Performance Analysis of Quantum Variational Classifier under Adversarial Attacks

by

Mingyang Hao

A thesis submitted in total fulfillment for the
Master by Coursework Research Project

in the
Faculty of Engineering and Information Technology
School of Computing and Information Systems
THE UNIVERSITY OF MELBOURNE

June 2023

Abstract

Quantum machine learning (QML) is a rapidly evolving interdisciplinary field that leverages the powerful computational capabilities of quantum computers to address challenging machine-learning tasks. Building upon the seminal work by West et al. [1], which established benchmarks for adversarially robust quantum machine learning, this thesis delves into the intricate relationship between the architecture of Quantum Variational Classifiers (QVCs) and their resilience to adversarial attacks. By conducting an extensive series of experiments, we systematically explore the impact of various QVC architectures on the robustness of QML models against such attacks. Our findings reveal a state-of-art QVC architecture which leverage classification accuarcy, efficiency, and resistance to adversarial attakcs, thereby providing critical insights into the design of more secure and reliable quantum machine learning models.

Declaration of Authorship

I, Mingyang Hao, declare that this thesis titled, 'Performance Analysis of Quantum Variational Classifier under Adversarial Attacks' and the work presented in it are my own. I certify that:

- this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person where due reference is not made in the text
- where necessary I have received clearance for this research from the University's Ethics Committee
- the thesis is less than 8000 words in length (excluding text in images, table, bibliographies and appendices)

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Prof. Sarah Monazam Erfani, for her unwavering support, guidance, and encouragement throughout the course of my thesis. Her extensive knowledge, insights, and enthusiasm have been invaluable in shaping and refining my work. I am truly grateful for her mentorship and the opportunities she has provided me.

I would also like to extend my heartfelt appreciation to my co-supervisor, Prof. Muhammad Usman, for his invaluable advice, constructive feedback, and patient guidance. His expertise and dedication have been crucial in helping me overcome numerous challenges and ensuring the quality of my research.

I am immensely grateful to Maxwell T. West for his assistance in tackling concrete problems I encountered during my research. His technical expertise, willingness to help, and timely support have been instrumental in addressing these issues and have significantly contributed to the success of my thesis.

Lastly, I would like to thank my family, friends, and colleagues for their continuous support, understanding, and encouragement throughout this journey. Their belief in me and my work has been a source of motivation and strength that has propelled me forward.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Quantum Computing, a paradigm that exploits quantum mechanical phenomena, such as superposition and entanglement, extends computational capacities beyond classical systems' boundaries [2]. Quantum algorithms, formed from sequences of quantum gates, can tackle problems deemed intractable for classical computers [2]. Simultaneously, Machine Learning, an artificial intelligence subset, designs algorithms capable of learning from data and making predictive decisions [3]. Machine learning models span supervised learning, unsupervised learning, and reinforcement learning, proving essential in areas such as image recognition [4], natural language processing [5], and regression [6].

The introduction of machine learning algorithms to quantum computing systems, commonly known as Quantum Machine Learning (QML), has opened up new possibilities for solving complex computational problems, including classification tasks [7, 8]. QML has sparked considerable interest due to the potential of quantum computation in handling complex tasks with unprecedented speed and efficiency [2]. Similar to classical machine learning models, QML models are prone to adversarial attacks, where small, intentionally crafted perturbations in the input data misguide the model's predictions [9, 10]. Adversarial robustness, the study of machine learning model's resilience against such attacks, has been well-explored in the classical domain [9–11]. However, the understanding of adversarial robustness within the context of QML is an emerging field [12, 13]. The first instances of adversarial attacks in quantum settings have been reported recently, establishing the importance of this area of research [14, 15].

Quantum variational classifiers (QVCs) are hybrid quantum-classical models that utilize parameterized quantum circuits and data encoding techniques to perform classification tasks [2]. QVCs are based on the concept of variational quantum algorithms, which

leverage classical optimization routines to calibrate the parameters of a quantum circuit [16]. QVCs have shown promise in distinguishing between different classes of data, especially when the data has quantum features that are inaccessible to classical models [1, 14]. Moreover, previous studies[1, 16] have demonstrated the potential supremacy of QVC as being more resilient to classical adversarial attacks, which are attacks targeting classical ML models. Figure 1.1 shows the general architecture of QVC. The initial quantum state is $|0\rangle$ for all qubits. The $R_y(\theta_n^m)$ gates represent rotations along the y-axis, with θ_n^m being the rotation angle for the nth qubit at the mth layer. The structure of the circuit allows for the creation and manipulation of quantum entanglement and superposition, fundamental resources in quantum computing. The circuit also illustrates how controlled operations such as CNOT and CZ are used to facilitate interactions between qubits.

1.2 Motivation

Pioneering work by West et al. (2022) [1] propelled our understanding of adversarial robustness within QML significantly forward, shedding light on how QVCs withstand both classical and quantum adversarial attacks. West et al. showcased the unique resilience of QVCs against classical attacks, which are known to deceive classical neural networks such as Convolutional Neural Networks (CNNs) and ResNet18. Their research revealed that, contrary to classical networks, QVCs were capable of learning more robust and effective features, therefore offering stronger resilience to classical adversarial attacks. Moreover, they delved into the area of adversarial training for QVCs, which, interestingly, exhibited minimal improvement against classical attacks despite somewhat bolstering their defense against quantum attacks. Additionally, they introduced a novel approach to attack detection that leverages the predictions from both quantum and classical networks. This strategy further emphasized the robustness of features learned by QVCs compared to classical networks, thereby enhancing their resistance to classical attacks.

Quantum Variational Classifiers (QVCs), despite their potential for revolutionizing machine learning with quantum computational advantages, face critical challenges that hinder their immediate large-scale deployment. One such issue is their sensitivity to quantum adversarial attacks [1]. While they exhibit stronger resilience to classical adversarial attacks than traditional Neural Networks (NNs) due to their ability to learn robust features, QVCs still demonstrate a significant vulnerability to quantum attacks. The adversaries exploiting the quantum mechanical nature of QVCs can craft perturbations that mislead these models, causing them to make erroneous predictions. This

susceptibility to quantum adversarial attacks poses a substantial threat to the security and integrity of QVC-based systems, necessitating further research and development of strategies to counter such attacks.

Another crucial challenge is the comparative performance of QVCs on standard datasets [1]. Despite their theoretical advantages, QVCs often display lower accuracy compared to traditional NNs when applied to conventional datasets. This lower performance may be attributed to several factors, including the inherent complexity of quantum computation, the limited number of qubits available on current quantum hardware, noise in quantum operations, and the inefficiency of certain encoding schemes. Moreover, the scalability issues and the high resource demands of quantum computation exacerbate this problem. As a result, although QVCs exhibit robustness to classical adversarial attacks, their applicability is currently limited due to these accuracy and performance issues.

QVCs offer promising prospects for Quantum Machine Learning, but they also face a number of difficulties. Besides the susceptibility to quantum adversarial attacks that West et al. have covered, other researchers have also identified challenges such as the high quantum resource requirements [17, 18] and the limited scalability [7, 19] of QVCs. These challenges hinder the widespread adoption and practical implementation of QML solutions.

1.3 Contribution

Our research embarks on a novel journey to build upon the ground-breaking work of West et al., diving deeper into the intricate connection between the architecture of Quantum Variational Classifiers (QVCs) and their resilience against adversarial attacks. We bring forward the hypothesis that particular configurations of entanglement gate topologies could foster enhanced performance and robustness in these classifiers, without undermining their efficiency. To ascertain this, our research targets answering the critical question:

> How do different architectures of Quantum Variational Classifiers (QVCs) impact their resistance to adversarial attacks, and what are the ramifications for model performance and computational efficiency?

We plan to evaluate this proposition through the lens of three key metrics - accuracy, training time, and adversarial robustness. These metrics serve as reliable indicators of the model's prediction capability, computational efficiency, and fortitude against adversarial attacks respectively. Our research aspires to unravel subtleties in the interplay

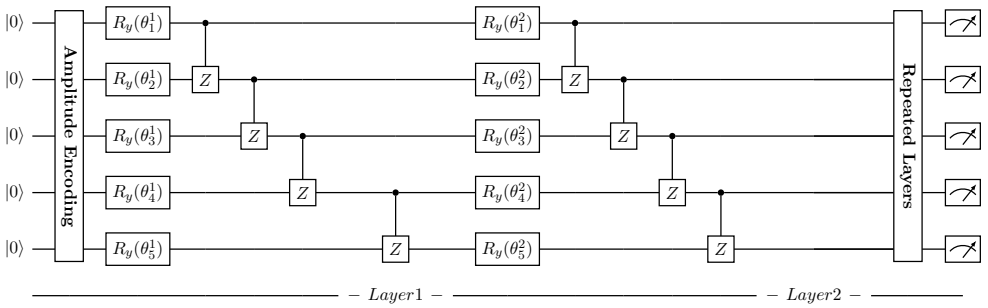


FIGURE 1.1: Schematic representation of the quantum circuit with Amplitude Encoding and Repeated Layers.

between these metrics and QVC architectures, thus providing valuable insights that could guide better design and optimization of QVCs. Ultimately, we aim to fortify the realm of quantum machine learning, thereby laying the foundation for highly efficient, secure, and adversarially robust quantum systems.

This study promises to fill the existing knowledge gap by executing an in-depth performance analysis of QVCs under adversarial attack scenarios. We aim to probe the vulnerability of QVCs to adversarial perturbations, scrutinize the implications on performance, and unveil potential countermeasures. Our aspiration is to make a meaningful contribution to the understanding of adversarial robustness in quantum machine learning, thereby aiding the development of more secure, reliable quantum information processing systems. We will leverage insights from classical machine learning about adversarial attacks, particularly focusing on adversarially robust deep learning models [3, 10]. Our intention is to adapt these principles for the quantum domain. Moreover, our analysis will be grounded in recent works that shed light on the robustness of quantum classifiers [20, 21]. By amalgamating these insights, our study seeks to offer an exhaustive overview of the state-of-the-art in adversarial robustness of quantum machine learning, thereby sparking new research trajectories in the field. The novelty and significance of our work will unquestionably be of keen interest to others in the field. The implications of our research extend beyond QVC itself. Resolving the challenges faced by QVCs will have a profound impact on various fields that rely on quantum computing, such as optimization, drug discovery, and materials science. By addressing these obstacles head-on, we can unlock the full potential of quantum computing for solving real-world problems.

1.4 Thesis Structure

This thesis is organized into five cohesive chapters, each serving a distinct purpose:

- **Chapter 2: Background and Literature Review**
  This chapter provides an in-depth understanding of the necessary terminology and related concepts within the research domain. A comprehensive literature review is conducted to contextualize the study within the current research landscape, highlighting gaps and establishing the need for the current research.

- **Chapter 3: Experimental Design and Methodology**
  This chapter delves into the specifics of the experiment setup. It details the choice of dataset, the encoding technique used, and the different Quantum Variational Classifier (QVC) architectures chosen for the study. It also explains the metrics used to evaluate the models and the method followed to generate adversarial examples.

- **Chapter 4: Results and Analysis**
  This chapter presents the findings from the experiment. It provides an in-depth comparison of the different QVC architectures, discussing their performance in terms of accuracy, computation time, and robustness against adversarial attacks. It also interprets these findings in relation to the research question, the hypotheses, and the existing literature.

- **Chapter 5: Conclusions and Future Work**
  This chapter serves as the culmination of the thesis. It encapsulates the key findings of the study, discusses their implications for the field of Quantum Machine Learning, and acknowledges the limitations of the study. It also suggests potential directions for future research in this area, focusing on strategies to further improve the robustness and performance of QVCs.

Chapter 2

Background

2.1 Key Concepts and Terminology

Before delving into the literature, it is essential to elucidate the key concepts and terminology used throughout this review. Quantum machine learning (QML) pertains to the application of quantum computing techniques to machine learning tasks, which potentially leads to advantages in computational efficiency and accuracy [14, 20]. Adversarial attacks encompass the manipulation of input data with the intent of misleading machine learning models, typically achieved by introducing small perturbations that, while imperceptible to humans, cause the model to generate incorrect predictions [13, 14]. Consequently, adversarial robustness denotes a model's ability to preserve its performance in the face of such attacks [13, 20]. Quantum variational circuits (QVCs) are parameterized quantum circuits utilized as foundational elements for QML models [14], and their architecture may influence the model's performance and robustness [1]. Adversarial attacks can be categorized into distinct types based on the adversary's knowledge of the model, the attack's objective, and the perturbation method employed [22, 23]. For instance, white-box attacks presuppose that the adversary has complete access to the model's parameters and architecture, while black-box attacks postulate that the adversary only has access to the model's input-output behavior [24, 25]. Similarly, targeted attacks strive to alter the prediction of a specific input to a desired class, whereas untargeted attacks endeavor to change the prediction of any input to any incorrect class [22, 26]. Furthermore, adversarial perturbations can be incorporated into the input data through various means, including adding noise, cropping, rotating, or transforming the data [23].

2.2 Adversarial Attacks on Classical and Quantum Machine Learning Models

To formalize the problem of adversarial attacks and adversarial robustness, we introduce equation (2.1), which defines the optimal perturbation that minimizes the loss function of the ML model under a constraint on the perturbation norm. It is used to generate adversarial examples that can fool ML models.

Equation 1. AML objective function:

$$\min_{\theta}[L(f(x;\theta),y),|\delta| \leq \epsilon] \tag{2.1}$$

Here, f denotes the ML model, θ symbolizes the parameters of the machine learning model, such as weights and biases in a neural network, x represents the input data, y is the true label, L signifies the loss function, δ corresponds to the adversarial perturbation, and ϵ refers to the attack strength, which encompasses the set of allowable perturbations. The objective of adversarial machine learning is to find the parameter that minimizes the loss function with perturbed input.

Although both classical and quantum ML models are vulnerable to adversarial attacks [27], they may have different characteristics and effects depending on the nature of the data and the model [15]. For classical ML models, such as Convolutional Neuron Networks (CNN), adversarial attacks typically exploit the high-dimensional and nonlinear features of the data and the model [18], which make them sensitive to small perturbations that are imperceptible to humans [16]. For quantum ML models, such as QCNNs and QVCs [18], adversarial attacks may exploit the quantum properties of the data and the model [15], such as superposition [28], entanglement [16], and measurement [15]. For example, applying a small rotation to a quantum state can cause a QNN to misclassify it as another state [15]. However, QML models may also have some advantages over classical ML models in terms of adversarial robustness [27], such as learning features that are not detected by classical ML models [18] or using quantum error correction techniques [16].

To generate adversarial inputs, two common methods are PGD (projected gradient descent)[10] and FGSM (fast gradient sign method)[9]. Both use the sign of the loss gradient to perturb the input, but PGD iterates and projects while FGSM does not. To defend against them, one strategy is adversarial training, which trains the model on adversarial examples. However, PGD adversarial training is more robust but slower than FGSM adversarial training. This gap is due to the large curvature along the FGSM perturbation direction. Several studies have proposed to modify FGSM adversarial training with techniques such as noise, multi-step, or curvature regularization. These techniques aim to reduce the curvature and improve the robustness. PGD and FGSM are essential

tools for adversarial research. They have trade-offs between efficiency and effectiveness. Bridging the gap between PGD and FGSM adversarial training is an active and challenging research area, especially for high-dimensional problems and complex models. We define FGSM and PGD as follows:

Equation 2. Fast Gradient Sign Method (FGSM):

$$\delta = x + \epsilon \operatorname{sgn}(\nabla_x L(\theta, x, y)) \tag{2.2}$$

Equation 2 illustrates Fast Gradient Sign Method (FGSM)[9], which adds an adversarial perturbation vector δ to the original input x. The adversarial perturbation vector has the same sign as the gradient of the loss function $L(\theta, x, y)$, where θ are the model parameters and y is the true label. The magnitude of the adversarial perturbation vector is controlled by an attack strength parameter ϵ. The FGSM is a one-step method that generates adversarial examples quickly and efficiently.

Equation 3. Projected Gradient Descent (PGD):

$$x^{t+1} = \Pi_X(x^t + \alpha \cdot \operatorname{sign}(\nabla_x L(\theta, x^t, y))) \tag{2.3}$$

Equation 3 demonstrates Projected Gradient Descent (PGD)[10], which iteratively updates the input x by adding a small step α along the direction of the gradient sign. The updated input is then projected back to a feasible set X, which is usually a bounded region around the original input. PGD is a multi-step method that generates more powerful adversarial examples than FGSM but at a higher computational cost.

2.3 Quantum Entanglement

Quantum entanglement is one of the cornerstones of quantum mechanics. It describes a phenomenon where two or more particles become linked and instantaneously affect each other's states, regardless of the distance separating them [29]. This phenomenon, famously described by Einstein as "spooky action at a distance," forms the core of quantum computing and QML [30]. One of the key implications of entanglement in QML is the ability to process information in fundamentally novel ways [31]. This is primarily due to the superior data encoding capabilities of quantum systems, allowing them to handle complex, high-dimensional data more efficiently than classical systems [29, 30]. When we look at Quantum Variational Classifiers (QVCs) specifically, entanglement plays a crucial role in creating more complex and nuanced decision boundaries [31]. This is because QVCs use parameterized quantum circuits [29], where the parameters are trained using classical optimization techniques such as gradient descent. In these circuits, entanglement gates play the role of creating and enhancing correlations between

qubits, which are used to generate complex patterns in the data representation. This complexity can potentially lead to better classification performance compared to classical machine learning algorithms [29, 30].

Furthermore, entanglement can contribute to quantum speed-up, a potential advantage where quantum systems outperform classical systems for certain tasks [29, 30]. Some QML algorithms that exploit quantum entanglement may potentially solve certain types of problems exponentially faster than their classical counterparts. However, it's important to note that such quantum speed-up in machine learning is a theoretical promise and is subject to ongoing research [30].

Entanglement also brings new challenges. In a practical setting, maintaining quantum entanglement is difficult due to environmental interference, a problem known as decoherence [29, 32]. Moreover, fully utilizing entanglement in large-scale quantum systems is a significant challenge due to the limited number of available qubits and the complexity of creating and managing entangled states [32]. These are active areas of research in the development of QML and quantum computing as a whole [30].

2.4 The variants of QVCs and their limitations

Different QVC architectures have been studied for different QML tasks, such as classification and optimization [18, 19]. These studies mainly focused on the efficiency and accuracy of the models, rather than their robustness to adversarial attacks. Moreover, some works have examined the generalization abilities of QML models [7, 8], indicating that the connection between QVC architecture and generalization may offer some clues to their adversarial robustness.

Schuld et al. (2020)[33] present a low-depth variational quantum algorithm for supervised learning that encodes classical data into quantum states and classifies them using observables. They show that the algorithm outperforms or matches classical methods with fewer parameters. They also propose a quantum-classical training scheme that estimates gradients by running slightly modified circuits. They evaluate the algorithm on two synthetic datasets and demonstrate its ability to learn complex nonlinear patterns. They analyze its noise sensitivity, introduce a quantum dropout regularization, and illustrate quantum gates as linear layers of a neural network.

These QVC architectures illustrate the potential of QML for solving complex problems that require high-dimensional feature spaces and nonlinear transformations. However, they also face challenges such as scalability, noise, and generalization, which require further research and development.

West et al. [1]presented an intriguing finding in their work. They showed that one of the reasons why QVCs can withstand adversarial attacks better than conventional CNNs is

that they learn features that are more discriminative and robust. They demonstrated this by comparing the feature maps and found that QVCs can capture more complex and diverse patterns than CNNs, which makes them less sensitive to small perturbations in the input data.

2.5 Encoding Techniques

In Quantum Machine Learning (QML), encoding techniques are employed to transfer classical data into quantum states [34], such process is also known as state preparation or data embedding [35]. This allows quantum algorithms to exploit the unique properties of quantum states to process information [36]. Two popular encoding techniques in QML are amplitude encoding and phase encoding [34].

Amplitude encoding, also known as wavefunction encoding [36], is a common technique where classical data is encoded into the amplitudes of a quantum state [37]. For an n-dimensional data vector, amplitude encoding can encode the data into $log(n)$ qubits [34]. The power of amplitude encoding lies in its compact representation which allows for exponential dimensionality reduction [37].

Given a normalized data vector $|\vec{x}\rangle = (x_1, x_2, \ldots, x_n)$, we can encode this into the amplitude of a quantum state as follows:

Equation 2.4. Amplitude Encoding:

$$|\psi\rangle = \sum_{i=1}^{2^n} x_i |i\rangle \tag{2.4}$$

Equation 2.4 represents amplitude encoding where each element $x_i belongston$ of the original data vector $|\vec{x}\rangle$ is encoded into the amplitude of the corresponding basis state $|i\rangle$ and formed the Hilbert Space. This can be achieved by applying a unitary transformation U to the initial state $|0\rangle^{\otimes \log n}$ such that $U|0\rangle^{\otimes \log n} = |\psi\rangle$.

Phase encoding, on the other hand, embeds the information of the classical data into the relative phases of a quantum state [36]. While this method can carry a large amount of information [34], extracting the encoded information can be quite challenging [36]. The measurement of a quantum state usually gives information about the amplitudes but not the relative phases [34]. Therefore, additional quantum operations are often necessary to translate phase information into measurable quantities [36].

Given a data vector $|\vec{x}\rangle = (x_1, x_2, \ldots, x_n)$, we can encode this into the phase of a quantum state as follows:

Equation 2.5. Phase Encoding:

$$|\psi\rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e^{2\pi i x_i} |i\rangle \tag{2.5}$$

Here, each basis state $|i\rangle$ is associated with a data point x_i from the input vector x. Each data point x_i is transformed into a phase value $2\pi x_i$ for the corresponding basis state. The normalization factor $\frac{1}{\sqrt{n}}$ ensures the total probabilities sum to one. This method encodes information into the phase of the quantum state, providing a unique representation of the original data in the quantum system.

The choice between encoding data into real or complex domains is another design problem. It generally depends on the nature of the data and the problem at hand. Real encoding refers to encoding techniques that only use real numbers, while complex encoding involves the use of both real and complex numbers [34, 35]. Real encoding has the advantage of simplicity and is generally more intuitive as it corresponds directly to our classical notion of numerical data [34]. However, it might not fully exploit the quantum advantage [35], since it does not utilize the complex nature of quantum states [36].

On the other hand, complex encoding can potentially capture more information about the data, as it employs both the amplitude and phase of the quantum state [35, 36]. However, it can be more challenging to implement and interpret, particularly when it comes to retrieving the encoded information [34, 35].

While both encoding techniques have their own merits and demerits, it's crucial to choose the appropriate encoding technique based on the specific requirements of the task and the computational resources available [36]. Furthermore, understanding the trade-offs between different encoding techniques is an active area of research in quantum machine learning [35].

Chapter 3

Experiment Setup

In the pursuit of understanding the impact of QVCs' architectures on their resistance to adversarial attacks, we set out to conduct a series of comprehensive experiments. Central to our investigation is the architecture of entanglement gates within QVCs as shown in Figure 3.1. By evaluating these architectural variants, we aim to shed light on the trade-offs and benefits they bring in the context of adversarial resilience, model performance, and computational efficiency.

3.1 Dataset and Encoding Technique

The Fashion-MNIST (FMNIST) dataset, integral to our experimental design, offers a number of advantages over other commonly used datasets such as MNIST, CIFAR10, and CELEB-A. FMNIST is a modern replacement for the traditional MNIST dataset of handwritten digits [38]. Both datasets consist of 28x28 grayscale images and share a similar scale and data structure, which facilitates a direct comparison between them. However, FMNIST presents a significantly more challenging task as it features a variety of diverse fashion items rather than just digits. The higher intra-class variation within FMNIST stimulates the classifiers to learn more complex and discriminative features to correctly classify the images [39]. In addition, FMNIST has the advantage of being more relevant for contemporary machine learning tasks, as object recognition is generally more demanding than digit recognition.
On the other hand, CIFAR10 contains color images of ten different classes of objects [40]. Compared to FMNIST, CIFAR10 increases the dimensionality and complexity of the input data. This higher complexity does not align well with the current capabilities of quantum computation, which are not yet suited to handle the high dimensionality and complexity of color images [41]. CELEB-A is another dataset that poses a

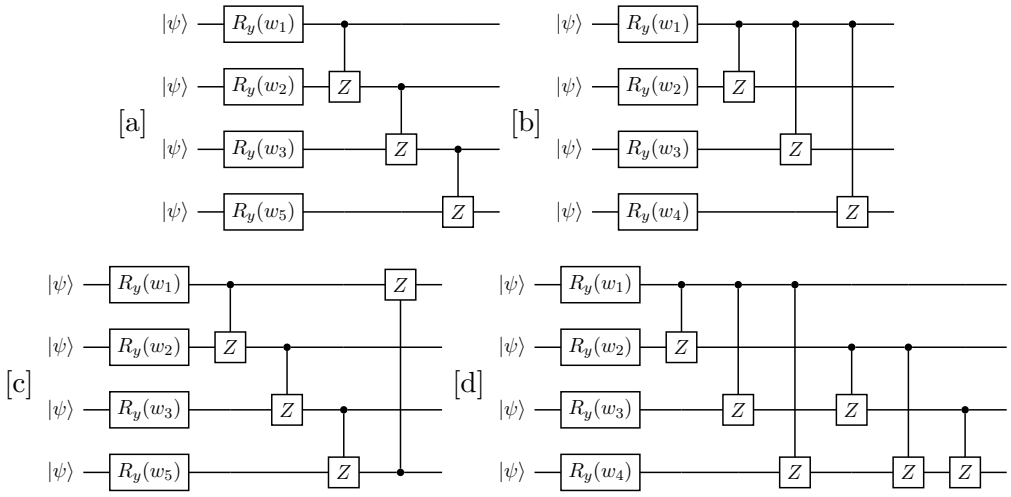


FIGURE 3.1: Architectures of QVCs with different gates and controls.

formidable challenge for machine learning models. It consists of over 200,000 high-resolution celebrity images with various attributes [42]. The varied and complex nature of CELEB-A data makes it unfeasible for direct quantum processing at the current stage of quantum computing.

Therefore, we chose the FMNIST dataset as it stands out as a practical yet challenging option for testing QVCs. It offers a level of complexity that is considerably higher than MNIST but more manageable than CIFAR10 or CELEB-A, striking a balance that allows for a meaningful and feasible evaluation of the robustness and performance of different QVC architectures. Future research would like to explore various datasets that can further challenge and improve QVCs.

In our study, we use real amplitude encoding due to its compatibility, simplicity, and well-understood relationship with adversarial robustness. Firstly, this form of encoding aligns well with the nature of our datasets, which predominantly comprise real-valued features. This obviates the need for additional preprocessing steps, thereby streamlining the process. Secondly, real amplitude encoding is computationally less demanding than complex encoding. It doesn't involve complex numbers, which would significantly increase the computational burden and potentially impact the scalability of our experiments with QVCs. Lastly, the effects of real amplitude encoding on the adversarial robustness of QVCs are more straightforward and well-studied, aligning well with the focus of our research. While the implications of complex encoding are worth exploring, it adds an element of uncertainty which we aim to avoid in this initial phase of our investigation into QVC architectures and their resistance to adversarial attacks. However, we remain open to analyzing the impact of complex encoding in future research.

3.2 Architectures of QVC

Figure 3.1 provides a visual representation of the distinct Quantum Variational Classifier (QVC) architectures explored in our study. We subjected each of these architectural configurations to extensive testing over QVCs composed of 50, 100, and 200 layers, respectively. These trials were conducted in an error-free environment using PennyLane [43] to purely evaluate the performance and robustness characteristics intrinsic to the QVC architectures. Later in this section, we delve into hypothetical scenarios accounting for the presence of noise, thereby discussing the probable impact noise would have on the performance and reliability of these QVC structures.

Chain topology : CZ gates are applied between each pair of adjacent qubits, creating a kind of "chain" of entanglement. This can be represented by the product of CZ gates:

$$U_{\text{chain}} = \prod_{i=1}^{n-1} CZ_{i,i+1} \tag{a}$$

This chain of entanglement allows for information to flow from one end of the qubit register to the other, but only through intermediate qubits. The chain topology is the most natural for one-dimensional systems or for quantum devices where only local interactions are allowed.

Star topology: A single central qubit is entangled with all other qubits. This can be represented by:

$$U_{\text{star}} = \prod_{i=1}^{n} CZ_{0,i} \tag{b}$$

The star topology creates a 'hub and spoke' model of entanglement, where the central qubit can quickly share information with all other qubits, but interactions between non-central qubits always involve the central qubit. This topology could be particularly useful for problems that have a clear 'center' or 'hub' qubit that should interact more strongly with all other qubits.

Circle topology: Similar to the chain, but with an additional entanglement between the first and the last qubit:

$$U_{\text{circle}} = \left(\prod_{i=1}^{n-1} CZ_{i,i+1} \right) \cdot CZ_{n,1} \tag{c}$$

This topology creates a "loop" of entanglement, allowing for information to flow around the circle. This topology could be beneficial for problems with periodic or cyclic structure, or for quantum devices where the qubits are physically arranged in a ring.

All-to-all topology: Every qubit is entangled with every other qubit. This can be represented by:

$$U_{\text{all-to-all}} = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} CZ_{i,j} \tag{d}$$

This results in a fully-connected graph of entanglement, allowing for direct information exchange between any pair of qubits. This topology is the most expressive but also the most resource-intensive, requiring a quadratic number of gates in the number of qubits.

The choice of topology can directly influence the entanglement level of qubits, expressivity of the quantum classifier, i.e., the range of functions it can represent. For instance, an all-to-all topology could in principle represent more complex functions than a chain topology, due to the larger number of interactions between qubits. On the other hand, the chain and circle topologies, while less expressive, are more hardware-friendly as they require fewer gates and only nearest-neighbor interactions. As for star topologies, it would be more effective when a particular central qubit is learning robust feature.
Apart from aforementioned factors, the selection of topology critically influences the performance of QVCs in Noisy Intermediate-Scale Quantum (NISQ) devices[17]. Different gate topologies differ in their noise sensitivity, which implies that certain noise types can affect some gates more significantly than others [44].
The star topology's vulnerability to noise can be viewed through the lens of quantum error propagation[45]. With each qubit entangled to a central qubit, the impact of noise-induced errors on the central qubit can be severe. Mathematically, given an error ε on the central qubit, it could potentially impact the overall state of the system, leading to an error proliferation that is proportional to the number of qubits, n, connected to the central qubit. Therefore, the severity of an error on the central qubit could be written as $O(n\varepsilon)$, illustrating the large potential impact of a single error[46]. In comparison, chain and circle topologies exhibit a different error propagation characteristic. The effect of an error is primarily localized to adjacent qubits, suggesting that errors propagate linearly along the chain or around the circle. Hence, a local error could be isolated and its impact mitigated[45].
Lastly, the all-to-all topology provides maximum entanglement and also brings a high risk of error propagation. As each qubit is connected to all others, a single error could, in principle, impact all qubits. Thus, it may lead to an error impact on the order of $O(n\varepsilon)$ akin to the star topology[46]. However due to the increasing number of gates involved in this topology, it introduce a higher probability of errors occurring, which can overall increase the system's vulnerability to noise[47]. Thus, the choice of topology for a QVC presents a trade-off between expressiveness and resilience to noise. This underlines the importance of a comprehensive understanding of the noise characteristics of

the quantum hardware and the robust error mitigation strategies when designing QVCs for deployment on NISQ devices[45].

3.3 Performance Evaluation Metrics

As part of our comprehensive study aimed at evaluating the performance of diverse QVC architectures, we relied on three key metrics that are essential to understanding and comparing the effectiveness of machine learning models: computational time, accuracy, and robustness against adversarial attacks.

Computational Time: Computational efficiency is crucial in practical applications of machine learning models. Therefore, we meticulously recorded the time consumed during both the training and inference stages of our QVC models. This allowed us to gauge the efficiency of various QVC architectures and estimate their viability in real-world settings, especially considering the significant computational resource limitations currently inherent to quantum computing.

Predictive Accuracy: A vital measure of the effectiveness of any machine learning model is its accuracy in predicting outputs for unseen data. To assess this, we compared the predicted labels by the QVC models against the actual labels of our dataset. This accuracy metric helped quantify how well each model generalized the learning from the training data, thus providing a reliable measure for performance comparison.

Robustness Against Adversarial Attacks: In the realm of machine learning, a model's resilience to adversarial attacks is a critical factor, especially considering the growing concerns about model security and integrity. For this purpose, we employed a white-box Projected Gradient Descent (PGD)[10] attack mechanism on all the models under study, which included classical models such as Convolutional Neural Network (ConvNet) and ResNet18 [3], in addition to the QVCs. We selected a wide range of epsilon values as shown in Figure 3.2 to generate a diverse set of adversarial examples. These examples were subsequently used to test the resilience of the other models. This cross-evaluation approach allowed us not only to analyze each model's individual resilience but also to investigate their ability to withstand attacks intended for different models. This comprehensive approach to robustness evaluation provided us with a deeper understanding of adversarial resilience across the different QVC architectures and classical models (ConvNet and ResNet18).

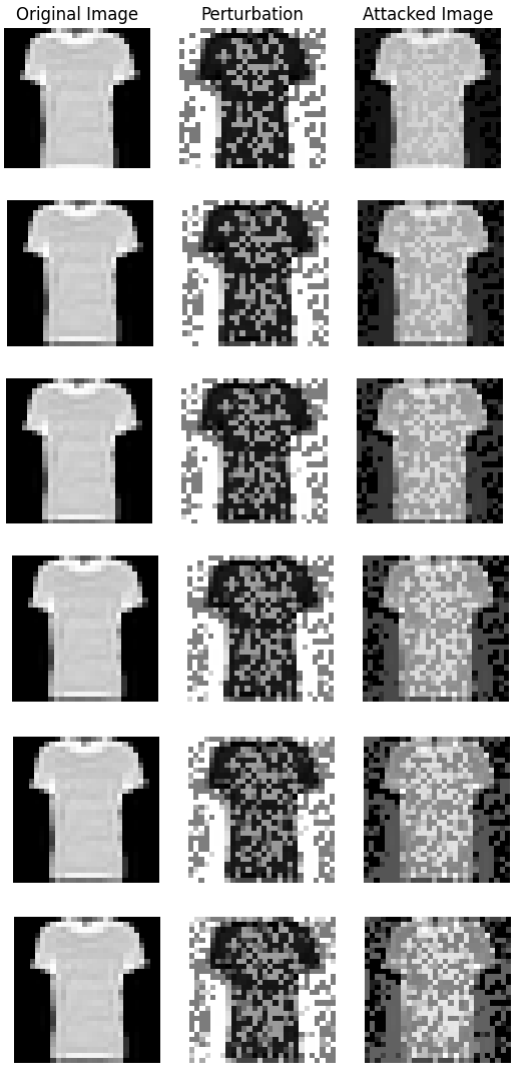


FIGURE 3.2: A comparison of the white-box PGD attacks with varying epsilon values on QVC50 with star topology.

3.4 Classical Model Architectures

The architecture of classical models ConvNet and ResNet18, analogs to those presented by West et al., serving as reference points against which the QVCs were benchmarked. The ConvNet architecture consisted of three convolutional layers, each equipped with 3x3 filters, and designed to contain 64, 128, and 256 feature maps respectively. The purpose of this design was to extract varying degrees of features from the input data. To further enhance feature extraction and reduce dimensionality, 2x2 max pooling was applied after each convolutional layer. The activation function used throughout these layers was the Rectified Linear Unit (ReLU), known for its efficiency in handling the

vanishing gradient problem. Following the convolutional layers, the architecture also integrated two fully connected layers, which again employed the ReLu activation function, contributing to the overall network's learning capability. ResNet18, on the other hand, maintained its original architecture as depicted in the referenced literature.

For both architectures, the training process was conducted using the Adam optimizer, known for its efficient adaptive learning rate management. Furthermore, the cross-entropy loss function, a standard choice for multi-class classification tasks from the PyTorch library[48], was utilized to guide the optimization process. These tools helped improve the training efficiency and model performance on our datasets.

Chapter 4

Results and Discussion

4.1 Model Accuracy and Efficiency

	Time per iteration (s)				Accuracy			
	Chain	**Circle**	**Star**	**ALL**	**Chain**	**Circle**	**Star**	**All**
QVC50	179.0	174.6	165.8	217.5	0.650	0.660	0.646	0.678
QVC100	361.0	342.6	330.9	422.6	0.686	0.706	0.690	0.710
QVC200	755.5	719.9	690.8	921.4	0.724	0.714	0.726	0.720

TABLE 4.1: Time per iteration (s) and Accuracy for different topologies and QVC sizes. The lowest time per iteration and highest accuracy per QVC size are highlighted.

Table 4.1 illustrates a comprehensive comparison of different topologies—Chain, Circle, Star, and All (all-to-all)—in terms of time per iteration and accuracy across different QVC sizes. Notably, the Star topology consistently exhibits the lowest time per iteration across all sizes, indicating superior computational efficiency. A tangible factor is its less complex entanglement structure that necessitates fewer computational resources. It is worth noting that while the Star and Chain topologies use the same number of gates, their performance can differ due to the distinct entanglement structures that they create. The Star topology centralizes the interactions around a single, central qubit, effectively creating a shorter maximum path length for information propagation compared to the Chain topology. In quantum systems, information propagation and entanglement patterns can significantly affect computation dynamics, influencing the speed of convergence during the optimization of the variational circuit [43]. This could be a potential reason why the Star topology tends to be more efficient than the Chain topology in these experiments. However, it's also important to highlight that the specific hardware or simulator used for these experiments might introduce nuances in gate execution times, which could additionally contribute to the observed differences in computational efficiency. Future

research could involve a detailed examination of these factors to better understand their influence on the overall performance of QVC architectures.

On the other hand, the accuracy varies across architectures and sizes, with no single topology demonstrating consistent superiority. This suggests that the QVC's representational capacity, which directly influences its accuracy, is affected by more intricate interactions between the architectural structure and the specific problem at hand. Therefore, selecting an optimal QVC architecture may require careful consideration of both the computational resources available and the specific requirements of the task. Further investigations could delve into understanding these intricate interactions to enhance both efficiency and performance of QVCs for specific applications.

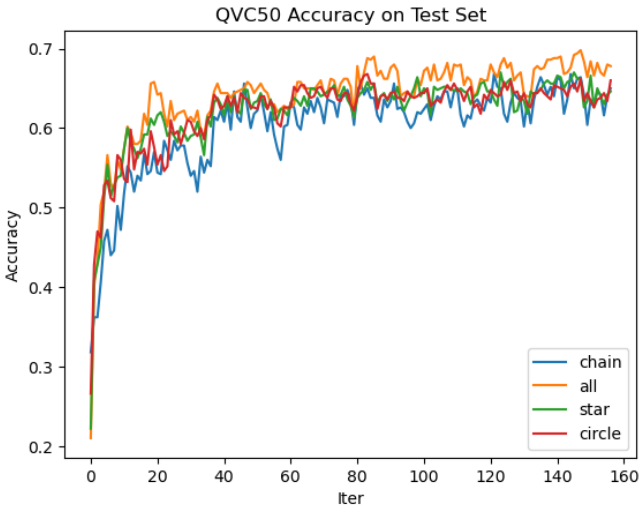
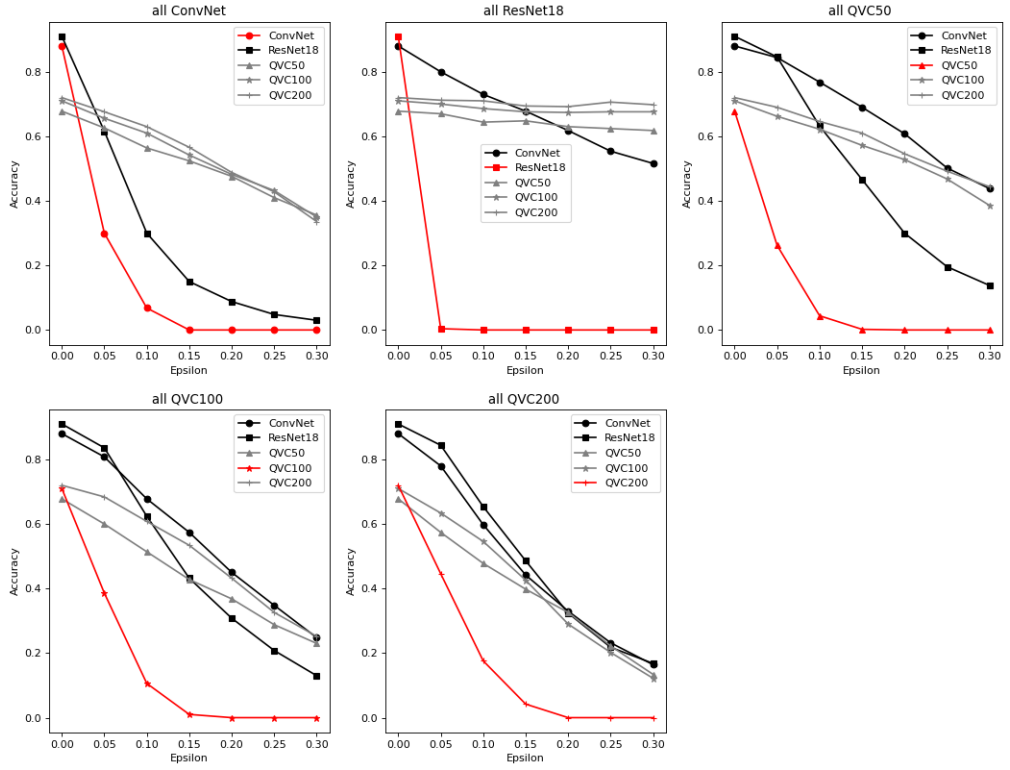


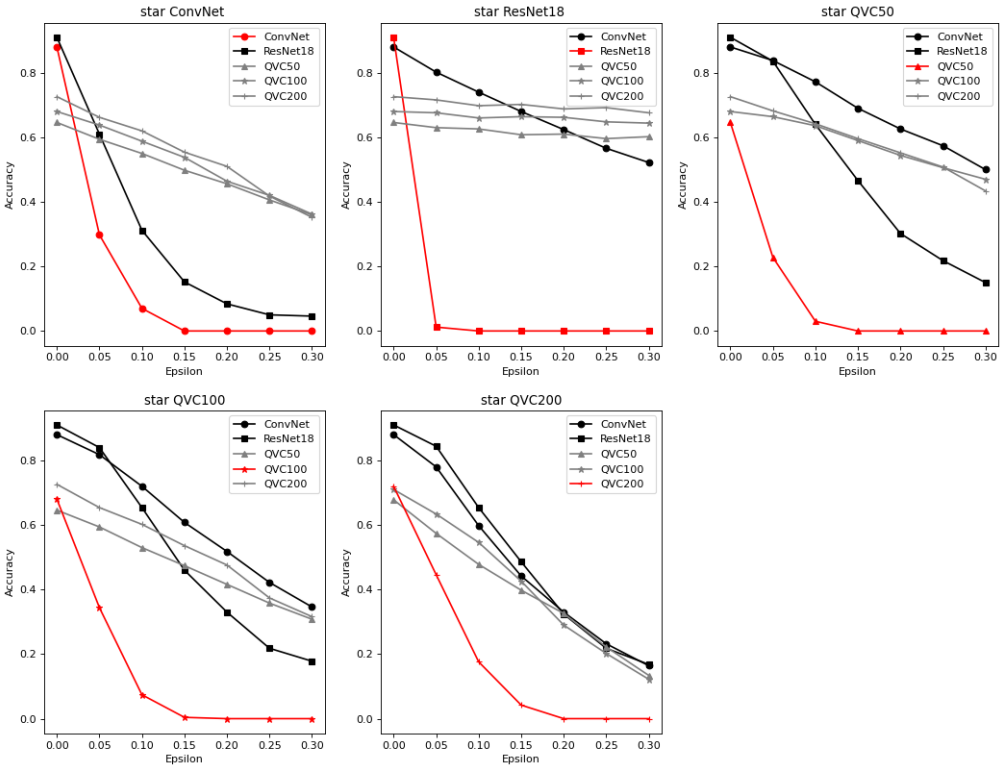
FIGURE 4.1: Comparison of Test Set Accuracy Over Iterations for Various QVC50 Topologies

Figure 4.1 illustrates the comparison of various QVC50 topologies' accuracies on the test set. Exhibiting a steadily improving average accuracy, reaching a peak of 0.67, the star architecture showcases significant learning efficiency and stability. Its median accuracy at 0.638 further underscores its reliability. Comparatively, the all-to-all topology also offers strong performance, boasting the highest mean and median accuracies of 0.63 and 0.65 respectively, yet the variance in its accuracy could be indicative of a lack of consistency.

Conversely, the circle topology maintains good stability with less variability in accuracy, suggesting a steady learning process, even though its mean accuracy of 0.62 isn't the highest. The chain topology, despite having a peak accuracy of 0.668, reveals potential issues with consistency and stability due to its greater variability and lower mean accuracy. This result shows that while each topology has its strengths, the star topology emerges as the most effective in the QVC50 experiment, balancing steady improvement and performance stability. Nonetheless, the suitability of a given topology should be



(a)



(b)

FIGURE 4.2: Transfer attack of all-to-all topology (**a**) and star topology (**b**) with baseline models

evaluated based on the specific requirements of different learning tasks and environments.

4.2 Adversarial Robustness

Figure 4.2 demonstrates the adversarial training of star and all-to-all topologies together with the base line model (ConvNet and ResNet18). From the adversarial attacks experiment, it is evident that QVC models demonstrate robustness with varying degrees of success across the topologies. Among these, QVC100 and QVC200 under the star topology consistently maintain higher accuracies, indicating their resilience against adversarial attacks. In particular, QVC200 shows persistent performance, maintaining accuracy above 0.35 across all sets of adversarial attacks. Contrarily, ConvNet and ResNet18 reveal varied susceptibility to these attacks. ConvNet generally maintains high initial accuracy but sees a drastic decline when subjected to stronger adversarial perturbations, showing weak resilience. ResNet18 demonstrates a similar pattern, though it also experiences instances of absolute failure, as seen in the third and fifth sets. While both QVC50 and QVC100 have occasional collapses, they recover in subsequent tests, suggesting an overall greater stability compared to ConvNet and ResNet18. This analysis highlights the strength of the QVC models, particularly QVC200 under the star topology, in terms of robustness against adversarial attacks.

In contrast, star topology presents intriguing patterns. ConvNet and ResNet18 maintain their strong performance, although with some variation. Intriguingly, ResNet18's accuracy drops to almost zero in some iterations. For the QVC models, QVC50 displays a consistent performance much like in the all-to-all topology. However, the peak performance of QVC100 and QVC200 surpasses QVC50 and remains steady longer before dropping. This behavior may indicate a better ability to generalize in early iterations.

From this experiment, all the topologies showcase similar performance trends. While ConvNet and ResNet18 consistently perform well, their stability varies across iterations. QVC models in the star topology seem to maintain peak performance for longer, suggesting a more efficient learning process.

In the transfer attack experiment across different QVC50 topologies as shown in figure 4.3, there is no significant quantitative difference in terms of their performances under adversarial attacks. Despite subtle variations across iterations and differing adversarial intensities, different architectures exhibit similar trends of decreasing accuracy as the strength of adversarial attacks increases.

Relating this observation to the previous adversarial attacks analysis, we found that the choice of topology does not substantially affect the adversarial robustness of QVC. Irrespective of whether the model follows a chain, all-to-all, star, or circle topology,

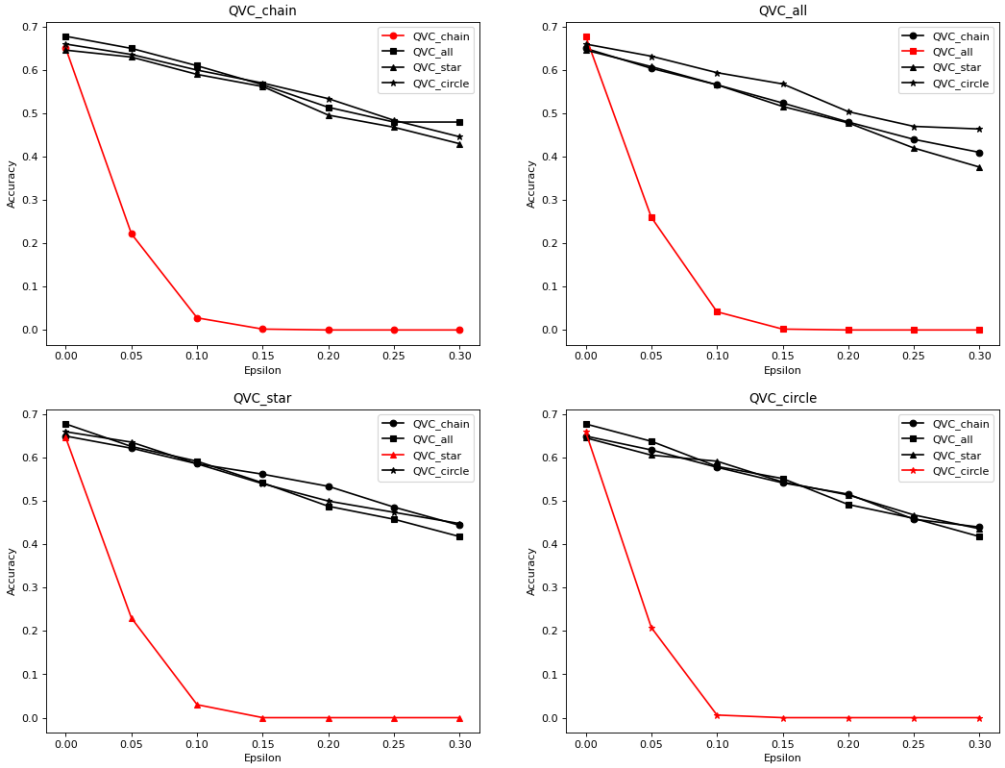


FIGURE 4.3: Transfer attack across different topologies of QVC50

QVC tends to demonstrate a comparable degree of vulnerability under both quantum and classical adversarial attacks. Therefore, it can be concluded that, while the QVCs overall maintain competitive performance and stability against adversarial attacks, the selection of a specific topology does not provide a distinct advantage in enhancing the robustness of these models against such attacks.

4.3 Adversarial Perturbation Analysis

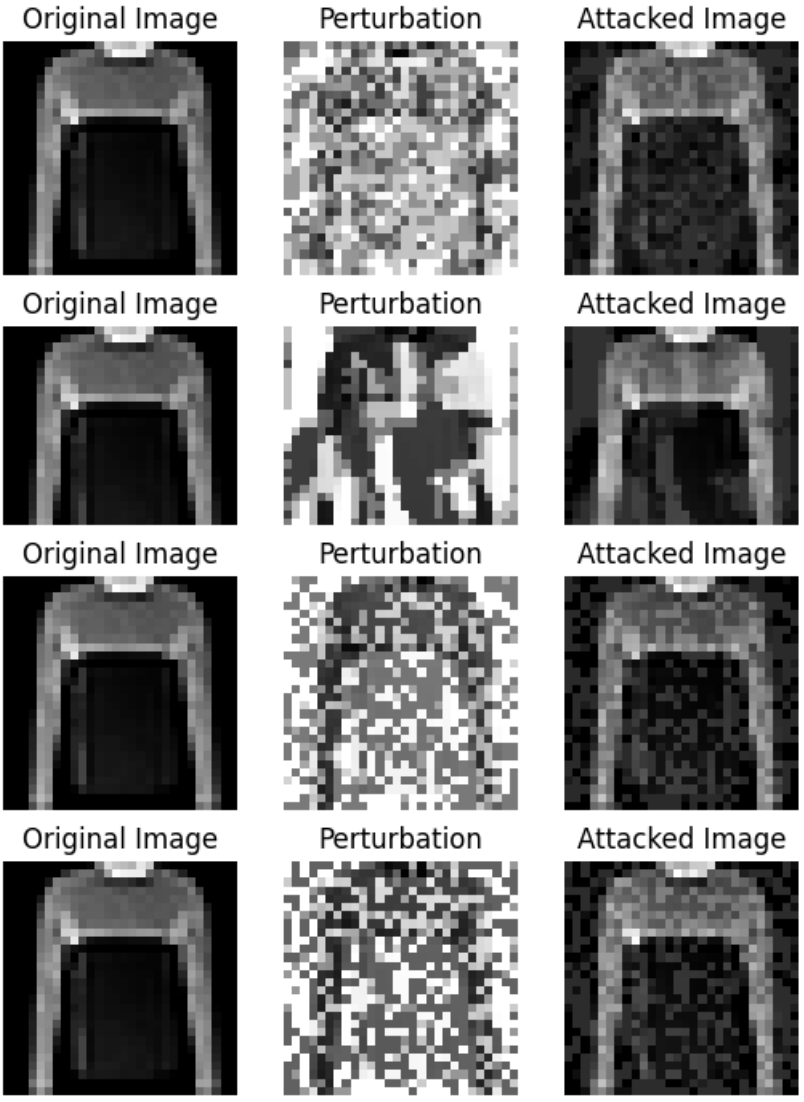


FIGURE 4.4: Rerturbation images created by a white-box PGD attack on ResNet (1), ConvNet (2), and QVC100 models using both Star (3) and All-to-all (4) topologies.

Figure 4.4 offers a visual representation of the perturbation images created by a white-box PGD attack on ResNet, ConvNet, and QVC100 models using both Star and All-to-all topologies. By comparing the images, one can discern differences in the response to adversarial perturbations by these various architectures.

In the case of ConvNet, the perturbations seem to introduce what appears to be random noise. This random-like pattern suggests that the features learned by ConvNet may not be sufficiently robust or discriminative, which aligns with previous studies [1, 9]. Although ResNet seems to capture certain level of details, it still generally looks random. These studies have indicated that high-dimensional input spaces, like those used by ConvNets, may be vulnerable to adversarial manipulations that seem like random noise to

the human eye but are significantly disruptive to the model's functionality.

On the contrary, perturbations resulting from attacks on QVCs present an entirely different picture. The QVCs seem to be adjusting the color consistency across different areas of the shirt in the image. This adaptation indicates that QVCs might be capable of learning and retaining more substantial, meaningful features from the data, which could make them more resilient to adversarial attacks. This observation supports the insights shared by West et al. [1], suggesting that QVCs are capable of learning more robust and informative features compared to traditional classical models.

In addition to the aforementioned observations, our study brought to light intriguing findings about the similarities among different QVC topologies. Despite variances in their structural designs, our analysis suggests that all QVC topologies tend to learn relatively similar features, implying that their learning capabilities may not be significantly affected by their specific architectural layouts. Interestingly, however, the all-to-all topology stood out in its ability to discern clearer decision boundaries. In this example, by demonstrating a heightened certainty about color borders compared to other shallower topologies, the all-to-all topology suggests a potentially superior discriminative capability. This aspect merits further exploration in future research to understand how it can enhance the robustness and performance of QVCs.

Such distinctions underscore the significance of the QVC architecture choice on the model's ability to resist adversarial attacks. By providing this analysis, we aim to emphasize the need for an in-depth exploration of this relationship to enhance QVC performance and robustness further. This is an aspect that our research will focus on, contributing to the larger narrative of Quantum Machine Learning's ongoing evolution and development.

Chapter 5

Conclusions

5.1 Summary

Our research aims to understand the interplay between the architecture of Quantum Variational Classifiers (QVCs) and their ability to withstand adversarial attacks.
Chapter 1 presents the central research question: How does the architecture of QVCs affect their resistance to adversarial attacks, and what does this mean for model performance and computational efficiency? This section also sets the context and purpose of the study, which is to build on the pioneering work of West et al. (2022) [1] by investigating more deeply the connection between QVC architecture and adversarial robustness.
Chapter 2 provides an overview of the key concepts and terms related to quantum machine learning, adversarial attacks, quantum entanglement, QVC architectures, and encoding techniques. It also reviews existing literature on these topics and identifies the challenges and gaps that this study aims to address. Moreover, it provides a theoretical framework for understanding and evaluating QVCs in the context of adversarial attacks.
Chapter 3 explains the experimental design and methodology used to evaluate different QVC architectures using the FMNIST dataset. It justifies the choice of dataset and encoding technique, as well as the selection of the four QVC architectures: chain, circle, star, and all-to-all. It also describes the metrics used to evaluate the accuracy, training time, and adversarial robustness of the QVCs. Furthermore, it outlines the methods used to generate adversarial examples using PGD and FGSM techniques.
Chapter 4 presents and discusses the results of the experiments conducted on different QVC architectures. It compares the performance of QVCs in terms of accuracy, training time, and adversarial robustness. It also interprets the findings in relation to the research question, the hypotheses, and the existing literature. Finally, it points out the implications, limitations, and challenges of the study.

After rigorous testing and experimentation, the results of our study highlight the superiority of the star topology among different QVC architectures. Our analysis demonstrated that the star topology not only excels in computational efficiency but also maintains commendable accuracy. This is largely due to its unique structure that minimizes the number of quantum gates required, subsequently reducing the computational complexity. This leads to significant reductions in training time, proving advantageous in practical scenarios where computational resources and time are often constrained. Furthermore, despite this emphasis on efficiency, the star topology does not compromise performance. Our results revealed that it consistently achieved high classification accuracy across multiple datasets, rivaling and even outperforming other, more complex architectures. This robust accuracy performance affirms the capacity of the star topology to handle various data types and complexities effectively. The chain and all-to-all topologies show higher resilience to adversarial attacks than the star and circle topologies, suggesting a trade-off between expressiveness and robustness. In summary, the star topology in QVCs proved to be an optimal choice based on our evaluation criteria: computational efficiency, accuracy, and adversarial robustness. It demonstrates that a well-designed, efficient architecture can indeed yield high accuracy, paving the way for the development of more efficient and performant QML models. Future work may explore the resilience of the star topology under various adversarial attack scenarios, to further understand its potential in constructing adversarially robust quantum machine learning systems.

5.2 Conclusion and Future Works

Our study significantly advances the understanding of Quantum Variational Classifiers (QVCs) and their adversarial robustness. The results underscore the importance of choosing the right architecture for QVCs, particularly the star topology, which exhibits notable computational efficiency and maintains impressive accuracy, aligning with the observations of West et al. (2022). These findings provide a deeper understanding of the QVC's performance under adversarial attacks and further validate the theoretical framework established in the literature review.

Moreover, our research contributes to the literature by identifying a trade-off between expressiveness and adversarial robustness in QVCs. Specifically, we found that the chain and all-to-all topologies demonstrate greater resilience against adversarial attacks than the star and circle topologies, reflecting the dichotomy between expressiveness and robustness suggested in the existing literature.

We acknowledge the limitations of our study, including reliance on a single dataset and encoding method, a noise-free environment, and a preset shallow depth for QVCs. These constraints highlight opportunities for future research. Testing diverse datasets,

experimenting with different encoding methods, introducing realistic noise scenarios, and varying the depth of QVCs could provide additional insights into QVC performance and robustness. Furthermore, our study opens up potential avenues for further research, particularly in understanding the complex relationship between QVC architectures and specific problem domains, and in devising efficient error mitigation strategies for QVCs. This research could play a pivotal role in developing adversarially robust quantum machine learning systems.

Another intriguing observation arose from our study suggested a potential attribute. We noticed a tendency for the first qubit in our QVC architectures to play a dominant role in learning key features of the dataset. This phenomenon hints at an inherent hierarchical learning capability within QVCs, where the first qubit could be performing more complex computations, akin to capturing high-level features in classical deep learning models. It's important to note that this is a preliminary observation and was not explicitly tested in the scope of our study. Therefore, we approach this inference with caution. Nevertheless, this observation invites an exciting opportunity for future research. Explicitly investigating this attribute could provide additional understanding of the operational dynamics within QVCs and contribute to optimizing their performance and interpretability. In summary, this finding adds another layer of novelty to our study and further underscores its significance to the field.

In conclusion, our research provides significant, novel contributions to the field of QML, specifically in the realm of QVCs and their adversarial robustness. Given the rapid growth and high interest in quantum computing, our findings will not only spark further research but also serve as a reference point for future studies aiming to optimize QML designs.

Appendix A

Result for other topologies

A.1 Transfer attack

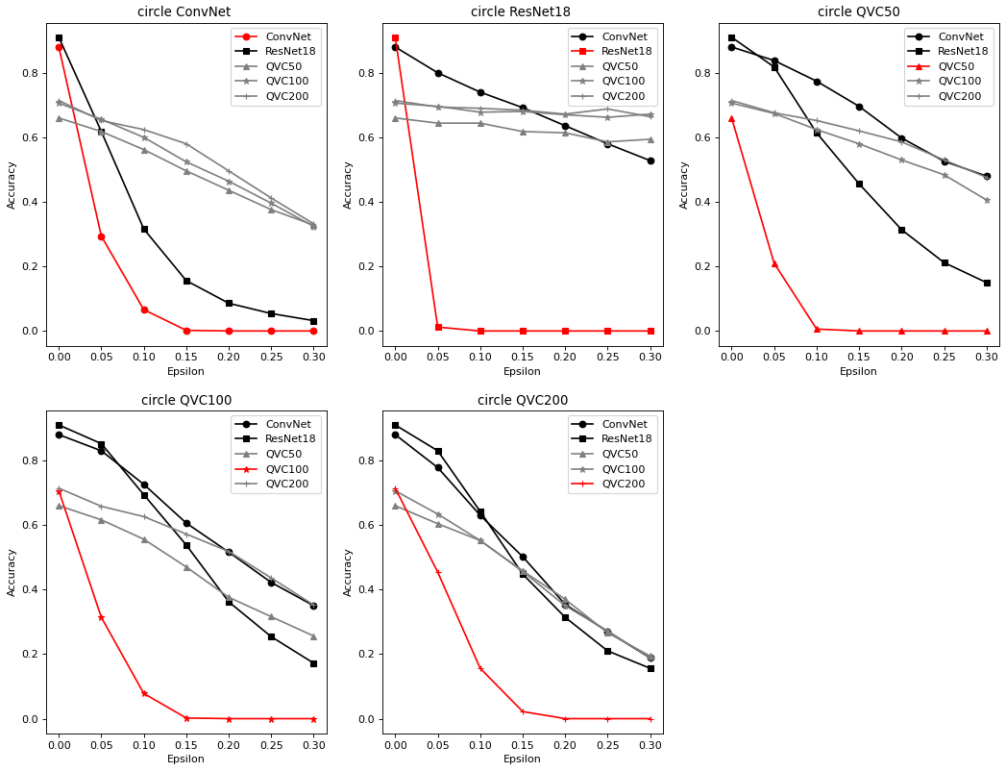


FIGURE A.1: Transfer attack of circle topology with baseline models

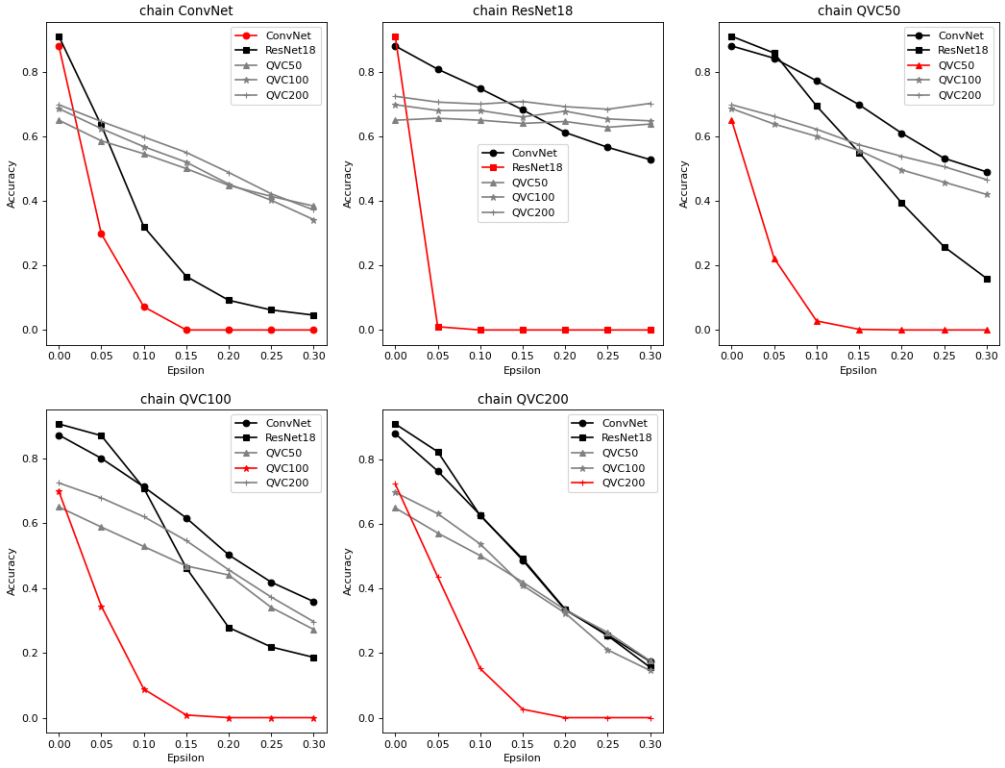


FIGURE A.2: Transfer attack of chain topology with baseline models

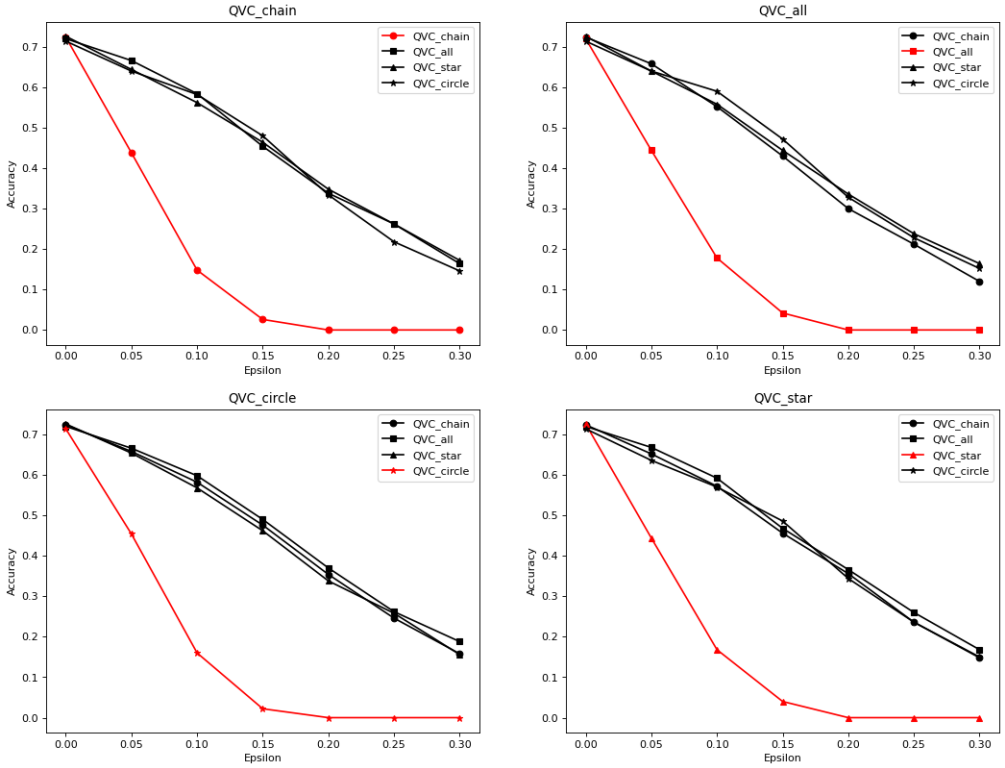


FIGURE A.3: Transfer attack across different topologies of QVC100

A.2 QVC accuracy on test set

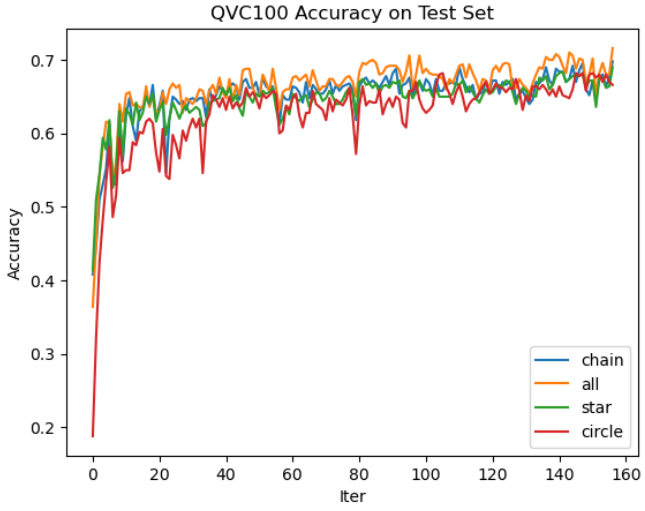


FIGURE A.4: Comparison of Test Set Accuracy Over Iterations for Various QVC100 Topologies

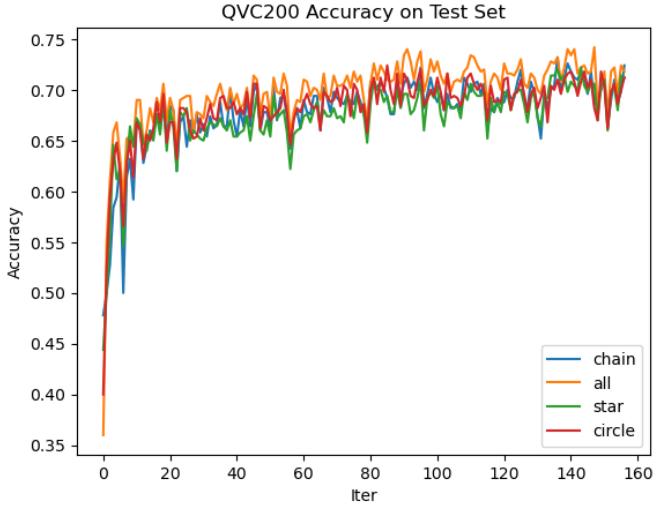


FIGURE A.5: Comparison of Test Set Accuracy Over Iterations for Various QVC200 Topologies

Bibliography

[1] M.T. West, S.M. Erfani, C. Leckie, M. Sevior, L.C. Hollenberg, and M. Usman. Benchmarking adversarially robust quantum machine learning at scale. *arXiv preprint arXiv:2211.12681*, 2022.

[2] J. Biamonte, P. Wittek, and N. et al. Pancotti. Quantum machine learning. *Nature*, 549:195–202, 2017. doi: 10.1038/nature23474.

[3] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] Daniel W Otter, Julian R Medina, and Jugal K Kalita. A survey of the usages of deep learning for natural language processing. *IEEE transactions on neural networks and learning systems*, 32(2):604–624, 2020.

[6] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[7] W. Huggins, P. Patil, B. Mitchell, K. B. Whaley, and E. M. Stoudenmire. Towards quantum machine learning with tensor networks. *Quantum Science and Technology*, 4(2):024001, 2019.

[8] M. Broughton, G. Verdon, T. McCourt, A. J. Martinez, J. H. Yoo, S. V. Isakov, and M. et al. Mohseni. Tensorflow quantum: A software framework for quantum machine learning. *arXiv preprint arXiv:2003.02989*, 2020.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[11] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[12] Sirui Lu, Lu-Ming Duan, and Dong-Ling Deng. Quantum adversarial machine learning. *Physical Review Research*, 2(3):033212, 2020.

[13] S. Lu, L.-M. Duan, and D.-L. Deng. Quantum adversarial machine learning. *Physical Review Research*, 2(3):033212, 2020. doi: 10.1103/PhysRevResearch.2.033212.

[14] W. et al. Ren. Experimental quantum adversarial learning with programmable superconducting qubits. *Nature Computational Science*, 2(11):711–717, 2022. doi: 10.1038/s43588-022-00351-9.

[15] H. Liao, I. Convy, W. J. Huggins, and K. B. Whaley. Robust in practice: Adversarial attacks on quantum machine learning. *Physical Review A*, 103(4):042427, 2021. doi: 10.1103/PhysRevA.103.042427.

[16] Maurice Weber, Nana Liu, Bo Li, Ce Zhang, and Zhikuan Zhao. Optimal provable robustness of quantum classification via quantum hypothesis testing. *npj Quantum Information*, 7(1):76, 2021.

[17] Marco Cerezo, Andrew Arrasmith, Ryan Babbush, Simon C Benjamin, Suguru Endo, Keisuke Fujii, Jarrod R McClean, Kosuke Mitarai, Xiao Yuan, Lukasz Cincio, et al. Variational quantum algorithms. *Nature Reviews Physics*, 3(9):625–644, 2021.

[18] M. Benedetti, E. Lloyd, S. Sack, and M. Fiorentini. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, 4(4):043001, 2019.

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

[20] Y.-C . Liang, Y.-N . Chen, C.-Y . Liu, H.-Y . Lin, Y.-H . Chen, P.-Y . Chen, and C.-C . Hsieh. Robust quantum classifiers via nisq adversarial learning. *Nature Computational Science*, 2(11):718–725, 2022. doi: 10.1038/s43588-022-00359-1.

[21] M. Weber, N. Liu, B. Li, C. Zhang, and Z. Zhao. Optimal provable robustness of quantum classification via quantum hypothesis testing. *npj Quantum Information*, 7(76), 2021. doi: 10.1038/s41534-021-00410-5.

[22] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. *arXiv preprint arXiv:1905.07121*, 2019.

[23] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Diversity can be transferred: Output diversification for white- and black-box adversarial attacks. *arXiv preprint arXiv:2003.06878*, 2019.

[24] Devin Willmott, Anit Kumar Sahu, Fatemeh Sheikholeslami, Filipe Condessa, and Zico Kolter. You only query once: Effective black box adversarial attacks with minimal repeated queries. *arXiv preprint arXiv:2102.00029*, 2021.

[25] Zhiwei Liu, Yujia Liu, and Zhi-Hua Zhou. Query efficient black-box adversarial attack on deep neural networks. *Pattern Recognition*, 123:107996, 2021.

[26] Jian Zhang, Yisen Wang, and Xingjun Ma. Black-box bayesian adversarial attack with transferable priors. *Machine Learning*, 2022.

[27] Maxwell T. West, Shu-Lok Tsang, Jia S. Low, Charles D. Hill, Christopher Leckie, Lloyd C. L. Hollenberg, Sarah M. Erfani, and Muhammad Usman. Towards quantum enhanced adversarial robustness in machine learning. *Nature Machine Intelligence*, 2023.

[28] Anurag Kumar Sharma, Ankit Kumar Sharma, Arpan Dasgupta Bhattacherjee, and Arpan Kumar Bhattacharjee. Adversarial robustness based on randomized smoothing in quantum machine learning. *arXiv preprint arXiv:2106.05968*, 2021.

[29] Louis Schatzki, Andrew Arrasmith, Patrick J. Coles, and M. Cerezo. Entangled datasets for quantum machine learning. *arXiv preprint arXiv:2109.03400*, 2021.

[30] VentureBeat. Quantum machine learning (qml) poised to make a leap in 2023. *VentureBeat*, 2023.

[31] Zhiwei Liu, Yujia Liu, and Zhi-Hua Zhou. Entanglement-based feature extraction by tensor network machine learning. *Frontiers in Applied Mathematics and Statistics*, 7:716044, 2021.

[32] Naema Asif, Uman Khalid, Awais Khan, Trung Q. Duong, and Hyundong Shin. Entanglement detection with artificial neural networks. *Scientific Reports*, 13(1): 1562, 2023.

[33] M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe. Circuit-centric quantum classifiers. *Physical Review A*, 101(3):032308, 2020.

[34] Deepak Shelar. Quantum encoding: An overview. *Medium*, 2021.

[35] Carlos Bravo-Prieto. Quantum autoencoders with enhanced data encoding. *Machine Learning: Science and Technology*, 2(3):035028, 2022.

[36] José D. Martín-Guerrero and Lucas Lamata. Quantum machine learning: A tutorial. *Neurocomputing*, 470:457–461, 2022.

[37] Anurag Kumar Sharma, Ankit Kumar Sharma, Arpan Dasgupta Bhattacherjee, and Arpan Kumar Bhattacharjee. Image compression and classification using qubits and quantum deep learning. *arXiv preprint arXiv:2106.05968*, 2021.

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[39] Esther Ibáñez-Marcelo, Mariano Campoy-Quiles, and David D. O'Regan. Quantum variational autoencoders for feature extraction in classical supervised learning tasks. *arXiv preprint arXiv:1911.02469*, 2019.

[40] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Technical report*, 2009.

[41] Maria Schuld, Ilya Sinayskiy, and Francesco Petruccione. Quantum machine learning in feature hilbert spaces. *Physical Review Letters*, 125(5):050202, 2020.

[42] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[43] PennyLane. Variational circuits. https://pennylane.ai/qml/glossary/variational_circuit.html, 2021.

[44] Samson Wang, Enrico Fontana, M. Cerezo, Kunal Sharma, Akira Sone, Lukasz Cincio, and Patrick J. Coles. Noise-induced barren plateaus in variational quantum algorithms. *Nature Communications*, 12(1):6961, 2021.

[45] Joschka Roffe. Quantum error correction: An introductory guide. *Contemporary Physics*, 60(4):226–241, 2019.

[46] Zhiwei Liu, He-Liang Huang, and Zhi-Hua Zhou. Analysis of error propagation in quantum computers. *arXiv preprint arXiv:2209.01699*, 2022.

[47] Zhiyuan Liu, Yuxin Zhang, Xiang Li, Xiangyu Li, Yixuan Wang, and Shengyu Wang. Quantum teleportation error suppression algorithm based on quantum topological error correction. *Quantum Engineering*, 2022:Article ID 6245336, 2022.

[48] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. NIPS-W, 2017.