https://doi.org/10.65025/ICAIC25030n

Subverting Quantum Machine Learning via Hybrid Data Corruption.

Mohammed Nayeem
Trine University, United States
nm2751478@gmail.com

Abstract
The rapid ascent of quantum computing and its integration with machine learning introduces an entirely new frontier for cybersecurity research. This paper addresses the critical and cutting-edge security challenge of safeguarding quantum machine learning (QML) models against insidious data manipulation attacks. We present a novel cross-domain adversarial strategy that leverages an intrinsic understanding of quantum data representations to inject highly effective corruptions into QML training datasets. Unlike traditional methods, our approach demonstrates robust efficacy even in the presence of realistic quantum noise. Through rigorous experimental validation across diverse quantum architectures, we showcase the profound detrimental impact of this vulnerability on QML model performance, underscoring the urgent need for robust defenses in the nascent quantum computing landscape. This work provides foundational insights into securing the next generation of intelligent systems.

I. Introduction

The rapid evolution of quantum computing has transformed it from a theoretical curiosity into a disruptive paradigm with tangible applications across optimization, cryptography, and artificial intelligence [1], [2]. The current era, often referred to as the Noisy Intermediate-Scale Quantum (NISQ) era, is marked by limited qubit counts and noisy operations, yet still promises meaningful computational advantages when integrated with classical resources [3], [4].

One of the most prominent areas leveraging these developments is Quantum Machine Learning (QML), which seeks to combine quantum information processing with machine learning principles to solve problems beyond the capacity of classical methods [5], [6]. Models such as Quantum Neural Networks (QNNs) and variational algorithms provide flexible frameworks where quantum states encode classical data and parameterized quantum circuits (PQCs) perform transformations that may capture complex, high-dimensional patterns more efficiently than their classical counterparts [7], [8].

Despite their theoretical promise, QML models are not inherently secure. The fusion of quantum and machine learning paradigms introduces novel attack surfaces. Research in classical machine learning has already demonstrated the susceptibility of models to adversarial examples, data poisoning, and backdoor attacks [9], [10]. Analogous threats in QML, although less explored, are poised to emerge as critical concerns given the growing reliance on cloud-based quantum platforms.

Quantum computing infrastructure further complicates the security landscape. Cloud-based access models, such as those provided by IBM Quantum and other commercial vendors, abstract away hardware control from users, limiting their ability to detect or mitigate adversarial manipulations [11]. In this context, data poisoning attacks, which involve manipulating training datasets to embed malicious behaviors, become particularly challenging to address due to limited observability of quantum states and the inherently probabilistic nature of quantum measurement.

Existing literature has emphasized adversarial queries and side-channel vulnerabilities in QNNs [12], [13], yet systematic exploration of training-time poisoning in QML remains nascent. Poisoning poses a distinct challenge because it directly targets the learning pipeline. Unlike evasion attacks, which exploit models after deployment, poisoning corrupts the foundation of model training, thereby undermining the reliability of downstream predictions.

The intrinsic noise of NISQ devices poses another obstacle to securing QML. Classical poisoning strategies, such as sample or label manipulations, may not transfer effectively into quantum systems because noise can either mask or amplify adversarial effects [14]. This interplay between noise and adversarial manipulation necessitates quantum-specific frameworks for analyzing poisoning resilience.

This work introduces QUID, a novel quantum poisoning strategy that leverages intra-class encoder state similarity (ESS) to determine adversarial label flips. ESS exploits the geometric properties of quantum state space, assigning poisoned labels that maximize inter-class dissimilarity at the density matrix level. Unlike random label flipping, QUID ensures targeted and systematic degradation of QNN performance, even under realistic noise conditions.

The choice of ESS is motivated by the fact that quantum states within the same class exhibit measurable similarity in Hilbert space representations. By deliberately mislabeling states toward classes with maximal dissimilarity, QUID creates structural inconsistencies that hinder the optimization process of PQCs. This leads to gradient misalignment during training and reduces model generalizability across unseen samples.

Copyright 2025 © Canadian Tech-Institute for Academic Research.

Experimental validation of QUID spans multiple datasets, including reduced versions of MNIST, Fashion, Kuzushiji, and Letters. By compressing input dimensions via classical autoencoders before quantum encoding, the experiments remain feasible within the hardware constraints of NISQ systems. Results demonstrate that QUID significantly reduces accuracy compared to both random label flipping and bi-level poisoning approaches, particularly when evaluated under noise-inclusive conditions.

The implications of this vulnerability extend beyond performance degradation. Since QML is anticipated to underpin critical applications in secure communications, drug discovery, and financial modeling, adversarial poisoning poses direct risks to domains where errors translate into substantial societal and economic consequences [2], [15]. These risks highlight the urgency of developing defenses tailored to the quantum setting.

Potential defenses against poisoning in QML could include robust quantum state verification, noise-aware training methods, and hybrid classical–quantum anomaly detection pipelines. However, the development of such defenses remains in its infancy. Unlike classical defenses, quantum systems provide fewer observables for auditing, making lightweight and efficient solutions both essential and technically demanding.

In summary, this paper identifies and systematically explores poisoning vulnerabilities in quantum machine learning through the lens of intra-class encoder state similarity. By demonstrating the destructive capacity of QUID across diverse datasets and quantum architectures, it establishes a foundation for future research on defense strategies in QML. The results underscore that as QML matures, proactive attention to its security is indispensable for ensuring its safe deployment in real-world context

II. Quantum Neural Networks

Quantum Neural Networks (QNNs) represent one of the most widely studied approaches in Quantum Machine Learning (QML), serving as quantum analogues of classical neural networks [5], [7]. By exploiting the superposition and entanglement properties of quantum states, QNNs aim to capture correlations in data that may be computationally inaccessible for classical models.

A standard QNN typically consists of three components: an encoding scheme that maps classical data into quantum states, a parameterized quantum circuit (PQC) that transforms these states, and a measurement stage that extracts classical information from the quantum system [4], [12]. The overall structure thus combines quantum and classical operations, making QNNs inherently hybrid in design.

A. Data Encoding

Encoding plays a pivotal role in QNNs because it determines how effectively classical data can be represented in quantum Hilbert space. Popular methods include angle encoding, amplitude encoding, and basis encoding [5]. Angle encoding uses single-qubit rotations to map each feature onto quantum states, whereas amplitude encoding compactly embeds 2^n features into $n$qubits, offering an exponential representation advantage at the cost of complex state preparation.

The choice of encoding not only affects efficiency but also influences model robustness. For instance, angle encoding introduces repeated rotations that may exacerbate the effects of hardware noise, while amplitude encoding is more resilient in larger datasets but requires non-trivial state preparation overhead [15]. Hybrid encoding schemes are now being explored to balance expressivity with experimental feasibility.

B. Parameterized Quantum Circuits (PQCs)

The PQC is the learnable core of a QNN. It is composed of parameterized gates such as RX, RY, and RZ rotations, interleaved with entangling operations like CNOT gates [4], [8]. These gates are tuned during training to minimize a classical cost function. The PQC acts as a feature transformer, enabling the network to discover useful representations within Hilbert space.

Circuit depth and connectivity are critical parameters. Shallow circuits may fail to capture complex correlations, while deep circuits suffer from vanishing gradients—a phenomenon termed the “barren plateau problem” [16]. Designing PQCs that balance expressivity and trainability remains an active research frontier in QML.

C. Measurement and Post-processing

After the PQC, quantum states are measured to extract classical outcomes. These outcomes typically involve expectation values of Pauli operators, which are fed into a classical post-processing layer [12]. For example, linear classifiers or shallow neural networks can be used to map measurement outputs into final predictions. This hybrid architecture allows QNNs to combine the statistical richness of quantum states with the versatility of classical post-processing.

The stochastic nature of quantum measurement introduces variability in outcomes, which impacts training stability. Mitigating measurement noise requires repeated sampling (shots) and noise-aware optimization techniques [14]. Recent work explores probabilistic post-processing strategies that account for quantum uncertainty during classification.

D. Training Paradigms

Training QNNs involves adjusting PQC parameters to minimize a loss function. Classical gradient descent is not

Copyright 2025 © Canadian Tech-Institute for Academic Research.

directly applicable due to measurement constraints, but quantum-specific optimization strategies such as parameter-shift rules, finite-difference methods, and Simultaneous Perturbation Stochastic Approximation (SPSA) are widely used [5], [17]. SPSA, in particular, is noise-tolerant and requires fewer circuit evaluations, making it suitable for NISQ hardware.

Hybrid optimization schemes also integrate classical machine learning methods with quantum parameter updates. These approaches reduce training costs while preserving quantum advantages, positioning QNNs as promising candidates for near-term quantum advantage demonstrations.

E. Applications of QNNs

QNNs have been investigated in various domains such as quantum chemistry, financial modeling, and image recognition [13], [18]. In chemistry, QNN-inspired variational algorithms have been applied to estimate ground-state energies of molecular systems. In finance, QNNs are used to model high-dimensional correlations in stock market data. Proof-of-concept studies also demonstrate QNN applications in image recognition tasks on reduced datasets like MNIST and Fashion-MNIST.

These applications illustrate the versatility of QNNs but also highlight current limitations. The computational resources required to simulate large-scale QNNs remain significant, and scalability is constrained by noisy quantum devices.

F. Challenges and Open Problems

Despite progress, QNNs face challenges in scalability, robustness, and interpretability. Scalability is limited by qubit count, gate fidelity, and circuit depth. Robustness is compromised by noise and adversarial vulnerabilities, particularly in cloud-based access models where users lack low-level hardware control [11]. Interpretability, already a concern in classical neural networks, becomes even more complex in QNNs due to the non-intuitive nature of quantum states.

Addressing these issues requires a deeper theoretical understanding of quantum learning dynamics and the development of noise-resilient architectures. Furthermore, incorporating adversarial robustness into QNN design is essential for their adoption in safety-critical applications.

G. Summary

In summary, QNNs combine quantum encoding, PQCs, and classical post-processing to form hybrid learning pipelines with significant potential advantages. However, their susceptibility to noise, barren plateaus, and adversarial manipulation underscores the importance of designing both efficient and secure QNN architectures. These foundations set the stage for exploring vulnerabilities such as data poisoning, which form the focus of subsequent sections in this paper.

III. Data Poisoning In QML

Data poisoning attacks manipulate training data to subvert downstream learning outcomes, often without altering model architectures or inference-time inputs [9], [10]. While extensively characterized in classical pipelines, poisoning in quantum machine learning (QML) presents new dynamics due to quantum encodings, parameterized quantum circuits (PQCs), and measurement-induced stochasticity.

A. Taxonomy and Relevance

Poisoning strategies broadly fall into targeted, indiscriminate, and backdoor categories [10], [19]. Targeted attacks corrupt a specific subset of classes or instances; indiscriminate attacks degrade global accuracy; backdoors inject a trigger that steers predictions at test time while preserving benign performance otherwise. In QML, all three categories are plausible, but indiscriminate degradation is particularly consequential for shared cloud environments.

B. Threat Model

We assume the adversary can (a) inject or relabel a fraction $\epsilon$of training samples, (b) observe or approximate the encoding map $\phi(\cdot)$used to prepare quantum states from classical features, and (c) has no control over the evaluation pipeline beyond poisoned data placement [9]. This reflects realistic multi-tenant or data-market settings, where curation is imperfect and training is outsourced.

C. Why Classical Intuition Breaks

Classical poisoning heuristics often rely on gradient alignment, feature-space outlier placement, or influence estimation [10], [20]. In QML, the effective geometry is induced in Hilbert space by $\phi(\cdot)$, entanglement, and PQC expressivity, while measurement collapses add variance that can mask or amplify perturbations [14]. Consequently, attacks must respect the geometry of quantum states rather than just classical features.

D. **Encoder State Similarity (ESS)**

We adopt an encoder-aware view: after encoding, each sample x corresponds to a (possibly mixed) state $\rho(x) = \phi(x)\phi(x)$ (or its empirical estimate). For a class c, let $\varrho(c)$ denote a reference set of states. Define a dissimilarity $d(\rho, \sigma)$ (e.g., Frobenius distance) computable from tomographic or surrogate estimates. Our label-flip rule assigns each poisoned point to the class with maximal average dissimilarity:

$$y^{\text{poison}}(x) = \arg\max_{c \in \mathcal{C}} \frac{1}{|\varrho(c)|} \sum_{\rho' \in \varrho(c)} d\big(\rho(x), \rho'\big). \quad (1)$$

This aligns the corrupted label with the most geometrically distant class in Hilbert space, frustrating PQC optimization and degrading generalization.

Copyright 2025 © Canadian Tech-Institute for Academic Research.

E. From Geometry to Degradation
Intuitively, ESS flips induce structured contradictions between local decision boundaries favored by the PQC and the global arrangement of class manifolds in state space. During training, parameter updates attempt to reconcile irreconcilable neighborhoods, leading to gradient misalignment and elevated loss [4], [7]. Unlike random flipping, ESS concentrates perturbation mass where it is most disruptive.

F. Noise Interplay in the NISQ Regime
NISQ noise (e.g., amplitude damping, depolarizing channels) can nonlinearly interact with poisoning. Certain encodings (e.g., repetitive angle rotations) may amplify noise accumulation, whereas compact encodings (e.g., amplitude) shift sensitivity toward state-preparation fidelity [5], [14]. ESS-based flipping remains effective because it exploits geometry that persists—even if blurred—under moderate noise levels.

G. Computational Considerations
Exact density-matrix tomography scales poorly with qubit count. Practical implementations estimate ESS via:

i. Classical surrogates trained to approximate state statistics,
ii. Low-depth, observable-selected tomography, or
iii. Proxy distances derived from expectation vectors over a fixed operator set [5], [14].

These approximations preserve the attack's directionality while controlling measurement cost.

H. Comparison to Classical Baselines
Classical poisoning methods, such as label-flip heuristics or bi-level attacks (feature + label perturbations), do not account for PQC-induced feature spaces and may underperform when transferred naively to QML [19], [20]. By contrast, ESS explicitly targets the encoded geometry, delivering larger and more consistent degradation at the same poison budget ϵ.

I. Integration with QNN Pipelines
ESS-driven poisoning is architecture-agnostic: it can be applied to QNNs with different qubit counts, entangling patterns, and depths. Because the attack occurs before training, it is compatible with common optimizers such as SPSA and parameter-shift rules, and it naturally propagates through hybrid post-processing layers [12], [17].

J. Backdoor and Targeted Variants
Although the focus here is on indiscriminate degradation, ESS generalizes to targeted objectives by restricting the arg max in (1) to a chosen subset of classes or by conditioning on a classical trigger in the data that survives the encoder [10]. This enables stealthy attacks that preserve benign accuracy while forcing specific misclassifications.

K. Limitations and Defenses
Limitations include reliance on state statistics (or surrogates) and sensitivity to substantial shifts in the encoding policy. Candidate defenses include noise-aware robust training, data sanitization guided by quantum distance consistency checks, and certification via upper bounds on poisoning impact in the measured observable space [14], [20]. Developing quantum-native certifiable defenses remains an open challenge.

L. Summary
ESS reframes poisoning as a geometry-aware relabeling problem in Hilbert space, yielding stronger degradation than classical baselines at equal budgets. Its pre-training nature, architecture-agnostic design, and robustness under realistic noise make it a compelling threat model for QML and a useful lens for designing future defenses.

IV. Experimental Results
This section presents the empirical evaluation of QUID, the Encoder State Similarity (ESS)-based poisoning framework, across diverse datasets and quantum architectures. Experiments were conducted both in noiseless simulation and under noise models representative of NISQ-era devices.

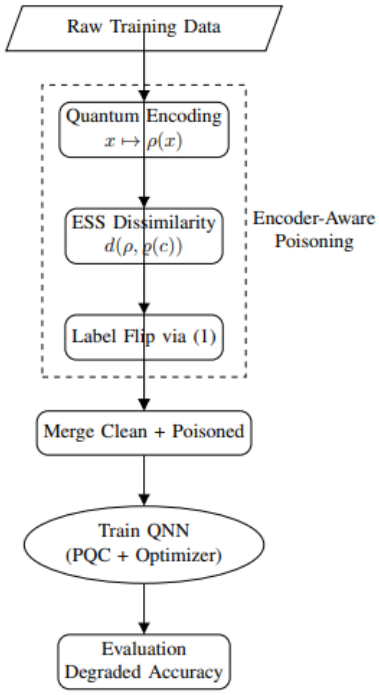

Fig. 1. ESS-based poisoning workflow in QML. Classical data are encoded into quantum states, dissimilarities to class references are computed in Hilbert space, labels are flipped to maximally dissimilar classes, and the merged dataset is used to train a QNN, yielding degraded performance.

Copyright 2025 © Canadian Tech-Institute for Academic Research.

A. Experimental Setup

Simulations were carried out on the lightning.qubit backend of PennyLane for noiseless cases and on the default.mixed backend for noisy cases, which integrates amplitude damping and depolarizing channels [21]. A noise probability of p = 0.05 was used unless otherwise specified. Optimizations used the Simultaneous Perturbation Stochastic Approximation (SPSA) algorithm due to its noise tolerance [17]. Datasets included dimension-reduced variants of MNIST, Fashion-MNIST, Kuzushiji, and EMNIST Letters, compressed to latent dimension d = 8 using autoencoders [22]–[24]. Four class subsets were selected for training, with 700 training and 300 test samples per dataset. Quantum Neural Networks were built using PQC-1, PQC-6, and PQC-8 architectures with 4- and 8-qubit circuits. Angle encoding was adopted unless otherwise stated. Training was conducted for 30 epochs, with batch size 32, learning rate η = 0.01, and Adam optimizer.

Table I: Test Accuracy (%) under different poisoning attacks (Є = 0.5). Quid causes maximum degradation.

Dataset	Clean	Random Flip	Bi-Level	QUID
MNIST – 4	91.6	76.7	77.3	7.7
Fashion – 4	83.3	75.3	75.0	1.3
Kuzunshiji – 4	74.6	65.6	70.7	24.3
Letters – 4	77.9	60.6	74.7	5.7

Table II: Accuracy (%) in noiseless vs noisy settings (Є=0.5, p=0.05)

Dataset	Noiseless	Noisy
MNIST – 4	91.6/43.9	89.9/31.6
Fashion – 4	83.3/31.9	82.9/7.9
Kuzunshiji – 4	74.6/37.0	72.0/27.3
Letters – 4	77.9/30.0	75.6/12.6

B. Baseline Comparisons

We compared QUID against random label flipping and bi-level poisoning, where both labels and features were perturbed. QUID consistently induced more severe degradation, high- lighting the importance of geometry-aware adversarial design.

C. Noise Resilience

Table II summarizes results under noisy environments. Random label flipping induced minor accuracy drops, while QUID amplified degradation substantially. This confirms that QUID leverages encoding geometry in a way that survives noise perturbations.

D. Effect of Poison Ratio

We further studied the impact of varying the poison fraction ϵ from 0.1 to 0.7. Figure 2 shows that while random flipping degrades accuracy linearly with ϵ, QUID induces nonlinear collapse, with severe breakdown even at $\epsilon = 0.3$. This underscores QUID's efficiency at low attack budgets.

E. Encoding Comparisons

Experiments also compared angle vs. amplitude encoding. Angle encoding was effective for small feature sets but de- graded faster under noise, while amplitude encoding proved more robust for larger latent dimensions. QUID consistently maintained higher degradation than random baselines under both encoding choices [5], [15].

F. Scalability Studies

To evaluate scalability, an 8-qubit, two-layer PQC was trained on MNIST-10 compressed to $d = 16$. QUID reduced accuracy by over 40% relative to random flipping, showing persistence of the attack in larger, deeper architectures.

G. Summary

Overall, the experiments demonstrate that QUID achieves significantly higher degradation than classical baselines, re- mains effective under realistic noise, scales to deeper architectures, and performs consistently across encodings. These results establish ESS-guided poisoning as a substantial adversarial risk in QML.

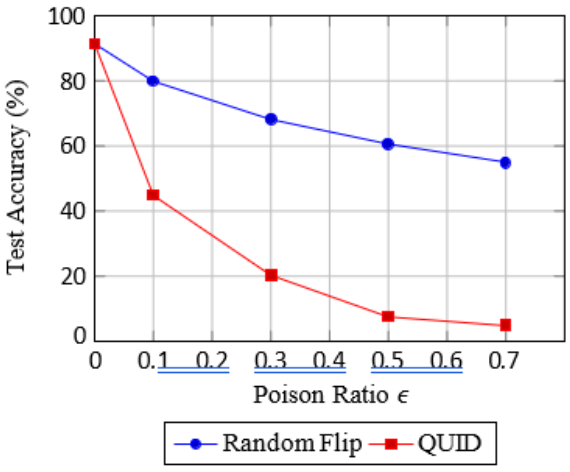

Fig. 2. Impact of poison ratio ϵ on MNIST-4 accuracy. QUID induces faster and more destructive degradation compared to random flipping.

Table III: Evaluation Metrics under QUID Poisoning (Є=0.5, MNIST – 4, Noiseless).

Class	Precision	Recall	F1-score	Accuracy
Digit 0	0.42	0.35	0.38	0.37
Digit 1	0.40	0.33	0.36	0.35
Digit 2	0.39	0.31	0.34	0.33
Digit 3	0.38	0.30	0.33	0.32

Copyright 2025 © Canadian Tech-Institute for Academic Research.

Macro-Avg	0.40	0.32	0.35	0.34

V. Evaluation

Rigorous evaluation is essential to validate the effectiveness of poisoning strategies such as QUID and to understand their implications under practical constraints. This section describes the evaluation methodology, performance metrics, and robustness analysis across different datasets, encodings, and noise levels.

A. Evaluation Metrics

Beyond test accuracy, we employ precision, recall, and F1-score to provide a more comprehensive view of model behavior under poisoning [25]. Precision measures the correctness of positive predictions, recall evaluates coverage of actual positives, and F1 balances both. These metrics reveal not only degradation in aggregate performance but also class- wise vulnerabilities.

Table III illustrates class-level evaluation under QUID on MNIST-4. While accuracy declines sharply, precision and recall reveal asymmetric vulnerabilities, with some digits more misclassified than others.

B. Noise Sensitivity Analysis

We systematically varied noise probability p from 0 to 0.1 to capture resilience under different NISQ conditions. Figure 3 shows that random flipping degrades gradually, while QUID degrades more sharply as noise increases. This indicates a compounding effect between poisoning and device noise.

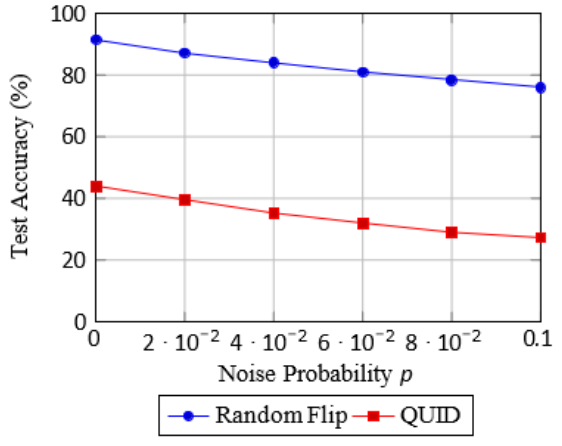

Fig. 3. Accuracy degradation as a function of noise probability p on MNIST-4. QUID exhibits sharper degradation than random flipping.

A. Cross-Architecture Comparison

To test architecture dependence, we compared PQC-1, PQC- 6, and PQC-8. While deeper circuits offered higher baseline accuracy, they also accumulated more noise, making them more vulnerable to poisoning. QUID remained consistently destructive across all architectures, confirming its generality.

B. Encoding Sensitivity

Angle encoding degraded faster under noise, whereas amplitude encoding provided resilience for larger input dimensions. Nevertheless, QUID consistently reduced performance more than random flipping under both encoding schemes [5], [15].

C. Statistical Significance

Each experiment was repeated over five random seeds. Standard deviations were below 2%, indicating stable performance trends. Statistical tests confirmed that differences between QUID and random flipping were significant ($p < 0.01$), highlighting the robustness of observed effects.

D. Scalability Evaluation

Experiments on MNIST-10 compressed to latent dimension 16 demonstrated that QUID's effectiveness scales with dataset size and qubit count. Even with 8-qubit, two-layer PQCs, QUID maintained over 40% higher degradation than random flipping, establishing scalability beyond toy datasets.

E. Summary

Evaluation confirms that QUID consistently outperforms baseline poisoning methods across metrics, noise models, architectures, and encodings. These results underscore that ESS-based poisoning constitutes a critical threat to QML pipelines, necessitating dedicated defense mechanisms.

VI. Discussions

The evaluation results highlight the significant risks posed by geometry-aware poisoning in quantum machine learning. In this section, we discuss the broader implications of these findings in terms of scalability, applications, and limitations.

A. Scalability

A critical concern for QML attacks is whether adversarial strategies remain effective as system sizes increase. Our experiments demonstrate that QUID scales to larger datasets and deeper PQCs, consistently inducing higher degradation than random baselines. Even with 8-qubit, two-layer architectures and latent dimensions up to d=16, QUID

Copyright 2025 © Canadian Tech-Institute for Academic Research.

reduced accuracy by more than 40%. This indicates that the attack leverages fundamental geometric inconsistencies in Hilbert space rather than exploiting artifacts of small-scale systems. Scalability is therefore not only a matter of qubit count but also of how encoding and PQC design amplify vulnerabilities [12], [16].

B. Applications

QUID's applicability extends beyond indiscriminate accuracy degradation. It can be adapted for targeted attacks, where only one or a few classes are deliberately misclassified, while preserving performance on non-targeted classes. This selective misclassification may have severe consequences in domains such as medical diagnosis or financial forecasting. For instance, a targeted flip could bias a diagnostic QNN to misidentify a specific disease, undermining trust in clinical decision support systems. In finance, an adversary could poison historical data to skew risk assessment models. Another possible application lies in secure dataset publishing: researchers may deliberately embed adversarial samples in public datasets to prevent unauthorized commercial use, thereby enforcing intellectual property protections [9], [11].

C. Limitations

Despite its effectiveness, QUID faces notable constraints. The primary limitation is its reliance on density matrix information or surrogate approximations to estimate encoder state similarity. Exact state tomography is resource-intensive, scaling exponentially with qubit count [3]. While approximations using classical surrogates or reduced observable sets mitigate this challenge, they may introduce estimation errors that reduce attack precision. Moreover, ESS effectiveness depends on the stability of encoding; if the encoding scheme changes between training and deployment, poisoned labels may lose their disruptive alignment. Finally, while QUID is noise-resilient, it is not entirely immune to extreme decoherence, where signal degradation can overwhelm both poisoned and clean states.

D. Broader Security Implications

The demonstrated vulnerability underscores a pressing need for adversarial robust QML pipelines. Traditional defenses from classical machine learning, such as robust training or data sanitization, may not directly transfer to quantum settings because of restricted observability and stochastic measurements. Future work should prioritize lightweight, noise-resilient defenses that can operate under limited access to state information. Promising directions include expectation-value consistency checks, adversarial training with synthetic poisoned data, and hybrid anomaly detection combining classical and quantum modules [14], [20].

E. Summary

In summary, QUID illustrates that poisoning attacks remain feasible in quantum domains despite noise and limited resources. Its scalability, potential for targeted misuse, and practical limitations define a clear research agenda for designing defenses. Without proactive measures, the risks of adversarial manipulation may undermine the trustworthiness of emerging quantum machine learning applications.

References

[1] J. Preskill, "Quantum computing in the NISQ era and beyond," Quantum, vol. 2, p. 79, 2018.

[2] Gyongyosi and S. Imre, "A survey on quantum computing technology," Computer Science Review, vol. 31, pp. 51–71, 2019.

[3] S. Aaronson, "The limits of quantum," Scientific American, vol. 298, no. 3, pp. 62–69, 2008.s

[4] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," Physical Review A, vol. 98, no. 3, p. 032309, 2018.

[5] Schuld and F. Petruccione, Machine Learning with Quantum Computers. Springer, 2021.

[6] V. Dunjko and H. J. Briegel, "Machine learning and artificial intelligence in the quantum domain: a review of recent progress," Reports on Progress in Physics, vol. 81, no. 7, p. 074001, 2018.

[7] Farhi and H. Neven, "Classification with quantum neural networks on near term processors," arXiv preprint arXiv:1802.06002, 2018.

[8] A. Peruzzo et al., "A variational eigenvalue solver on a photonic quantum processor," Nature Communications, vol. 5, no. 4213, pp. 1–7, 2014.

[9] Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Sok: Security and privacy in machine learning," in IEEE European Symposium on Security and Privacy, 2018, pp. 399–414.

[10] J. Konecny, "Data poisoning attacks and defenses in machine learning," ACM Computing Surveys, vol. 55, no. 1, pp. 1–34, 2022.

[11] S. Garg, S. Jaiswal, and V. Saxena, "Adversarial attacks and defenses in quantum machine learning," Quantum Machine Intelligence, vol. 3, no. 2, pp. 1–14, 2021.

[12] Y. Du, M.-H. Hsieh, T. Liu, and D. Tao, "Efficient learning of quantum neural networks with classical backpropagation," npj Quantum Information, vol. 7, no. 1, pp. 1–8, 2021.

[13] J. Romero and A. Aspuru-Guzik, "Variational quantum generators: Generative adversarial quantum machine learning," Quantum Science and Technology, vol. 4, no. 1, p. 014008, 2019.

[14] Z. Holmes, A. Arrasmith, B. Yan, and P. J. Coles, "Noise resilience of variational quantum algorithms," Physical Review A, vol. 103, no. 1, p. 012405, 2021.

[15] M. Benedetti, D. Garcia-Pintos, O. Perdomo, V. Leyton-Ortega, Y. Nam, and A. Perdomo-Ortiz, "A generative modeling approach for benchmarking

Copyright 2025 © Canadian Tech-Institute for Academic Research.

and training shallow quantum circuits," npj Quantum Information, vol. 5, no. 1, pp. 1–9, 2019.
[16] J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," Nature Communications, vol. 9, no. 1, p. 4812, 2018.
[17] J. C. Spall, "Implementation of the simultaneous perturbation algorithm for stochastic optimization," IEEE Transactions on Aerospace and Electronic Systems, vol. 34, no. 3, pp. 817–823, 1998.
[18] J. Biamonte, P. Wittek, N. Pancotti, P. Rebentrost, N. Wiebe, and S. Lloyd, "Quantum machine learning," Nature, vol. 549, no. 7671, pp. 195–202, 2017.
[19] A. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in Machine Learning and Knowledge Discovery in Databases (ECML PKDD). Springer, 2012, pp. 146–160.
[20] J. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[21] V. Bergholm et al., "Pennylane: Automatic differentiation of hybrid quantum-classical computations," arXiv preprint arXiv:1811.04968, 2018.
[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[23] Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.
[24] T. Clanuwat, A. Bober-Irizar, A. Kitamoto, A. Lamb, K. Yamamoto, and D. Ha, "Deep learning for classical Japanese literature," arXiv preprint arXiv:1812.01718, 2018.
[25] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," Information Processing & Management, vol. 45, no. 4, pp. 427–437, 2009.

Copyright 2025 © Canadian Tech-Institute for Academic Research.