Towards Robust Quantum Machine Learning

Saeefa Rubaiyet Nowmi*†, Jesus Lopez†, Md Mahmudul Alam Imon†, Viviana Cadena†,
Shahrooz Pouryousef‡, and Mohammad Saidur Rahman†
†Department of Computer Science, University of Texas at El Paso, TX, USA
{snowmi, jlopez126, mimon, vcadena1}@miners.utep.edu, msrahman3@utep.edu
‡University of Massachusetts Amherst, Amherst, MA, USA shahrooz@cs.umass.edu

Abstract

Quantum Machine Learning (QML) integrates quantum computing with machine learning to address complex, high-dimensional problems that are challenging for classical methods. Models such as Quantum Multilayer Perceptrons (QMLPs), Quantum Convolutional Neural Networks (QCNNs), and Quantum Variational Circuits (QVCs) have shown strong potential in applications like image recognition and malware classification. However, similar to their classical counterparts, these models remain vulnerable to adversarial manipulation. Motivated by this vulnerability, this work systematically investigates how adversarial attacks ranging from training-time data poisoning to inference-time evasion (e.g., FGSM and PGD) manifest within quantum architectures.

We examine how classical adversarial strategies translate into quantum systems through parameterized encodings, hybrid quantum-classical designs, and probabilistic measurement processes. Our evaluation spans multiple quantum encodings (angle and amplitude) and varying circuit depths to assess robustness, generalization, and sensitivity to perturbations. Experimental results reveal that both encoding scheme and circuit depth significantly influence model performance. On the AZ dataset, amplitude encoding with 10 layers achieved 51.62% accuracy, outperforming shallow (2-layer) circuits but remaining below MNIST performance. On MNIST, amplitude encoding reached 76.62% accuracy with 10 layers, compared to 64.68% for angle encoding. These findings highlight that deeper circuits enhance representational capacity and robustness. While amplitude encoding provides superior discriminative power, particularly beyond a certain circuit depth under noiseless conditions; angle encoding shows better performance with shallower depth in noisy condition.

To mitigate adversarial vulnerabilities, we benchmark defense mechanisms such as randomized encoding, quantum adversarial training, and engineered noise injection using depolarizing and amplitude damping channels. Overall, this work advances the development of trustworthy and scalable QML systems by integrating noise-aware design, encoding diversity, and adversarial robustness, paving the way toward secure, certifiably reliable quantum artificial intelligence.

*Corresponding author.