. The robust model is selected using validation behavior rather than test performance. The final test set is reserved for the clean and adversarial evaluation reported in the Results section.

3.10 Threat model, fairness controls, and reproducibility

The attacker is assumed to have white-box access to the evaluated model and its input gradients. This is a deliberately strong first-order threat model because it avoids the ambiguity of transfer-only attacks and directly tests the local sensitivity of each classifier. Perturbations are applied after the common preprocessing pipeline in the four-dimensional angle-scaled space. The budgets $\epsilon = 0.02\pi$, 0.05π , and 0.10π therefore have an explicit interpretation relative to the [0, π] feature domain.

Fairness between quantum and classical models is enforced at the representation and evaluation levels. Both model families receive the same four transformed features, use the same ten random seeds, and are evaluated at the same attack budgets. The comparison is not intended to equalize every parameter count or optimization detail; instead, it asks whether the proposed quantum defense remains competitive against a compact classical neural model trained on exactly the information available to the quantum model. This control is essential because superiority over an undefended baseline would not establish that the quantum architecture itself contributes to robustness.

Reproducibility is supported through deterministic seed control, per-seed outputs, checkpointing, cached preprocessing, attack diagnostics, statistical-result files, and an archived implementation package. The notebook and supporting experimental outputs are available through Zenodo (Bhoge et al., 2026). The reported GPU option, when used, accelerates classical simulation of quantum circuits and must not be interpreted as execution on a quantum processing unit. The controlled experimental configuration, including data splits, attack budgets, and matched evaluation settings, is summarized in Table 2.

Table 2. Controlled experimental configuration

Component	Setting	Purpose
Dataset	UNSW-NB15, binary label	Benchmark network intrusion detection
Quantum input	4 PCA components scaled to $[0, \pi]$	Match four-qubit circuit
Feature map	ZZFeatureMap	Nonlinear quantum encoding
Ansatz	RealAmplitudes	Trainable variational circuit
Quantum learner	EstimatorQNN / VQC-style classifier	Input-gradient white-box evaluation
Classical control	Compact MLP	Matched reduced-feature baseline
Seeds	11, 23, 37, 41, 53, 67, 71, 83, 97, 101	Repeated-trial stability
Attack budgets	0.02π , 0.05π , 0.10π	Increasing Loo perturbation strength
Evaluation attacks	FGSM and 7-step PGD	One-step and iterative white-box attacks
Robust training	50% clean + 25% FGSM + 25% 3-step PGD	PGD-aware mixed adversarial training
Statistics	95% CI, paired t, Wilcoxon, Cohen's dz, Holm correction	Magnitude and significance of paired differences

3.11 Evaluation metrics and interpretation

Accuracy measures the overall fraction of correct predictions, while precision and recall distinguish false-alarm behavior from missed attacks. F1 provides their harmonic balance, and ROC-AUC summarizes ranking behavior across thresholds. In an intrusion-detection context, recall is particularly important because false negatives correspond to malicious flows that evade detection; however, a model with high recall and excessive false positives may still be operationally unsuitable. For this reason, the study reports a set of complementary metrics rather than selecting a single favorable measure.

Attack success rate (ASR) is computed relative to samples that were classified correctly before attack. This denominator prevents already-wrong clean predictions from being counted as successful adversarial manipulations. The transition audit additionally records correct-to-wrong, correct-to-correct, wrong-to-correct, and wrong-to-wrong movements. Perturbation diagnostics report the achieved L-infinity magnitude, the achieved-to-requested epsilon ratio, and boundary clipping. These quantities help determine whether a low attack success rate reflects genuine model resistance or simply an attack that failed to use the allowed budget.

The statistical analysis is intentionally paired. For each seed and condition, the clean-trained and robust-trained models are compared using the corresponding seed result. The paired t-test evaluates the mean paired difference, whereas the Wilcoxon signed-rank test reduces dependence on normality assumptions. Cohen's d_z standardizes the paired difference by the standard deviation of within-seed differences. Because several attack/metric comparisons are tested, Holm-adjusted p-values are reported alongside raw p-values. Statistical significance is not treated as a substitute for effect magnitude; the discussion considers both.

4. Results

4.1 Overall clean and adversarial accuracy

The ten-seed clean and adversarial accuracy results for all four model conditions are summarized in Table 3.

Table 3. Accuracy (%) across ten matched seeds (mean $\pm$ SD).

Condition	Clean QSL	PGD-aware QSL	Clean MLP	Robust MLP
CLEAN	70.73 $\pm$ 2.98	61.57 $\pm$ 11.21	69.50 $\pm$ 3.23	67.33 $\pm$ 2.15
FGSM 0.02 π	43.53 $\pm$ 11.06	48.23 $\pm$ 11.16	60.97 $\pm$ 6.25	66.53 $\pm$ 2.31
FGSM 0.05 π	31.03 $\pm$ 9.44	45.13 $\pm$ 6.20	47.27 $\pm$ 16.49	64.73 $\pm$ 3.05
FGSM 0.10 π	44.47 $\pm$ 9.08	49.10 $\pm$ 3.51	37.47 $\pm$ 22.61	54.43 $\pm$ 9.79
PGD 0.02 π	41.83 $\pm$ 11.74	46.33 $\pm$ 11.49	61.70 $\pm$ 5.68	66.70 $\pm$ 2.27
PGD 0.05 π	9.10 $\pm$ 5.79	26.90 $\pm$ 15.06	48.13 $\pm$ 15.52	64.27 $\pm$ 3.60
PGD 0.10 π	2.60 $\pm$ 1.22	20.73 $\pm$ 18.55	37.93 $\pm$ 22.24	50.50 $\pm$ 12.57

Clean QSL achieved 70.73% mean accuracy, while PGD-aware QSL achieved 61.57%, a reduction of 9.17 percentage points. Under stronger attacks, however, adversarial training materially improved the quantum model. At FGSM 0.05 π , accuracy increased from 31.03% to 45.13%. At PGD 0.05 π , it increased from 9.10% to 26.90%, and at PGD 0.10 π from 2.60% to 20.73%. Thus, the defense produces a clear clean-accuracy versus robust-accuracy trade-off. The FGSM accuracy trends across the three perturbation budgets are shown in Figure 3.

The corresponding PGD accuracy trends, which expose the strongest degradation of the clean-trained QSL model, are shown in Figure 4.

4.2 Statistical evidence for PGD-aware QSL

The selected paired statistical comparisons between clean-trained and PGD-aware QSL are reported in Table 4.

Table 4. Selected paired comparisons between clean and PGD-aware QSL.

Condition	Clean QSL	Robust QSL	Δ pp	paired t p	Wilcoxon p	d_z
CLEAN	70.73	61.57	-9.17	0.0222	0.0273	-0.87
FGSM 0.05 π	31.03	45.13	+14.10	0.0117	0.0137	1.00
PGD 0.05 π	9.10	26.90	+17.80	0.0070	0.0137	1.10
PGD 0.10 π	2.60	20.73	+18.13	0.0146	0.0215	0.95

The largest robustness gains occur under moderate and strong PGD. At PGD 0.05 π , the 17.80-point accuracy increase has paired t p = 0.0070, Wilcoxon p = 0.0137, and d_z = 1.10. At PGD 0.10 π , the 18.13-point increase has paired t p = 0.0146 and d_z = 0.95. FGSM 0.05 π also shows a 14.10-point gain with d_z $\approx$ 1.00. Conversely, the clean-accuracy loss is statistically detectable (p = 0.0222, d_z = -0.87).

4.3 Attack success rate and perturbation audits

Attack success rates and perturbation-constraint diagnostics for the quantum models are summarized in Table 5.

Table 5. Selected quantum attack diagnostics.

Model	Attack	ASR (%)	Achieved ϵ ratio	Boundary-touch samples (%)
Clean_QSL	FGSM 0.05 π	61.43	1.000	1.53
V3_PGD_Aware_QSL	FGSM 0.05 π	48.09	1.000	1.50
Clean_QSL	PGD 0.05 π	87.12	0.986	1.30
V3_PGD_Aware_QSL	PGD 0.05 π	69.78	0.979	1.13
Clean_QSL	PGD 0.10 π	96.70	0.924	3.33
V3_PGD_Aware_QSL	PGD 0.10 π	78.61	0.920	2.90

PGD-aware training reduces ASR most clearly at stronger PGD budgets: from 87.12% to 69.78% at 0.05 π and from 96.70% to 78.61% at 0.10 π . The attack audit confirms that all generated adversarial samples remain within the requested L_∞ bounds and $[0,\pi]$ feature domain. PGD uses approximately 97.9% of the allowed budget at 0.05 π and 92.0% at 0.10 π for the robust QSL model, while boundary-touch rates remain low. These checks reduce the likelihood that apparent robustness is caused by an incorrectly implemented or heavily clipped attack.

4.4 Matched classical comparison

The classical control changes the interpretation of the quantum results. Robust_MLP preserves 67.33% clean accuracy and reaches 64.27% at PGD 0.05π and 50.50% at PGD 0.10π , substantially exceeding the corresponding PGD-aware QSL accuracies of 26.90% and 20.73%. The result therefore supports the effectiveness of adversarial training for QSL but not a claim of superior absolute robustness. The matched accuracy and statistical comparison between PGD-aware QSL and the robust classical MLP is reported in Table 6.

Table 6. Accuracy comparison between PGD-aware QSL and Robust MLP.

Condition	Robust QSL	Robust MLP	Δ pp	d_z	Holm t p	Holm W p
CLEAN	61.57	67.33	-5.77	-0.56	1.0000	1.0000
FGSM 0.02π	48.23	66.53	-18.30	-1.63	0.0183	0.0820
FGSM 0.05π	45.13	64.73	-19.60	-2.34	0.0017	0.0820
PGD 0.02π	46.33	66.70	-20.37	-1.77	0.0110	0.0820
PGD 0.05π	26.90	64.27	-37.37	-2.03	0.0043	0.0820
PGD 0.10π	20.73	50.50	-29.77	-1.04	0.1699	0.1348

After Holm correction, paired t-tests still indicate differences favoring Robust_MLP at FGSM 0.02π , FGSM 0.05π , PGD 0.02π , and PGD 0.05π . Holm-adjusted Wilcoxon tests are more conservative and do not cross 0.05 in these comparisons. The standardized paired differences are nevertheless large at several budgets, including $d_z = -2.34$ at FGSM 0.05π and $d_z = -2.03$ at PGD 0.05π . This mixed inferential pattern is reported transparently rather than selecting only favorable tests.

4.5 Robustness-accuracy trade-off across model families

The ten-seed aggregate results reveal two distinct effects. First, PGD-aware training improves the attacked performance of the QSL model relative to its clean-trained counterpart, particularly under the stronger PGD conditions. Second, the robust classical MLP retains markedly higher absolute attacked accuracy than the robust QSL model. These findings should be interpreted together: adversarial training is effective for the tested quantum architecture, but the experiment does not support a quantum-advantage claim.

The clean-accuracy cost is also material. The clean QSL model achieves 70.73% mean accuracy, whereas the PGD-aware QSL model achieves 61.57%, a reduction of 9.16 percentage points. By comparison, the clean MLP achieves 69.50% and the robust MLP 67.33%, a smaller 2.17-point reduction. Under PGD at 0.05π , however, the robust QSL improves from 9.10% to 26.90%, while the robust MLP

reaches 64.27%. At PGD 0.10π , the corresponding values are 2.60%, 20.73%, and 50.50%. The defense therefore shifts the QSL decision boundary in a useful direction but does not close the classical robustness gap.

These results also illustrate why a clean-only comparison would be misleading. Clean QSL and clean MLP accuracy are similar in magnitude, yet their adversarial degradation patterns differ substantially. Robustness must therefore be treated as an independent evaluation axis. Reporting only the clean result could suggest parity between the model families while concealing a large difference under iterative white-box attack.

4.6 Interpretation of statistical and attack-audit evidence

Repeated seeds reduce the risk that a single favorable initialization determines the conclusion. The paired design is especially appropriate because each quantum/classical condition is evaluated under the same seed index and attack budget. Paired t-tests quantify mean differences under approximate parametric assumptions, while Wilcoxon signed-rank tests provide a nonparametric counterpart. Cohen's d_z reports standardized paired effect magnitude, and Holm adjustment limits family-wise error across multiple comparisons. The manuscript therefore emphasizes effects that are supported by magnitude, direction, and corrected inference rather than isolated unadjusted p-values.

The attack audits provide a second layer of validity. A robustness result is only meaningful if the generated examples satisfy the intended perturbation constraint. The recorded diagnostics verify the Loo bound and the $[0, \pi]$ domain, and attack-success statistics quantify transitions from initially correct to adversarially incorrect predictions. Together, these checks reduce the risk that apparent robustness is caused by a malformed attack, invalid transformed features, or inconsistent clipping.

5. Discussion

The interpretation is also consistent with the wider QAML literature: quantum classifiers can be vulnerable to carefully bounded perturbations (Lu et al., 2020; Ren et al., 2022), robustness can depend strongly on encoding, circuit structure, noise and training strategy (Winderl et al., 2024; Weber et al., 2021; West et al., 2023; Gong et al., 2024; Berberich et al., 2024; Huang et al., 2022; Georgiou et al., 2025; Dowling et al., 2026), and classical adversarial-training literature warns against evaluating a defense with only one weak attack or only clean accuracy (Goodfellow et al., 2015; Madry et al., 2018; Szegedy et al., 2014; Kurakin et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Tramèr et al., 2020; Croce & Hein, 2020). Accordingly, the present study treats the classical MLP control, ASR, perturbation audits, repeated seeds and corrected statistical tests as integral parts of the evidence rather than auxiliary analyses.

The central finding is not that the four-qubit QSL model is adversarially robust in an absolute sense. Instead, the evidence shows that targeted adversarial training changes its failure profile in a measurable

way. The clean QSL model collapses under moderate and strong PGD, whereas PGD-aware training recovers 17.8–18.1 percentage points of accuracy at 0.05π – 0.10π and substantially lowers PGD attack success. These gains are consistent with the broader adversarial-learning literature, in which robustness is obtained by explicitly optimizing against first-order adversaries (Madry et al., 2018), and with quantum studies showing that adversarial training can improve quantum classifiers (Ren et al., 2022; Montalbano & Banchi, 2025).

The clean-performance penalty is equally important. Robust QSL loses approximately 9.17 percentage points of clean accuracy, illustrating a robustness-accuracy trade-off. This is consistent with the need to treat robust performance as a multi-objective problem rather than optimize only clean accuracy. The validation criterion used here therefore combines clean, FGSM, and PGD F1 and includes worst-case F1 as a tie-break.

The matched MLP baseline provides a necessary reality check. Under the same compact four-dimensional representation and attack budgets, Robust_MLP is markedly stronger under PGD and usually stronger under FGSM. This means that the present evidence cannot support statements that quantum representation or variational circuits inherently confer practical adversarial superiority. Such restraint is especially important because current QML-IDS literature remains dominated by simulation and small-scale representations (Kaissar et al., 2026). Recent QML-IDS studies and surveys likewise emphasize controlled classical comparison, simulator limitations, and the need for realistic benchmarking (Kaissar et al., 2026; Abreu et al., 2024; Cirillo et al., 2026; Kim & Madhavi, 2024; Küçükkara et al., 2024; Aparcana-Tasayco et al., 2025; Sai et al., 2025; Abreu et al., 2025; S.S., 2025; Cirillo et al., 2026).

The non-monotonic FGSM response of the QSL models also deserves attention. Accuracy does not decrease monotonically with epsilon, whereas PGD exhibits the expected stronger degradation. Because the audit confirms that FGSM reaches the full requested Loo budget, the pattern is unlikely to be caused by an epsilon implementation failure. A plausible explanation is the nonlinear geometry of the learned decision regions, but this remains an empirical observation rather than a demonstrated mechanism.

5.1 Engineering implications

For near-term intrusion-detection research, the results favor selective use of quantum models rather than replacement of strong classical learners. A practical architecture could route uncertain, novel, or high-risk traffic to a quantum branch while retaining classical models for routine traffic. The present experiment does not validate such a deployment architecture, but the comparative result motivates it as future engineering work.

5.1.1 Why adversarial training helps QSL but does not yield quantum advantage

The improvement of the PGD-aware QSL model is consistent with robust-optimization intuition: exposing the learner to adversarial neighborhoods during training can reduce local sensitivity around observed samples (Madry et al., 2018). In the quantum setting, the defense changes the learned variational parameters while leaving the feature map and circuit family fixed. Consequently, the gain is attributable to the training objective rather than to a larger circuit or a different encoding. This controlled design strengthens the causal interpretation of the within-QSL comparison.

However, the robust MLP result demonstrates that the same four-dimensional representation contains enough information for a classical network to achieve a substantially stronger robustness-accuracy trade-off. The experiment therefore provides no evidence that Hilbert-space embedding alone confers adversarial protection. This interpretation is aligned with quantum-adversarial literature showing that quantum models can remain vulnerable and that robustness depends on architecture, training, encoding, noise, and threat model rather than on 'quantumness' as a generic property (Lu et al., 2020; West et al., 2023; Winderl et al., 2024; Kustiawan & Ghauth, 2026).

The negative comparative finding is scientifically useful. A robust classical control prevents overclaiming and identifies where further quantum research is required. Future QSL designs would need to demonstrate either improved absolute robustness, a favorable resource/accuracy trade-off, or a hardware-relevant advantage under a controlled benchmark. Until such evidence is obtained, the present contribution should be viewed as an adversarial-robustness characterization and defense study rather than a demonstration of quantum superiority.

5.1.2 Implications for practical intrusion-detection research

For practical IDS research, the results support a staged evaluation philosophy. A candidate quantum detector should first be compared with a strong classical model on the same information representation; second, it should be evaluated under attacks that directly target its gradients or decision function; third, the attack implementation should be audited; and fourth, repeated trials and effect sizes should accompany average accuracy. This sequence is more informative than comparing a quantum model with an arbitrarily weak classical baseline.

The four-feature compression used here is both a strength and a limitation. It enables tractable repeated quantum-gradient experiments and isolates the robustness question, but it also discards information present in the original UNSW-NB15 feature space. The classical MLP's strong robust performance on the same compressed input indicates that the comparison is not simply a consequence of the MLP having access to more raw features. Nevertheless, deployment-oriented studies should investigate whether the conclusions persist with larger quantum representations, alternative encodings, and additional datasets.

5.1.3 Relationship to prior adversarial-QML findings

The observed vulnerability of clean QSL is consistent with the broader finding that quantum classifiers are not inherently immune to adversarial perturbations. Lu et al. (2020) established the conceptual basis

for quantum adversarial machine learning, and Ren et al. (2022) demonstrated adversarial learning experimentally on superconducting hardware. The present IDS experiment extends that line of inquiry into a network-security setting with a direct gradient-based threat model. The severe degradation of clean QSL under stronger PGD reinforces the need to evaluate quantum cybersecurity models under attack rather than infer robustness from quantum state-space dimensionality.

The benefit from PGD-aware training is also compatible with studies exploring defenses and robust quantum models (West et al., 2023; Berberich et al., 2024; Gong et al., 2024). However, the magnitude of the classical MLP advantage shows why defense success must be interpreted comparatively. A defense can be effective in the within-model sense and still be unattractive in an engineering sense if a simpler classical model offers a better robustness-accuracy trade-off. This dual interpretation is a central contribution of the study: it recognizes the quantum defense improvement without converting that improvement into an unsupported superiority claim.

The result further suggests that future quantum IDS research should report resource-aware utility. If a quantum model eventually matches or exceeds a classical robust baseline, the next question will be whether it does so with a meaningful benefit in data efficiency, model complexity, trainability, latency, energy, or hardware scaling. Conversely, if classical models remain stronger, quantum experiments can still be valuable for identifying circuit designs, encodings, or training regimes that close specific gaps. This perspective aligns with NISQ-era caution that practical quantum value must be demonstrated under realistic constraints rather than inferred from representational arguments alone (Preskill, 2018; Cerezo et al., 2021).

5.2 Limitations and threats to validity

First, the quantum model is simulated on CPU and not executed on physical quantum hardware; hardware noise, shot noise, queue latency, and compilation effects are therefore outside the present evidence. Second, the experiment uses four PCA components and stratified subsets (800/200/300 per seed) to make repeated quantum-gradient evaluation feasible; this is a controlled benchmark rather than full-scale network deployment. Third, only binary classification, FGSM, and PGD are evaluated. Adaptive attacks, black-box transfer attacks, poisoning, and quantum-native perturbations remain untested. Fourth, ten seeds improve reliability but still yield limited power for nonparametric multiplicity-corrected tests. Fifth, the MLP is matched on input representation and attack protocol rather than parameter count or computational cost. Finally, robustness on UNSW-NB15 does not establish cross-dataset generalization. These limitations align with broader concerns about NISQ-era scalability, trainability, hardware noise, and the gap between simulator evidence and deployment evidence (Preskill, 2018; McClean et al., 2018; Sim et al., 2019; Pérez-Salinas et al., 2020; Abbas et al., 2021; Larocca et al., 2023; Küçükkara et al., 2024; Aparcana-Tasayco et al., 2025).

5.3 Future research directions

Several extensions follow naturally from the present evidence. First, the same ten-seed threat model should be applied to trainable quantum kernels or QKA so that robustness can be compared across

fundamentally different QML families. Second, QCNN-style architectures and data re-uploading circuits could test whether inductive bias or increased representational capacity improves the robustness-accuracy trade-off. Third, adaptive attacks, transfer attacks, and parameter-free ensembles such as AutoAttack-style evaluations would provide stronger evidence against gradient masking or attack-specific overfitting (Tramèr et al., 2020; Croce & Hein, 2020). Fourth, cross-dataset validation on CIC-IDS2017/2018 or another contemporary intrusion benchmark would test whether the observed ordering is specific to UNSW-NB15. Fifth, noisy simulation and carefully budgeted IBM Quantum hardware runs should be used to study how device noise changes both clean accuracy and attack sensitivity. Such hardware experiments should be reported separately from simulator acceleration. At last, scaling studies should examine qubit count, circuit depth, gradient variance, training time, and memory cost so that any future claim of practical quantum utility is supported by resource-aware evidence rather than accuracy alone.

5.4 Practical reporting recommendations for future QML-IDS studies

Based on the experimental findings and the literature reviewed above, future QML-IDS robustness papers should report at least five categories of information. First, preprocessing must state exactly which transformations are fitted on training data, how many features reach the quantum model, and what domain the attacker perturbs. Second, the threat model should identify attacker knowledge, norm constraint, epsilon values, step count, step size, initialization, clipping, and whether gradients are taken through the evaluated quantum model. Third, robustness claims should include clean performance, attacked performance, attack success rate, and a constraint audit. Fourth, repeated trials should be accompanied by uncertainty and effect-size reporting. Fifth, at least one competitive classical model should receive the same transformed features and attack budgets.

Such reporting would make comparisons across quantum IDS studies more meaningful. It would also reduce the risk of interpreting simulator acceleration, high-dimensional Hilbert-space embeddings, or a single favorable run as evidence of practical quantum advantage. The current study follows these principles within a deliberately small four-qubit experiment and makes its implementation outputs available for scrutiny. The same protocol can be extended to quantum kernels, QCNNs, larger variational circuits, noisy simulators, and physical quantum processors without changing the core requirements for fair adversarial evaluation.

6. Conclusion

This study provides a controlled adversarial evaluation of quantum supervised learning for network intrusion detection. A four-qubit EstimatorQNN using ZZFeatureMap and RealAmplitudes was attacked directly through white-box input gradients and trained with a 50/25/25 clean-FGSM-PGD mixture. Across ten matched seeds, PGD-aware adversarial training substantially improved QSL performance under moderate and strong PGD, including accuracy gains from 9.10% to 26.90% at 0.05π and from 2.60% to

20.73% at 0.10π . These improvements were accompanied by reduced attack success rates but also by a significant clean-accuracy penalty. A matched adversarially trained MLP remained considerably more robust in absolute terms. Accordingly, the contribution is evidence for improved robustness of the tested QSL model, not evidence of quantum advantage. Future work should examine quantum kernels and QCNNs, additional IDS datasets, stronger adaptive and transfer attacks, larger feature spaces, noisy simulation, and physical QPU validation.

Declarations

Funding:

We are not receiving funding from any organization for the submitted work.

Author Contribution

Rahul R. Bhoge: Conceptualisation, methodology, software development, experimental design, implementation, data curation, formal analysis, investigation, validation, visualisation, interpretation of results, and writing—original draft preparation. Ranjit R. Keole: supervision, conceptual guidance, methodology review, validation, interpretation of results, and writing—review and editing. Pravin P. Karde: Supervision, methodological guidance, validation, critical evaluation of the experimental results, and writing—review and editing. All authors contributed to the critical revision of the manuscript, read and approved the final version, and agreed to be accountable for the integrity and accuracy of the work.

References

1. Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., & Woerner, S. (2021). The power of quantum neural networks. *Nature Computational Science*, 1, 403–409. <https://doi.org/10.1038/s43588-021-00084-1>
2. Abreu, D. M., Rothenberg, C. E., & Abelém, A. J. G. (2024). QML-IDS: Quantum machine learning intrusion detection system. In *2024 IEEE Symposium on Computers and Communications (ISCC)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ISCC61673.2024.10733655>
3. Abreu, D. M., Rothenberg, C. E., & Abelém, A. J. G. (2025). QuantumNetSec: Quantum machine learning for network security. *International Journal of Network Management*. <https://doi.org/10.1002/nem.70018>
4. Anil, G., Vinod, V., & Narayan, A. (2024). Generating universal adversarial perturbations for quantum classifiers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(10). <https://doi.org/10.1609/aaai.v38i10.28963>
5. Aparcana-Tasayco, A. J., Deng, X., & Park, J. H. (2025). A systematic review of anomaly detection in IoT security: Towards quantum machine learning approach. *EPJ Quantum Technology*, 12.

6. Benedetti, M., Lloyd, E., Sack, S., & Fiorentini, M. (2019). Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4, 043001. <https://doi.org/10.1088/2058-9565/ab4eb5>
7. Berberich, J., Fink, D., Pranjić, D., Tutschku, C., & Holm, C. (2024). Training robust and generalizable quantum models. *Physical Review Research*, 6, 043326. <https://doi.org/10.1103/PhysRevResearch.6.043326>
8. Bhoge, R., Keole, R., & Karde, P. (2026). Adversarially Robust QSL on UNSW-NB15 experiment (Version 2.0.0) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.22244259>
9. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, 549, 195–202. <https://doi.org/10.1038/nature23474>
10. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39–57). IEEE. <https://doi.org/10.1109/SP.2017.49>
11. Cerezo, M., Arrasmith, A., Babbush, R., et al. (2021). Variational quantum algorithms. *Nature Reviews Physics*, 3, 625–644. <https://doi.org/10.1038/s42254-021-00348-9>
12. Cirillo, F., Esposito, C., & Seo, J. T. (2026). Benchmarking quantum machine learning methods for intrusion detection on noisy quantum computers. *Quantum Machine Intelligence*, 8, 27. <https://doi.org/10.1007/s42484-026-00379-4>
13. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning*, 2206–2216.
14. Dowling, N., West, M. T., Southwell, A., Nakhl, A. C., Sevier, M., Usman, M., & Modi, K. (2026). Adversarial robustness guarantees for quantum classifiers. *npj Quantum Information*, 12, 16.
15. Georgiou, P., Thomas, A. M., Jose, S. T., & Simeone, O. (2025). On the generalization of adversarially trained quantum classifiers. *arXiv*. <https://arxiv.org/abs/2504.17690>
16. Gong, W., Yuan, D., Li, W., & Deng, D.-L. (2024). Enhancing quantum adversarial robustness by randomized encodings. *Physical Review Research*, 6, 023020. <https://doi.org/10.1103/PhysRevResearch.6.023020>
17. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations*.
18. Havlíček, V., Córcoles, A. D., Temme, K., et al. (2019). Supervised learning with quantum-enhanced feature spaces. *Nature*, 567, 209–212. <https://doi.org/10.1038/s41586-019-0980-2>
19. Huang, J.-C., Tsai, Y.-L., Yang, C.-H. H., Su, C.-F., Yu, C.-M., Chen, P.-Y., & Kuo, S.-Y. (2022). Certified robustness of quantum classifiers against adversarial examples through quantum noise. *arXiv*. <https://arxiv.org/abs/2211.00887>
20. Kaissar, A., Bou Nassif, A., & Bouridane, A. (2026). Enhancing network intrusion detection with quantum machine learning: A comprehensive survey of methods, metrics, and applications. *Future Internet*, 18(5), 234. <https://doi.org/10.3390/fi18050234>

21. Kim, T. H., & Madhavi, S. (2024). Quantum intrusion detection system using outlier analysis. *Scientific Reports*, 14, 27114.
22. Kukliansky, A., Orescanin, M., Bollmann, C., & Huffmire, T. (2024). Network anomaly detection using quantum neural networks on noisy quantum computers. *IEEE Transactions on Quantum Engineering*, 5, 1–11. <https://doi.org/10.1109/TQE.2024.3359574>
23. Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial examples in the physical world. *International Conference on Learning Representations Workshop*.
24. Kustiawan, Y. A., & Ghauth, K. I. (2026). Adversarial robustness in quantum machine learning: A scoping review. *Computers*, 15(4), 233. <https://doi.org/10.3390/computers15040233>
25. Küçükara, M. Y., Atban, F., & Bayılmış, C. (2024). Quantum-neural network model for platform independent DDoS attack classification in cyber security. *Advanced Quantum Technologies*, 7, 2400084.
26. Larocca, M., Ju, N., García-Martín, D., Coles, P. J., & Cerezo, M. (2023). Theory of overparametrization in quantum neural networks. *Nature Computational Science*, 3, 542–551. <https://doi.org/10.1038/s43588-023-00467-6>
27. Lu, S., Duan, L.-M., & Deng, D.-L. (2020). Quantum adversarial machine learning. *Physical Review Research*, 2, 033212. <https://doi.org/10.1103/PhysRevResearch.2.033212>
28. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*.
29. McClean, J. R., Boixo, S., Smelyanskiy, V. N., Babbush, R., & Neven, H. (2018). Barren plateaus in quantum neural network training landscapes. *Nature Communications*, 9, 4812. <https://doi.org/10.1038/s41467-018-07090-4>
30. Montalbano, G., & Banchi, L. (2025). Quantum adversarial learning for kernel methods. *Quantum Machine Intelligence*, 7, 15. <https://doi.org/10.1007/s42484-025-00238-8>
31. Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2574–2582). IEEE. <https://doi.org/10.1109/CVPR.2016.282>
32. Moustafa, N. (2017). Designing an online and reliable statistical anomaly detection framework for dealing with large high-speed network traffic [Doctoral dissertation, University of New South Wales].
33. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. In *2015 Military Communications and Information Systems Conference (MilCIS)* (pp. 1–6). IEEE. <https://doi.org/10.1109/MilCIS.2015.7348942>
34. Payares, E. D., & Martinez-Santos, J. C. (2021). Quantum machine learning for intrusion detection of distributed denial of service attacks: A comparative overview. *Proceedings of SPIE*, 11699, 116990B. <https://doi.org/10.1117/12.2593297>
35. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., & Latorre, J. I. (2020). Data re-uploading for a universal quantum classifier. *Quantum*, 4, 226. <https://doi.org/10.22331/q-2020-02-06-226>

36. Preskill, J. (2018). Quantum computing in the NISQ era and beyond. *Quantum*, 2, 79.
<https://doi.org/10.22331/q-2018-08-06-79>
37. Qiskit Machine Learning Contributors. (2026). EstimatorQNN and quantum neural networks documentation (Version 0.9.1). Qiskit Machine Learning.
38. Razavi-Far, R., Meymani, M., Mahmoudinia, E., et al. (2026). Quantum adversarial machine learning: From classical adaptations to quantum-native methods. *Artificial Intelligence Review*, 59, 177.
<https://doi.org/10.1007/s10462-026-11578-7>
39. Ren, W., Li, W., Xu, S., et al. (2022). Experimental quantum adversarial learning with programmable superconducting qubits. *Nature Computational Science*, 2, 711–717.
<https://doi.org/10.1038/s43588-022-00351-9>
40. Sai, S., Goyal, I., Sharma, S., Manuri, S. H., Chamola, V., & Buyya, R. (2025). Quantum machine learning for cybersecurity: A taxonomy and future directions. *arXiv*.
<https://arxiv.org/abs/2512.15286>
41. Schuld, M., Bocharov, A., Svore, K. M., & Wiebe, N. (2020). Circuit-centric quantum classifiers. *Physical Review A*, 101, 032308. <https://doi.org/10.1103/PhysRevA.101.032308>
42. Schuld, M., & Petruccione, F. (2018). *Supervised learning with quantum computers*. Springer.
<https://doi.org/10.1007/978-3-319-96424-9>
43. Sim, S., Johnson, P. D., & Aspuru-Guzik, A. (2019). Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms. *Advanced Quantum Technologies*, 2, 1900070. <https://doi.org/10.1002/qute.201900070>
44. Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*.
45. Tramèr, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, 1633–1645.
46. Weber, M., Liu, N., Li, B., Zhang, C., & Zhao, Z. (2021). Optimal provable robustness of quantum classification via quantum hypothesis testing. *npj Quantum Information*, 7, 76.
<https://doi.org/10.1038/s41534-021-00410-5>
47. West, M. T., Tsang, S.-L., Low, J. S., et al. (2023). Towards quantum enhanced adversarial robustness in machine learning. *Nature Machine Intelligence*, 5, 581–589. <https://doi.org/10.1038/s42256-023-00661-1>
48. Winderl, D., Franco, N., & Lorenz, J. M. (2024). Quantum neural networks under depolarization noise: Exploring white-box attacks and defenses. *Quantum Machine Intelligence*, 6, 83.
<https://doi.org/10.1007/s42484-024-00208-6>

Figures

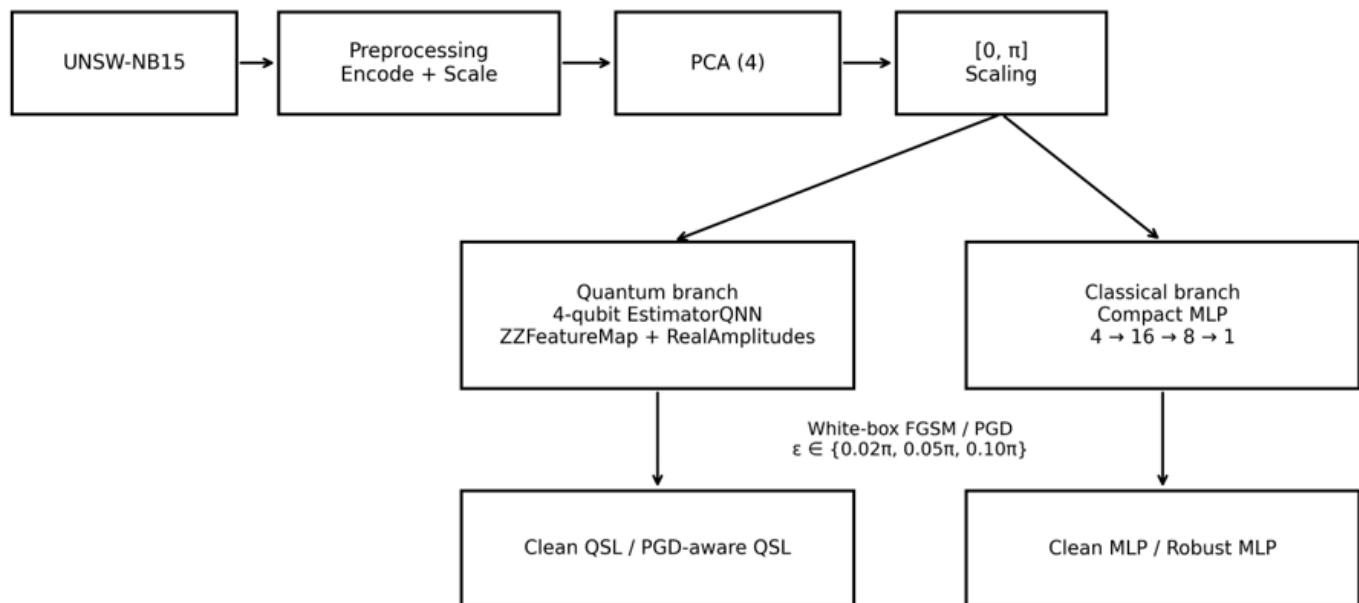


Figure 1. Controlled experimental framework for the matched QSL and MLP robustness study.

Figure 1

See image above for figure legend.

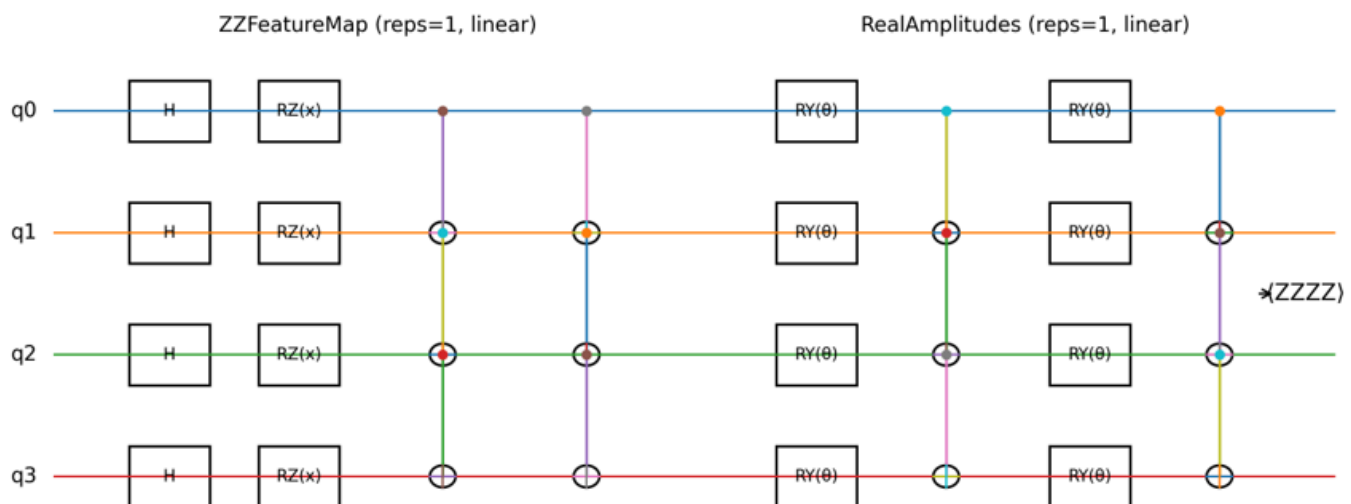


Figure 2. Conceptual circuit structure of the four-qubit QSL model used in the experiment.

Figure 2

See image above for figure legend.

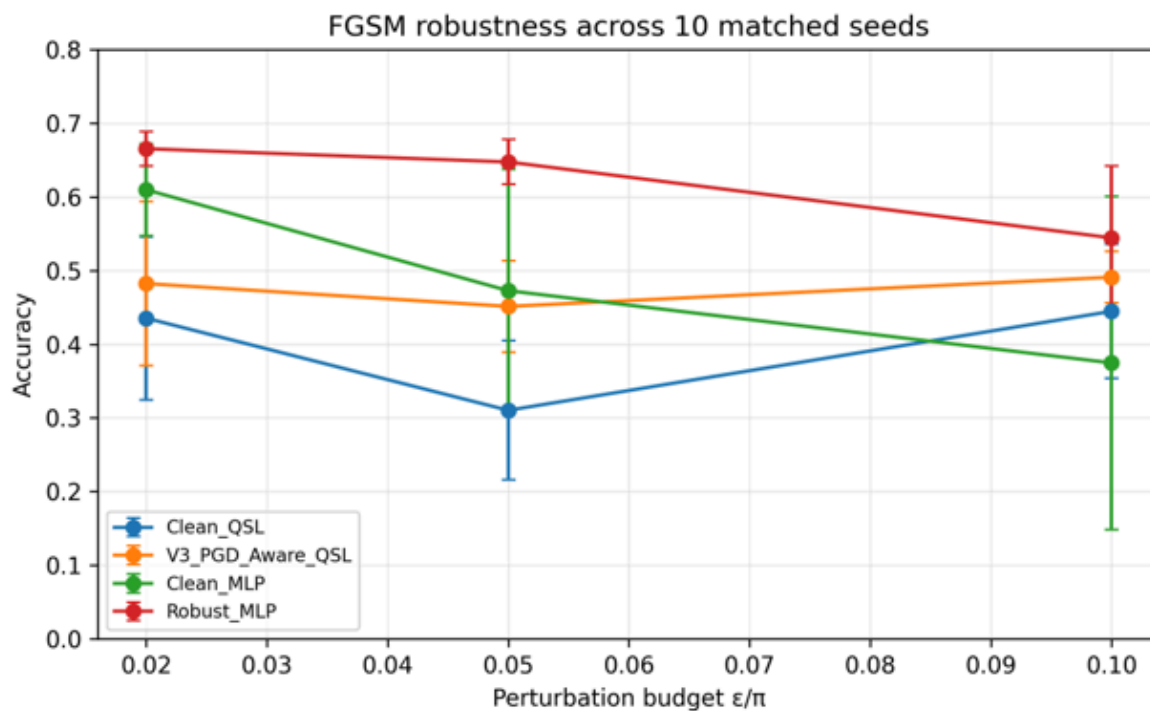


Figure 3. FGSM accuracy across ten matched seeds. Error bars show one standard deviation.

Figure 3

See image above for figure legend.

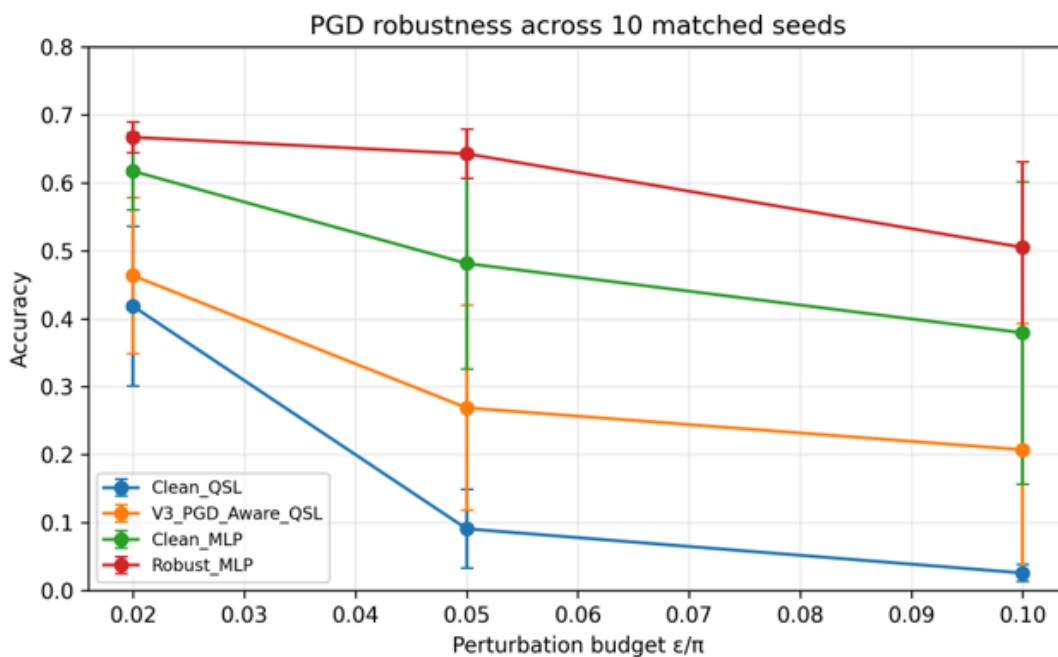


Figure 4. PGD accuracy across ten matched seeds. Error bars show one standard deviation.

Figure 4

See image above for figure legend.