Poster: Critical Evaluation of Quantum Machine Learning for Adversarial Robustness

Saeefa Rubaiyet Nowmi*, Jesus Lopez*, Md Mahmudul Alam Imon†,
Shahrooz Pouryousef‡, Mohammad Saidur Rahman*
*University of Texas at El Paso, †University of Dhaka, ‡University of Massachusetts Amherst
{srnowmi, jlopez126}@miners.utep.edu, mahmudulalam.imon@gmail.com,
shahrooz@cs.umass.edu, msrahman3@utep.edu

***Abstract*—Quantum Machine Learning (QML) integrates quantum principles into learning algorithms, but its adversarial robustness remains underexplored. We present the first systematization of adversarial robustness in QML across black-box, gray-box, and white-box threat models, implementing five representative attacks: label-flipping; QUID and the Huang backdoor; QTrojan and FGSM/PGD evasion. Amplitude encoding achieves 93% accuracy on MNIST in deep noiseless circuits but degrades below 5% under noise (p=0.01). FGSM/PGD compute gradients through the *entire* hybrid model, encoding determines how perturbations propagate, making it the critical security interface. QTrojan preserves clean accuracy but fails in multi-class settings, revealing a scalability limitation. These findings highlight quantum noise as a natural but unreliable defense in NISQ-era systems.**

1. Introduction

QML is transitioning from theory to deployment on cloud platforms like IBM Quantum [1]. Hybrid quantum-classical architectures inherit classical ML vulnerabilities while introducing quantum-specific attack surfaces like circuit tampering [2] and side-channel attacks [3]. Multi-tenant environments further broaden the attack surface through crosstalk [3] and data leakage [4]. No prior work has unified these threats across the full QML pipeline. This work asks: (1) What are the dominant threat vectors? (2) How do circuit depth and encoding influence robustness? (3) How resilient are QML models against black-box, gray-box, and white-box attacks?

2. Taxonomy

We decompose the adversarial landscape along three dimensions: access level (black-box, gray-box, white-box), mechanism type (classical-style vs. quantum-specific), and attack phase (training-time vs. inference-time). Figure 1 maps these attacks onto the QML pipeline.

Black-Box Attacks. The adversary has no internal visibility. Attacks include model extraction, crosstalk-induced side-channel attacks [3], membership inference, backdoor attacks, and data poisoning.

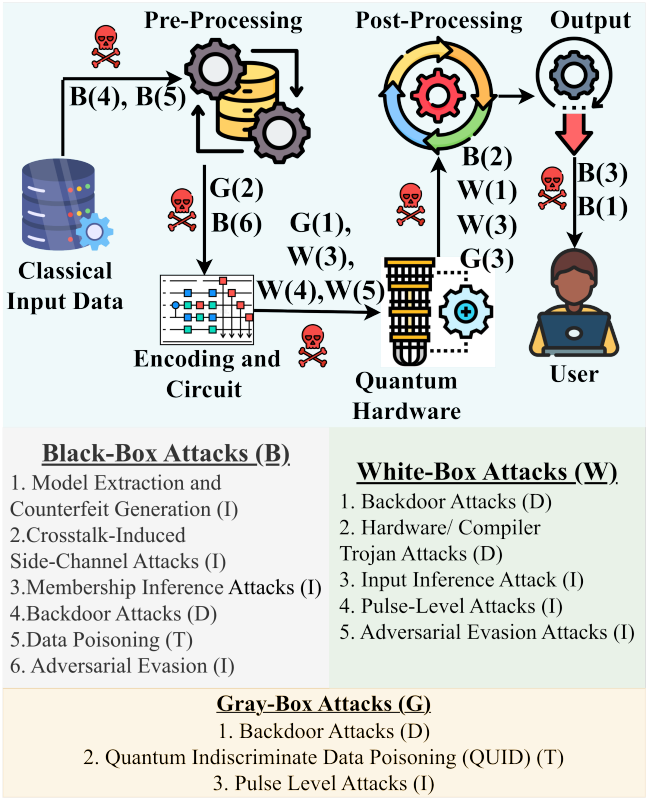


Figure 1: Adversarial attack deployment surface within the QML pipeline. **(B)**: Black-Box, **(G)**: Gray-Box, **(W)**: White-Box, **(T)**: Training-time, **(I)**: Inference-Time, **(D)**: Dual-Phase.

Gray-Box Attacks. The adversary has partial visibility into encoded quantum states or transpiled circuits. QUID [5] exploits geometric closeness of quantum states in Hilbert space to perturb labels, while the Huang backdoor [6] embeds triggers in trained weights via a proxy model.

White-Box Attacks. The adversary has full access to parameters, architecture, and hardware. QTrojan [2] inserts dormant gate layers around the encoding circuit. FGSM and PGD [7] craft minimal input perturbations.

TABLE 1: Selected results. Rel. Acc. = ratio to clean baseline, ASR = attack success rate, Rel. CDA = relative clean-data accuracy.

Setting	Configuration	MNIST	AZ-Class
Baseline Accuracy (%)			
Noiseless	Amplitude, 50L	92.6	67.0
Noisy	Angle, 2L	59.2	47.0
Noisy	Amplitude, 50L	9.7	4.8
Label-Flipping Rel. Acc.			
	QMLP-Amp., 50L	0.97	0.93
	CMLP baseline	0.50	0.51
QTrojan, Angle Enc. (Rel. CDA / ASR %)			
Noiseless			
	Angle, 2L	1.13 / 17.12	0.98 / 7.37
	Angle, 10L	1.03 / 6.50	1.01 / 1.63
Noisy (p=0.01)			
	Angle, 2L	1.24 / 20.83	1.00 / 6.85
	Angle, 10L	1.16 / 1.56	1.03 / 2.54

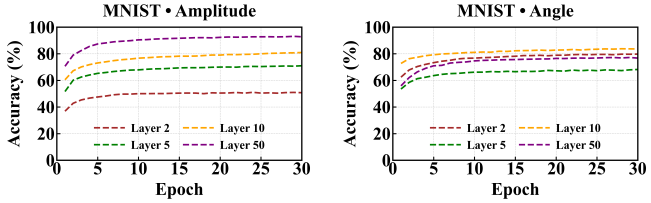


Figure 2: QMLP accuracy on MNIST with varying circuit layers and encoding under noiseless conditions.

3. Experimental Setup

For black-box, we use *label-flipping* (50% flipped) with *label smoothing*. For gray-box, we use *QUID* [5] with *Q-Detection* [8], and the *Huang backdoor* [6]. For white-box, we use *FGSM/PGD* and *QTrojan* [2]. QMLP models use 9-qubit circuits on MNIST and AZ-Class (reduced via PCA), varying encoding, depth (2–50 layers), and noise (p=0.01). A Classical MLP (CMLP) baseline enables comparison.

4. Results

Table 1 summarizes key results.

Baseline. Amplitude encoding reaches 92.6% on MNIST at 50 layers; angle peaks at 83.6% (10 layers) under noiseless conditions (Figure 2). Under noise, amplitude collapses to ∼10%, but shallow angle circuits retain 59.2%.

Label-Flipping (Black-Box). QMLPs retain 90–93% relative accuracy under 50% label-flipping, while CMLP loses ∼50%, partly due to NISQ-imposed PCA reduction acting as inadvertent regularization.

QTrojan (White-Box, Circuit-Level). QTrojan is stealthy (relative CDA 0.98–1.24) but ASR peaks at only 17.12% on MNIST and 7.37% on AZ-Class– at or below random chance– revealing a *scalability limitation* not characterized in the original 2–4 class evaluation [2].

FGSM/PGD (White-Box). At ϵ=0.01, amplitude-encoded QMLPs collapse to 30% relative accuracy due to single-shot embedding creating a fixed adversarial entry point, while CMLP retains 98%. At $\epsilon \geq 0.10$ on AZ-Class, shallow angle QMLPs outperform CMLP.

5. QML Security Guidelines

Encoding selection is the most consequential security decision: shallow angle encoding offers the best robustness-accuracy trade-off, while amplitude encoding requires quantum-aware validation [9]. Circuits should be protected via quantum logic locking (E-LoQ) [10], and hardware defenses should employ randomized qubit mapping [3] and pulse-channel verification. Quantum noise is asymmetric and unreliable as a passive defense, so active, threat-aware mechanisms are essential.

6. Conclusion

Shallow angle-encoded circuits offer the best robustness-accuracy trade-off under NISQ noise. QMLPs resist poisoning better than classical models but are more vulnerable to gradient-based evasion targeting the full hybrid model. Circuit-level attacks like QTrojan fail to scale beyond few-class settings, and quantum noise is an unreliable defense, underscoring the need for quantum-native robustness techniques.

References

[1] IBM Quantum, "IBM Quantum," https://www.ibm.com/quantum, 2025, accessed: 2025-08-04.

[2] C. Chu, L. Jiang, M. Swany, and F. Chen, "Qtrojan: A circuit backdoor against quantum neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[3] N. Choudhury, C. N. Mude, S. Das, P. C. Tikkireddi, S. Tannu, and K. Basu, "Crosstalk-induced side channel threats in multi-tenant nisq computers," in *Network and Distributed System Security (NDSS) Symposium*, 2025.

[4] C. Lu, E. Telang, A. Aysu, and K. Basu, "Quantum Leak: Timing side-channel attacks on cloud-based quantum services," in *Proceedings of the Great Lakes Symposium on VLSI (GLSVLSI)*, 2025.

[5] S. Kundu and S. Ghosh, "Adversarial data poisoning attack on quantum machine learning in the nisq era," in *Great Lakes Symposium on VLSI (GLSVLSI)*, 2025.

[6] C.-Y. Huang and S.-B. Zhang, "A backdoor attack against quantum neural networks with limited information," *Chinese Physics B*, 2023.

[7] M. T. West, S. M. Erfani, C. Leckie, M. Sevior, L. C. Hollenberg, and M. Usman, "Benchmarking adversarially robust quantum machine learning at scale," *Physical Review Research*, 2023.

[8] H. He, X. Lin, J. Chen, and Y. Xiao, "Q-Detection: A quantum-classical hybrid poisoning attack detection method," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

[9] S. Das and S. Ghosh, "Randomized reversible gate-based obfuscation for secured compilation of quantum circuit," *arXiv preprint arXiv:2305.01133*, 2023.

[10] Y. Liu, J. John, and Q. Wang, "E-loq: Enhanced locking for quantum circuit ip protection," in *IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 2025.

Critical Evaluation of Quantum Machine Learning for Adversarial Robustness

Saeefa Rubaiyet Nowmi*, Jesus Lopez*, Md Mahmudul Alam Imon†, Shahrooz Pouryousef‡, Mohammad Saidur Rahman*

*University of Texas at El Paso, †University of Dhaka, ‡University of Massachusetts Amherst

Introduction

Quantum Machine Learning (QML)

- Quantum computational principles → learning algorithms
- Improved representational capacity

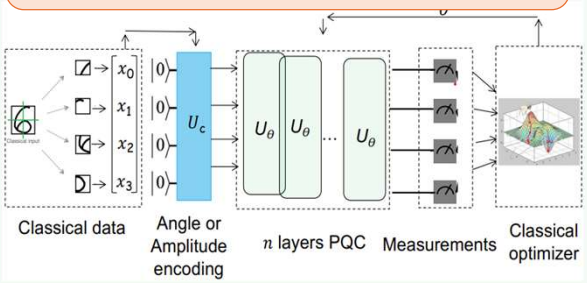
Security and robustness of QML remain largely underexplored

RQ1. Dominant threat vectors in hybrid quantum-classical architectures and their systematic classification.

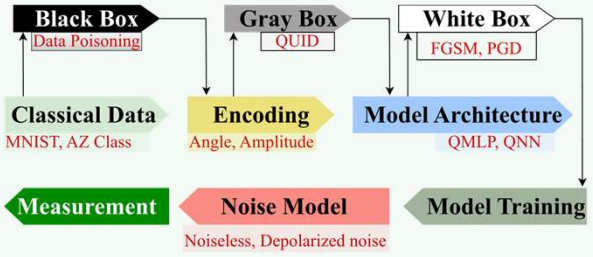
RQ2. Effects of circuit depth and data encoding on QML performance and adversarial robustness under NISQ constraints.

RQ3. Resiliency of current QML models and defenses against diverse attacks.

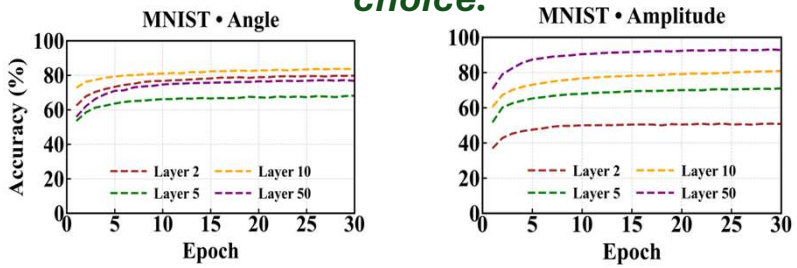
Parameterized Quantum Circuit as a QML model



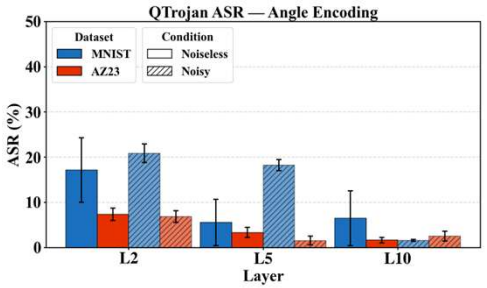
Experimental Setup



Deep amplitude encoded circuits achieve the best accuracy. But for NISQ devices, shallow angle encoded circuits are the most optimal choice.

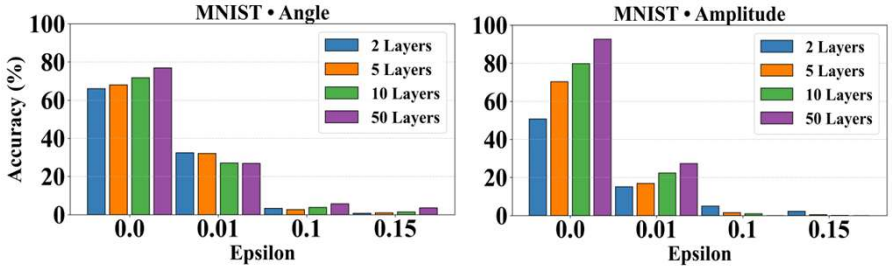


Qtrojan: Stealthy but fails to scale in multi-class classification

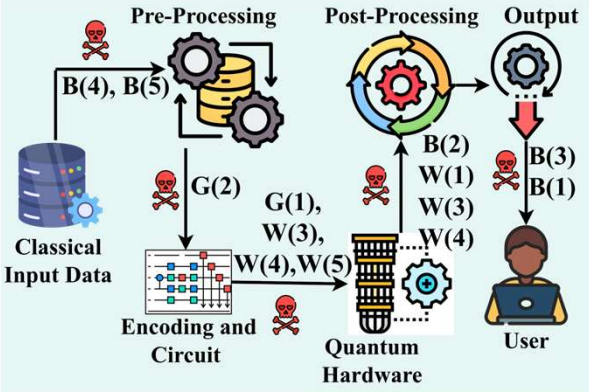


Dataset	L	Rel. CDA		ASR (%)	
		NL	Noisy	NL	Noisy
AZ-23	2	0.98	1.00	7.37	6.85
	5	1.00	1.02	3.33	1.54
	10	1.01	1.03	1.63	2.54
MNIST	2	1.13	1.24	17.12	20.83
	5	1.01	1.09	5.55	18.25
	10	1.03	1.16	6.50	1.56

Compared to CML, Shallow Angle-Encoded QML more Robust to FGSM



Attack Taxonomy



Black-Box Attacks (B)

1. Model Extraction and Counterfeit Generation
2. Crosstalk-Induced Side-Channel Attacks
3. Membership Inference Attacks
4. Backdoor Attacks
5. Data Poisoning (Label-Flipping)

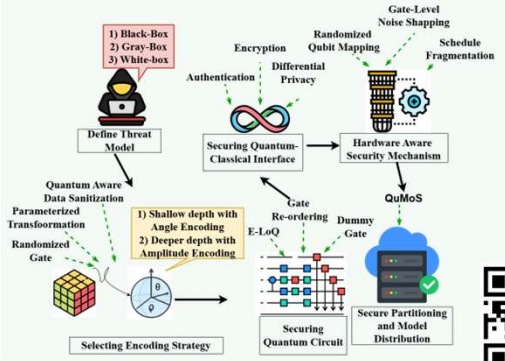
White-Box Attacks (W)

1. Circuit Level Backdoor Attacks
2. Hardware/ Compiler Trojan Attacks
3. Input Inference Attack
4. Pulse-Level Attacks
5. Adversarial Evasion Attacks

Gray-Box Attacks (G)

1. Backdoor Attacks (QuPTs and QTrojan)
2. Quantum Indiscriminate Data Poisoning (QUID)

Defense Pipeline



Scan me!