IJQCAI

https://ijqcai.org/

International Journal of Quantum Computing and Artificial Intelligence

Original Article

# Adversarial Attacks and Defence Mechanisms in Quantum Machine Learning Systems

Dzulfikar Ahmed[1], Mohamed Hafiz[2], Syed Abuthahir[3], Farid Ahmed[3]

[1,2,3] Department of Computer Science, Middle East College, Muscat, Oman

Received Date: 08 March 2026 Revised Date: 22 March 2026 Accepted Date: 04 April 2026

## Abstract

*Quantum Machine Learning (QML) is a new discipline that combines quantum computing and machine learning algorithms in order to take advantage of the computational benefits which are provided by quantum computing for performing complex problems. While similar to classical machine learning systems, quantum-models tend to be prone for adversarial attacks—craft perturbations designed purposely for causing model mispredictions. The intrinsic noise and probabilistic behavior of quantum systems— in particular the (Noisy Intermediate-Scale Quantum) NISQ—is only worsening these vulnerabilities. This paper provides an overview of adversarial threats in QML and illustrates how the quantum nature of any specific system to be attacked shapes both attack and defence strategies. We study on prevalent adversarial methods including gradient based perturbations, optimization attacks and transfer attacks among others, for their resonance in the quantum settings. And, we explore the influence of quantum noise, circuit layout, and hybrid architectures on system weaknesses. In order to develop solutions for such problems, we provide a survey of defines approaches including adversarial training noise-aware learning quantum error mitigation and robust optimization By incorporating probabilistic approaches like Bayesian inference, the ability to make decisions with uncertainty informs and strengthens resilience. Fields like cybersecurity, healthcare and autonomous systems showcase implications of adversarial robustness in QML. Lastly, the paper presents open challenges and future directions for research where some of them are on building scalable secured interpretable quantum AI systems.*

## Keywords

*QA, QML, QNN, NISQ, Quantum Machine Learning, Adversarial Attacks and Defences in the Quantum World: Robust AI vs. (Quantum) Security*

## Introduction

Quantum Machine Learning (QML) is a revolutionary area where Quantum Computing and Artificial Intelligence come together and bring up the ability to solve computational problems that cannot be solved by a classical system. Through quantum mechanical effects like superposition and entanglement, QML models can both encode and process information in Hilbert spaces of a very high dimension, allowing for more efficient traversal of complex solution landscapes. These qualities help mark QML as a future-proof paradigm for optimization, pattern recognition, and data analysis tasks. At the same time, with the increasing deployment of QML systems there are also major security risks that arise from adversarial attacks.

Adversarial attack is small, structured perturbations added to the input data so to trick machine learning models into making wrong or manipulated predictions. These attacks have been well studied in classical machine learning (e.g., by adding an imperceptible perturbation based on the gradient of a given classification loss, or even optimizing the input directly) and can lead to significant misclassification even with only minimal modifications. Yet, the transfer of adversarial attacks to QML systems presents new challenges with respect to their intrinsic probabilistic nature due to quantum-encoded operations in parallel and different aspects of quantum data encoding. Which basically means that since difference is anyway in the nature of QML based models (using the quantum interaction) then unlike classical systems where inputs are deterministically processed, the output from any quantum measurement possesses some kind of probabilistic properties thus distinguishing noise or unintentional natural variation from a potential adversarial attack becomes harder than expected.

Recent quantum hardware is defined by operating in the Noisy Intermediate-Scale Quantum (NISQ) regime. This framework encompasses, but is not limited to quantum devices that are small in scales, and therefore vulnerable to noise from DE coherence, imperfect gates, or environmental fluctuations. These noise sources, in turn,

induce variation in quantum computations during model training and inference stages. This noise serves a double purpose from the angle of security. It can usefully mask adversarial perturbations, possibly making them less potent; however, it may exaggerate these types of disturbances and hence cause models to be more sensitive to small attacks. As such, this intricate interplay of noise and adversarial behavior results in an exceedingly complicated baseline for the verification of QML robustness.

The widespread usage of hybrid quantum-classical architectures provides another major reason for the vulnerability of QML systems. This platform is hybrid repose with PQP which leads to system combing parameterized quantum cirque and its classical estimation trick. Although hybrids models help overcome some of these challenges, offering improved flexibility and scalability, they create new attack vectors. Adversarial perturbations can be focused on classical portions of a quantum model, for example those fabricating input preprocessing or in case of gradient-based optimization target the classical cost function based on the derivatives, and also target quantum components which could include data encoding schemes and/or circuit parameters. As an example, let a small perturbation happen in the classical input data which can reach time but instead of classical feature map it propagates through a quantum feature map and changes the final quantum resultant state after obtaining at last a false prediction. Such a system structure forms an interconnected frame that makes it necessary to analyze vulnerabilities at both levels instead of treating them separately.

Additionally, the high dimensional feature spaces adopted by QML can further increase their effects of adversarial perturbations. Many quantum encoding methods map classical data to a highly complex quantum state, meaning that small perturbations of the input can lead to a large variation in the encoded representation. Such sensitivity makes it easier to successfully execute adversarial attacks, especially when combined with the significant uncertainty associated with quantum measurements. The second challenge is the absence of mature security frameworks and standardized evaluation protocols in QML, making it cumbersome to develop resilient defines systems.

In light of these difficulties, systematic research on adversarial robustness in QML is more important than ever. These include: understanding what traditional attack strategies mean in the quantum context, identifying vulnerabilities that can be specific to quantum implementations and proposing defences which take into account both noise and uncertainty. Adversarial training, noise-aware learning and uncertainty quantification are some of the most promising techniques for making QML models more resilient. Incorporating probabilistic techniques such as Bayesian inference can also enhance robustness by allowing models to measure uncertainty and identify out-of-distribution examples.

In this paper, we aim to conduct a comprehensive overview of adversarial attacks and their defines mechanisms in such quantum machine learning systems. It investigates the possible forms of adversarial methods relevant to QML, inducement of more vulnerability and reviews current/future adversarial safeguards. The paper addresses these points to aid the increase of safe, stable and trustworthy quantum AI systems that can handle in malicious capes.

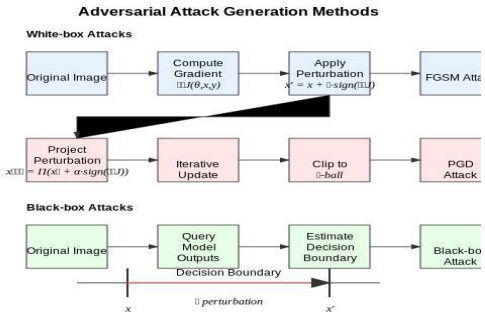


***Figure 1: Adversarial Attack Concept Input Perturbations***

## Query-Execution Frameworks for the Adversarial Quantum Machine Learning

Adversarial robustness in machine learning has become a mature field of study 10 years on, establishing itself as an essential pillar for understanding the vulnerabilities behind many state of the art intelligent systems. Adversarial attacks in classical machine learning demonstrate an inherent flaw of deep neural networks, which is that small perturbations can lead to very different model predictions while having little noticeable change to the input data. Gradient-based attacks, optimization-driven perturbations, and transfer-based methods have all shown that even high-performing models are very easy to break. Such outcomes have spurred significant work on attack

methods as well as defensive strategies, making adversarial machine learning a prominent field of research within Artificial Intelligence.

With the realization of Quantum Machine Learning many of these classical insights are being re-evaluated, and transformed into quantum-settings. QML models, particularly parameterised Q circuits have structural similarities to classical neural networks in that they typically consist of layers and the parameters are all trainable. This similarity makes them vulnerable to similar adversarial techniques. For example, adding perturbations to classical input data can shape the quantum encoding process, setting differences between those states and producing erroneous predictions. However, quantum system possess new security aspects that alter the way we think about adversarial attacks because quantum circuit parameters themselves can be targeted to manipulate the output etiquettes in an adversarial manner.

QD1 — perhaps the most defining feature of QML work — is the NISQ device framework: where quantum devices are affected by noise, modest quit counts and imperfect operations. Such restrictions have a considerable impact on the characteristics of adversarial attacks and the performance of defines strategies. In classical systems, noise is considered to be a deleterious artefact of the computation process but quantum noise is an intrinsic property of the computation. In turn, this creates a nontrivial relationship between adversarial perturbations and the non-deterministic nature of quantum processes. Noise can thus, in some cases mask the adversarial signals, and in others magnify their effects leading to a more vulnerable system. Recent research has been focused on you figuring out this interaction.

Another key development in this space is the rise of hybrid quantum-classical models, which combine quantum circuits with classical optimization methods. They are popular because they run on modern hardware and can use classical computing power. But they also inherit risks from both worlds. Traditional notions of adversarial attacks still apply to classical components, such as data preprocessing and gradient-based optimization, while the inclusion of quantum components opens additional new attack avenues due to errors in encoding and measurement. Such bifurcated characteristics necessitate a consolidated security stance that treats the system architecture — not its individual elements — as a singular whole.

Quantum-specific adversarial attack models have also been studied in some recent works. Those include perturbations that act directly on quantum states, control over probabilistic measurement outcomes, and the interference that occurs with quantum gate operations. These types of attacks take advantage of the unique characteristics of quantum systems such as superposition and entanglement to be successful. While the field is not matured yet but it shows there is requirement of specialized framework which could fill the gap between classical adversarial analyses.

The defines community has proposed various promising approaches to make QML systems more robust. One such approach, adversarial training — or exposing models to input examples with specially-crafted perturbations during the model's training protocol — has been adapted to quantum settings to produce resilient QNNs. Noise-aware learning methods are based on training the models under realistic hardware noise conditions which enable better model generalization in case of uncertain nature of operation. Quantum error mitigation techniques are designed to alleviate the effects of hardware flaws and, thus, indirectly enhance resilience against adversarial contamination as well.

Advancements in the field include incorporation of probabilistic methods such as Bayesian inference allowing for uncertainty aware predictions. This class of approaches enables models to estimate their own level of confidence, which leads to easier spotting of anomalous or adversarial inputs. Incorporating uncertainty estimation with robust optimization techniques, researchers are bringing QML systems into a more reliable and trustworthy direction

Thus, in broader sense, the background and related work in adversarial QML represents a collision of conventional principles from adversarial machine learning with the orders of different functionalities offered by quantum computing. Although there has been great progress, the field still has a lot of room for growth in areas such as scaling, safety, and general deployment. Further research in this domain is critical for supporting robust quantum AI systems able to run safely in hostile environments.

## Adversarial Attack Models in Quantum Machine Learning

As the implementations of quantum models become more developed and on the rise, adversarial attacks in Quantum Machine Learning (QML) emerge as a new security issue. Analogous to classical ML, QML systems can also be perturbed by custom crafted attacks that exploit specific properties of the underlying quantum hardware by modifying model predictions. However, in quantum systems the properties of superposition, entanglement and probabilistic measurement add a considerable complexity layer to these attacks. The synergy of adversarial perturbations and quantum noise makes systems more susceptible to vulnerabilities in the settings of Noisy

Intermediate-Scale Quantum (NISQ) devices. The major adversarial attack models in QML are discussed in this section using a classification approach.

### A. Gradient-Based Attacks

Gradient-based attacks are one of the most analysed adversarial methods. These techniques leverage the gradients of a loss function with respect to input data (for training-aided methods this loss is often model agnostic) to produce perturbations that increase the prediction score. Gradient information can be extracted with parameterized quantum circuits in QML using methods like the so-called parameter-shift rules. Fast gradient sign method (FGSM) and projected gradient descent (PGD) attacks can be modified to quantum systems by adding the perturbation on top of classical input before quantum encoding. These attacks are simple, but they work very well: Just like classical neural networks, QML models can be sensitive to a slight change in input data.

### B. Optimization-Based Attacks

The optimization-based attacks tend to be more sophisticated and cast adversarial example generation in terms of an optimization problem. Specifically, the goal is to compute the most minimal perturbation that results in a model making an incorrect assignment for a particular input. Includes C&W, These techniques are especially powerful because they allow you to optimize directly for success and small perturbation. These attacks can target both input data and quantum circuit parameters and are more general, which is difficult to resist against in QML. They use a more specific targeting of the model structure compared to gradient-based methods, which is what contributes to their effectiveness.

### C. Transfer Attacks

Transfer attacks utilize the observation that adversarial examples produced for some model may deceive other models although they have different architectures or training data. The importance of this transferability property is especially concerning for QML, since when hybrid quantum-classical systems share common components or representations of features. The attacker trains adversarial examples with a surrogate model and uses them on a targeted QML system without being able to access its internal parameters. Transfer attacks are nontrivial threats and potentially scalable, especially on the condition that the target model is unavailable or proprietary.

### D. Quantum-Specific Attacks

Quantum-specific attacks, on the other hand, are only found in QML and affect quantum elements for which there is no classical analogue. These range from perturbations to quantum states directly, gate parameter manipulation or measurement process interference. In the first case, for example, an attacker can change the parameters of a Variational quantum circuit and make it operate worse or inject errors into the output distribution. Furthermore, attacks on quantum encoding schemes could create a mismatch between how classical data is characterized in quantum states and the steps followed to make predictions. Much harder to detect and/or militate against, these attacks exploit the intrinsic properties of quantum systems for their purposes.

Conclusively, adversarial attack models in QML are both adaptations of classical techniques and completely new quantum-derived models. These attacks take advantage of the fact that QML systems are sensitive to small perturbations, alongside the complexity induced by quantum mechanics and hybrid architectures. This knowledge aids in constructing effective defines mechanisms, which is critical for the security and reliability of quantum machine learning systems in real-world applications

***Table 1: Adversarial Attack Models in QML***

| Attack Category | Examples | Target Component | Key Mechanism | Impact on QML Systems |
|---|---|---|---|---|
| Gradient-Based Attacks | FGSM, PGD | Input data | Uses gradients to create perturbations | High sensitivity to small input changes |
| Optimization-Based Attacks | C&W Attack | Input and parameters | Solves optimization problem for minimal perturbation | Highly precise and difficult to detect |
| Transfer Attacks | Cross-model attacks | Multiple models | Exploits shared representations across models | Works without direct model access |
| Quantum-Specific Attacks | Circuit perturbations | Quantum states and circuits | Targets quantum encoding and gate parameters | Unique vulnerabilities in quantum systems |

## Quantum Machine Learning Systems: Vulnerabilities

Recent Quantum Machine Learning (QML) systems provide large computational advantages, but these types of designs are ripe for a number of possible weaknesses stemming from both the quantum and classical components. In turn, quantum computation is probabilistic in nature while the current hardware has limitations as

well as architectures are often hybrid, all of which exacerbate these vulnerabilities. Realizing these vulnerabilities is crucial to designing reliable and secure QML models, especially when dealing with adversarial cases.

High-dimensional quantum feature spaces are one of the major weaknesses of QML systems. Since we can write quantum states in exponentially large Hilbert spaces, many quantum models transform classical data into a quit state with strong feature transformation and classification capacity for complex datasets. This representation in high-dimensional space, however, is also more susceptible to small perturbations. Slight variations in the input data can cause the quantum state itself to move substantially, despite producing predictions that is incorrect. This can be related to adversarial vulnerabilities on classical deep learning but is exacerbated with the complexity of quantum encoding schemes. Therefore, an attacker can exploit this sensitivity to create perturbations that are heavily affecting model outputs.

QML systems are also sensitized to noise and perturbations which constitutes another key vulnerability. Noisy Intermediate-Scale Quantum (NISQ): Due to the DE coherence, gate errors and measurement uncertainty on quantum devices, they inherently are noisy. The presence of such sources of noise leads to the introduction of randomness in computations and detracts from both the reliability and stability of model predictions. Although noise is an inherent property of quantum systems, it can couple to adversarial perturbations in a nontrivial manner. For the related task of detecting adversarial inputs, that noise can cover over the signals (i.e. motions) we would like to observe. Similarly, it may work in other cases, magnifying the effect of perturbations and making misclassification more likely. This dual interaction is a tricky space for making sure the model is robust.

By integrating two computational paradigms that differ radically (and whom already introduce their vulnerabilities), hybrid quantum-classical architectures bring new risks along. In these systems, the data is both encoded and (most importantly) transformed using quantum circuits while optimization and decision-making takes place classically. While that allows implementation on existing hardware, it also opens up multiple attack vectors. Specifically, classical elements that could be influenced by adversarial perturbations include input preprocessing or gradient-based methods of optimization; quantum elements consist of circuit parameters and measurement processes. This inter-relationship implies that we can transfer the weakness in one part of system to whole model which leads toward performance degradation and more expensive attack.

Combining the issues with over fitting and low generalization, these aspects make QML systems particularly vulnerable. Quantum models, similar to those in classical ML, involve complex structures; thus there is also an increased risk due undersized training data that models may learn the trends of the specific training dataset only instead of features generalizable. This can be more problematic for quantum systems as the training data are often expensive and scarce. Because over fitted models are very sensitive to small perturbations in input data, they are also more susceptible to adversarial attacks. Secondly, the problem of poor generalization may undermine its utility in real-world scenarios by making it less reliable at predicting cases when exposed to new or slightly different inputs. Another root cause which amplifies these vulnerabilities is the immaturity of security frameworks and lack of standardized evaluation protocols for QML systems. In contrast to classical machine learning, where adversarial robustness has been well-studied and established benchmarks available, QML is still a developing field. What this means is that multiple models are deployed without adequate testing against adversarial scenarios — thus making them susceptible to exploitation. And the lack of standardization metrics and evaluation means no apples-to-apples comparison is possible for how strong different models are not, nor what best practices exist for secure design.

To summarize, the joint effects of high dimensionality in feature representations, noise and perturbation sensitivity, hybrid architecture complications, training data inherent limitations, and generalization deficiencies underlie QML system vulnerabilities. Each of these creates a unique environment for introducing and maintaining robustness and security. Mitigating these vulnerabilities requires a comprehensive methodology that incorporates quantum and classical elements of the system, in addition to generating new techniques for robust training, noise suppression and adversarial defines.

## Quantum Machine Learning: Defense Mechanisms

The adversarial threats being an ever more relevant concern in QML, engaging the design of appropriate defines strategies is critical in maintaining enormous trust and protection within their models. QML is very different from classical systems in that it operates within a noisy and probabilistic setting—specifically NISQ. It complicates defines strategies from needing to deal with both adversarial perturbations and imperfections on the quantum hardware level. These modern approaches treat robustness as an end-to-end procedure by incorporating into the learning chain both adversarial training, noise aware learning, quantum error mitigation and advanced optimization techniques.

### A. Adversarial Training

One defence strategy which has been widely used in classical and quantum machine learning is adversarial training. This is done by augmenting the training dataset with adversarial examples to train model be able to identify and prevent in malicious perturbations. Adversarial training in QML can be done via perturbed classical inputs before the quantum encoding or knocked classical circuit parameters during training. This identifies the adversary data, and such attacks are durably protected by exposing the model to both: worsening (real) test data at inference. On the contrary, adversarial training increases computational overhead and is not always effective against previously unseen adversarial attacks so that it must be crafted accurately and kept updated.

### B. Noise-Aware Learning

Noise-aware learning employs the system noise that proves advantage training. Instead of being treated as a noise that should be completely eliminated, this model accounts for realistic noise assumed in the quantum simulation and training process. Through this, the QML models learn to function in noisy conditions and provide robustness towards hardware imperfections as well as adversarial perturbation. This technique works really well in NISQ devices where noise cannot be avoided. Also, noise-aware learning along with probabilistic methods like Bayesian inference help models learns uncertainty and then better identifies anomalous inputs.

### C. Quantum Error Mitigation

Quantum error mitigation techniques are designed to minimize the effects of noise without full-blown quantum error correction, which remains unfeasible. Commonly used to mitigate errors and enhance the fidelity of quantum computations are techniques like zero-noise extrapolation, probabilistic error cancellation, and measurement error mitigation. Specifically, error mitigation improves QML models stability against perturbations and can find use in adversarial defines. These techniques indirectly lower the performance of adversarial attacks that rely on system noise by improving the fidelity of quantum operations. But using error mitigation techniques often necessitates circuit repetitions and greater computational effort, resulting in overhead.

### D. Robust Optimization Techniques

The key features of robust optimization techniques are the modelling of uncertainty and adversarial conditions. These approaches include worst-case scenarios into the optimization procedure, making it so that when input to perturbations, the model stays invariant. Could QML be used in practice by using loss functions on quantum computers that the optimization algorithm cannot minimize, hence increasing robustness? Solutions include modify- captions of loss function, use of regularization strategies and even risk-sensitive learning approaches. So models can be trained to minimize not just prediction error, but also output sensitivity to input perturbations. This results in predictions that are much more robust and reliable, with low variability especially important in high-risk applications. Robust optimization is commonly driven by uncertainty-aware methods to increase model robustness.

In summary, defines mechanisms in QML need a multi-layered approach that includes both adversarial and quantum-special concerns. Adversarial training improves the model robustness against the known attacks; noise-aware learning increases adaptability to real-world scenarios; quantum error mitigation provides necessary stability for computation, and robust optimization secures the predictive performance under uncertainty. These strategies integrated with each other create a structured defence outline that is necessary to deploy QML systems securely. As quantum technologies keep evolving, much more work on scalable and adaptable defines mechanisms is essential to creating credible quantum AI tools.

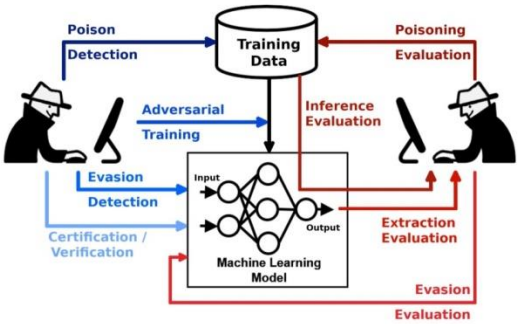


***Figure 2: Adversarial Training in QML Systems***

## Hybrid Quantum–Classical Security Methods

Hybrid quantum–classical security methods have arisen as a bulwark for Quantum Machine Learning (QML) systems against adversarial attacks. In summary, the potential of quantum computation combined with the maturity and robustness of classical machine learning defences leads to an optimal practical framework for security improvements. Purely quantum defences are typically not enough in today's NISQ (Noisy Intermediate-Scale

Quantum) era where the quantum hardware is limited and noise-riddled. Hybrid approaches overcome this limitation by deploying security mechanisms over both quantum and classical components, establishing a stronger system.

### A. Integrated Architecture for Security

In a hybrid QML system, there are generally quantum circuits responsible for encoding and transforming features followed by classical layers that perform optimization, classification or decision making. This architecture allows defences to be implemented at all stages of the pipeline which is great from a security perspective. For example, classical preprocessing steps can filter or normalize input data before encoding into quantum states, making systems less vulnerable to adversarial perturbations. In a similar fashion, classical post-processing modules can monitor quantum output and establish irregularities or inconsistencies which may trigger an attack. Using this layered approach increases the robustness of the overall system by adding safeguards that can compensate if one component is vulnerable.

### B. Hybrid Models for Classical Defence Techniques

The most significant benefit of hybrid systems is that they draw from well-established classical defences. There are different methods to incorporate these techniques, like adversarial training, input regularization and anomaly detection, into the classical part of the model. These approaches find and reduce adversarial inputs before they spread through the quantum circuit. Likewise, classical ML models can be trained to recognize the patterns corresponding to adversarial behavior and facilitate its monitoring and detection in real-time. Due to these strengths, hybrid QML systems inherit over decades of expertise in adversarial machine learning.

### C. Quantum-Level Security Enhancements

On the quantum nature of security strategies, they are dealing with how to increase error-resiliency of quantum circuits and minimize noise and perturbations. Tools used in this context include noise-aware circuit design, operator parameter regularization, and quantum error mitigation. This helps hybrid models access deeper circuits with fewer gate operations, thus little exposure to DE coherence and other sources of noise between measurements that make it easy for adversarial perturbations to acquire influence over outcomes. In addition, using probabilistic methods like Bayesian inference allows the system to express the uncertainty associated with quantum outputs, offering another line of define against potentially nefarious or anomalous inputs.

### D. Cross-Domain Robustness and Adaptability

Some of hybrid quantum-classical security approaches have the advantage of being suitable for processing various types of threats. As adversarial attacks fit on both quantum side and classical side, a universal defence is required for them. This is made possible through hybrid models that allow coordinated response across domains. In one example, if a classical anomaly detection system detects a potentially adversarial input, it can change the quantum circuit or optimization process hybrid management to relieve sensitivity. The interplay between various components provides the system with the means to adapt to emerging threats.

### E. Practical Implications and Applications

In practical applications, hybrid security methods have significant value when safety and reliability are of utmost importance. Keeping QML models reliable is non-negligible in several domains including healthcare, finance and autonomous systems because of the potential for a hazardous impact. In summary, hybrid architectures offer a viable route towards deploying practical QML systems on existing hardware and, thus the bridge between theoretical quantum advantages and physical use-cases. They also scale well, with classical components managing extremely large data processing while quantum components perform specialized tasks.

This greatly enhances the ability to evaluate security of QML systems, especially via novel hybrid quantum-classical methods. Combining classical defence mechanisms with improvements to the quantum level, these models are able to reach a good trade-off between robustness and efficiency / scalability. Hybrid approaches will be pivotal to enabling security and reliability of quantum machine learning systems in complex and adversarial environments as quantum technologies continue to mature.

## Experimental Evaluation

Experimental validation is crucial for probing the strength and reliability of QML systems being developed in the presence of adversarial attacks. However, in the Noisy Intermediate-Scale Quantum (NISQ) QML evaluation must account for quantum noise and adversarial perturbations simultaneously; hadamard a sophisticated evaluation framework will clarify to what extend a model can retain performance under attack, yet also take into account uncertainty and compromises in hardware.

### A. Evaluation Setup

- Datasets containing both clean and adversarial perturb samples
- QML models implementation such as quantum neural networks and hybrid architectures

- Simulation or execution over quantum hardware with realistic noise models
- Usage of various adversarial attacking strategies, such as gradient based and optimization-based methods
- Use of defines strategies adversarial teaching, error mitigation
- This arrangement guarantees evaluation under real world scenarios in which adversarial and noise threats jointly exist.

### B. Key Evaluation Metrics

- Accuracy: Classification accuracy using plain data before preprocessing.
- Robustness: Measures the model performance during adversarial perturbations
- Attack success rate (ASR): The percentage of the adversarial inputs that can mislead the model
- Uncertainty Calibration: Measures how closely predictive confidence matches what happens
- Noise Robustness: Evaluates performance stability at differing levels of noise
- These metrics give us a more holistic view of how a model is performing, not just accuracy but also its security and reliability.

### C. Performance Under Adversarial Conditions

- Adversarial examples derived via various attack strategies are used to test models
- Performance deterioration: performance degradation is researched and measured as compared with results before and after attacks.
- Robust models have little reduction in accuracy and high attack fail rate
- Performance erodes gracefully for many hybrid quantum-classical models, often benefiting from classical correction71–73
- Testing performance against your adversary is helpful for finding weaknesses, and provides direction on where to go with model design.

***Table 2: Evaluation Metrics for Adversarial QML Systems***

| Metric | Description | Purpose in Evaluation |
|---|---|---|
| Accuracy | Percentage of correct predictions on test data | Measures baseline model performance |
| Robustness | Ability to maintain performance under adversarial conditions | Evaluates resistance to attacks |
| Attack Success Rate | Proportion of adversarial inputs that cause misclassification | Quantifies effectiveness of attacks |
| Uncertainty Calibration | Alignment between predicted confidence and actual outcomes | Assesses reliability of predictions |
| Noise Resilience | Stability of performance under varying noise levels | Evaluates hardware robustness |

## Applications of Adversarially Robust Quantum Machine Learning

Adversarial robustness is not just a theoretical requirement in Quantum Machine Learning (QML); it is a practical necessity for deploying QML systems in real-world environments. As QML begins to influence critical sectors, ensuring that models can withstand adversarial manipulation becomes essential. Applications in cybersecurity, healthcare, and autonomous systems highlight both the potential of QML and the importance of integrating defines mechanisms against adversarial threats. These domains operate under high uncertainty and risk, making reliability and robustness key performance factors.

### A. Cybersecurity Applications

In cybersecurity, QML can be used for intrusion detection, anomaly detection, and threat intelligence analysis. Quantum-enhanced models have the potential to process complex network data and identify patterns that are difficult for classical systems to detect. However, adversarial attacks can manipulate input data, such as network traffic logs, to evade detection systems. Robust QML models address this issue by incorporating adversarial training and anomaly detection mechanisms, enabling them to distinguish between legitimate and malicious activity even under adversarial conditions. Additionally, uncertainty-aware techniques allow systems to flag suspicious inputs with low confidence, prompting further investigation and reducing the risk of undetected attacks.

### B. Healthcare Applications

Healthcare is another domain where QML can provide significant benefits, particularly in medical imaging, disease diagnosis, and personalized treatment planning. Quantum models can analyze large-scale biological and clinical datasets to identify subtle patterns that may not be visible using classical methods. However, adversarial perturbations in medical data—such as slight modifications to imaging inputs—can lead to incorrect diagnoses. This poses serious risks to patient safety. By integrating robust defines mechanisms, QML systems can detect

anomalies and provide confidence estimates for predictions, allowing healthcare professionals to make more informed decisions. The combination of quantum computational power and uncertainty-aware modelling enhances both accuracy and trustworthiness in medical applications.

### C. Autonomous Systems

Autonomous systems, including self-driving vehicles and intelligent robotics, rely heavily on real-time data processing and decision-making. QML can improve these systems by enabling faster and more efficient analysis of sensor data, such as images, liar signals, and environmental inputs. However, adversarial attacks can manipulate sensor data to mislead the system, potentially causing unsafe behavior. For example, slight alterations to visual inputs could cause a vehicle to misinterpret traffic signs or obstacles. Robust QML models mitigate these risks by incorporating adversarial defences and uncertainty estimation, allowing the system to recognize when inputs are unreliable and خاذات safer actions. This enhances both safety and reliability in dynamic environments.

***Table 3: Applications of Adversarial Robust Mel***

| Application Domain | Use Case | Adversarial Risk | Role of Robust QML | Benefits |
|---|---|---|---|---|
| Cybersecurity | Intrusion detection, anomaly detection | Evasion through manipulated network data | Detects anomalies and flags uncertain inputs | Improved threat detection and reduced false negatives |
| Healthcare | Medical imaging, disease diagnosis | Misleading inputs causing incorrect diagnoses | Provides confidence estimates and detects anomalies | Enhanced patient safety and diagnostic accuracy |
| Autonomous Systems | Self-driving vehicles, robotics | Manipulated sensor data leading to unsafe decisions | Enables risk-aware decision-making under uncertainty | Increased safety and operational reliability |

## Intelligent Noise Mitigation in Quantum Circuits Using Machine Learning

The field of adversarial robust Quantum Machine Learning (QML) remains an emerging area, and to facilitate the deployment of these systems in real applications, many key challenges remain. QML holds great potential for more computationally efficient ownership stakes, however incorporating adversarial robustness imposes further challenges as uncertainty and hardware limitations are also entangled with the environment. This chapter analyses some of the major new challenges arising and discusses future research directions that will determine how QML systems will be implemented so as to maximize their security and reliability.

Scalability is one of the biggest issues. The Noisy Intermediate-Scale Quantum (NISQ) is the regime of quantum systems currently available to us, where we have a limited number of quits and do not yet have low enough error rates. When it comes to supporting larger data sets and complex tasks QML models are relatively more computationally. Things get even more complicated with adversarial defines techniques like adversarial training or robust optimization since they may add extra computations and necessary evaluations. How to scale such techniques without losing performance is an open research problem. We will need to spend effort developing significantly lighter, quantified algorithms that can run on current and/or future quantum hardware.

A second important problem has to do with noise and uncertainty management. Since quantum systems are noisy by definition, the behavior of noise in combination with adversarial perturbations varies non-deterministically. Although noise can disguise adversarial attacks on occasions, it can also increase vulnerabilities and open the floodgates to models being attacked more easily. It is a tough challenge to design defines schemes that could deal with both quantum noise and adversarial inputs at the same time. There is a need for further research investigating advanced noise modelling methods, adaptive error mitigation approaches, and uncertainty-aware learning paradigms that can adapt to varying environmental conditions. In addition, embedding probabilistic tools such as Bayesian inference within QML models might lead way to a potential solution to these problems by improving uncertainty judgement and decision making.

Another common limitation limiting the field is the absence of standardized benchmarks and evaluation protocols. Note that for most classical machine learning, there are already a number of datasets and metrics for evaluating the ability of a model to protect itself against adversarial attacks, but such framework does not exist in QML yet. The dissimilarities in hardware platforms, noise models, and experimental setups make it challenging to compare results from different studies. This means that progress is slower and the proposed defines mechanisms may not always work. Defining consistent benchmarks, datasets and measures of evaluation are key to moving the field forward and making reproducible research possible.

This is where the idea of interdisciplinary integration comes to offer us both a problem and an opportunity. Adversarial QML NovemberQML combines quantum physics, machine learning, cybersecurity and optimization. By spanning these domains, bridging will need contributions from a wide-variety of corresponding researchers but also new theory uniting concepts stemming from separate fields. Future work should develop interdisciplinary methodologies leveraging principles from quantum information theory, robust ML and security engineering to tackle challenging problems in a unified approach.

A secure quantum-native is another avenue that is being developed. Despite many existing defines mechanisms being verged on classical machine learning, quantum systems also need expressly designed approaches. Such countermeasures may be, for example, quantum-specific where proven anomaly detection modes are deployed at a network layer (as does the Y390), secure quantum encoding or encoding methods designed at the circuit level and use properties of quantum physics such as entanglement and interference to defeat certain attacks at a physical layer. Investigating these quantum-native methods can provide more targeted and efficient security solutions reflecting how QML works.

Another significant area for future research is the incorporation of Explainability and trustworthiness into adversarial QML. When QML systems are deployed in sensitive domains, it is crucial to enable transparency and interpretability of their decisions. Explainable AI methods, possibly augmented with uncertainty-aware modelling, provide insights into both predictions and the risk (or confidence) associated with those predictions. This is crucial in applications including healthcare and autonomous systems where the decisions taken must be reliable and auditable.

The move from both experimental researches to real-world deployment, however, has practical challenges. The operational deployment of QML systems necessitates solid software frameworks, an efficient hardware interface, and a scalable deployment strategy. Threat modelling in these environments relates to both adversarial against models but also more general system-level threats, namely data integrity and communication security.

The future for adversarial robust QML is extremely bright, but developing it comes with considerable challenges around scalability, noise, standardisation and cross-disciplinary integration. Confronting these challenges is going to mean continued work in areas from algorithms, hardware and theory. Focusing on these contributions will enable the realization of more secure reliable and scalable quantum machine learning systems during operations in complex and adversarial environments.

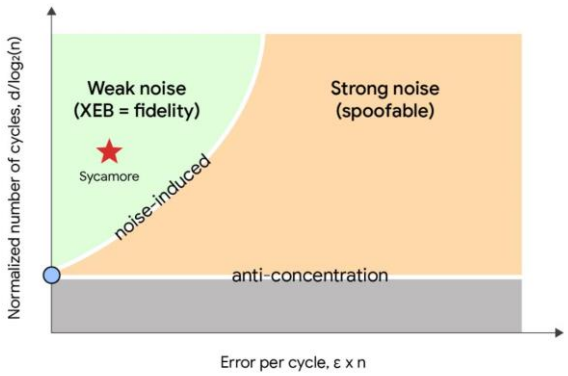


***Figure 3: Noise in Quantum Circuits (Error Sources)***

## Quantum-Adversarial Co-Design: Security in Qml Development

So, as Quantum Machine Learning (QML) systems will continue to mature and take hold, we are going to need to fix not only our reactive defines strategies but also transition to proactive security frameworks. An emerging direction of this is quantum-adversarial co-design, in which security concerns are incorporated into the design and training of QML models. In contrast to adversarial robustness being added on as an afterthought, co-design combines attack awareness, noise modelling and uncertainty estimation from the very first stages of system development. It is specifically relevant in the case of NISQ (Noisy Intermediate-Scale Quantum), where noise and adversarial risks must be mitigated at the same time.

Model architecture, training strategy, and security mechanisms optimized jointly is defined as quantum-adversarial co-design. Within this framework developers, predict the kinds of adversarial attacks that might be faced and build counter measures into the system. As an example, quantum circuits can have robustness constraints such that small perturbations in input data do not result in large difference of output. Equally, adversarial scenarios can form part of the training processes so that models learn to defend as well as execute other task specific goals.

This unified model minimizes the necessity for individual defines side, which leads to an overall improvement in system robustness.

Secure Data Encoding is one of the key elements of co-design. Feature encoding techniques are used to map classical data into quantum states in QML. Unfortunately, this also has the side effect of magnifying very small perturbations, which can be exploited by adversaries to trick systems into misclassifying data. During the design phase, co-design strategies dedicate their efforts to create encoding styles that are invariant with respect to input perturbations. That is, one can use redundant encoding schemes or noise-resistant mappings that guarantee (to a certain extent) that adversarial perturbations will not affect the resulting quantum state too much. Strengthening this first block significantly increases the resilience of the model.

A key feature of co-design is the embedding of adaptive learning mechanisms. Dynamic defines strategies can be implemented to QML models based on the threats it observes. During training itself, for instance, the model can learn to recognize characteristics of adversarial inputs and update its parameters. The Bayesian inference-inspired techniques can estimate uncertainty and detect anomalies in real-time. This enables the system to react proactively to novel attack methods instead of using only define which is predefined.

Co-design at the quantum level focuses on the improvement of circuit structures to achieve higher resistance. These approaches comprise reducing circuit depth to limit noise exposure, choosing gate sequences that are more robust against perturbations and integrating error mitigation into the circuit design. Circuit architecture matching with hardware limitations and security needs enables developers to build models that are naturally more robust. This also means low post-processing corrections, which makes the computations more reliable and faster.

Quantum-adversarial co-design has multiple benefits or advantages over conventional defines mechanisms. It formalizes a single setting where both noise and adversarial attacks are considered, the framework reduces the duplication of defensive models that has happened in the past years making it more efficient as it implies fewer epochs when training, enabling us to use one defensive mechanism at a time, this also helps us to adapt successfully to all types of attack on neural networks. Co-designed QML models can be more reliable and trustworthy in real-world applications, including cybersecurity, healthcare, and autonomous systems. Also, it adds to the scalability since security is built-in rather than deployed separately.

Putting everything together, quantum-adversarial co-design thus enables a somewhat-also-futuristic design paradigm to develop secure and robust QML systems. This approach accounts for the challenges presented by quantum environments through considering security at every stage of development, from data encoding to circuit design and training. In the long run, co-design principles will be essential to guarantee that quantum AI systems, which are already bound for greatness, are also robust and trustworthy in combating adversarial threats as QML matures.

## Ethical, Privacy and Trust Implications in Adversarial Quantum Machine Learning

With the growing interest towards the integration of Quantum Machine Learning (QML) systems into real-world applications, ethical, privacy and trust considerations have received more attention. Although most of the research in QML deals with performance, robustness and computational advantages, deploying such systems in sensitive domains needs careful consideration. Even more, adversarial attacks exist which can result in malicious manipulation of data and models leading to negative consequences. Thus, it is necessary to have ethically aligned and trusted QML systems for a responsible adoption.

Biased or Unfair Decision-Making: This is one of the main ethical concerns in quantum ML. Similar to classical machine learning, the QML systems are trained on data, which can come with built-in biases. If not properly mitigated, this can lead to biased or unfair models. Adversarial attacks can makes this worse by targeting certain affected groups or making use of existing bias found in the data. As one instance, even in healthcare applications, adversarial perturbations can cause a misdiagnosis of a certain population which raises serious ethical issues. There is then the need for curated data, fairness aware training techniques, and persistent monitoring of model outcomes to suppress bias.

Similar issue concerns with Privacy within QML systems are especially problematic for sensitive data, such as medical records, financial information, or even personally identifiable information. But quantum computing opens new doors and creates challenges here. Although quantum algorithms can improve data processing abilities, they also pose a threat to data security and privacy. For QML, adversarial attacks might be performed with the intention of leaking privileged information or in order to manipulate results by intervening within the encoding processes of data. To guarantee privacy in QML, we need secure data handling techniques like encryption, anonymization and privacy-preserving learning methods. They also enable identifying suspicious inputs that might indicate data leakage by incorporating uncertainty-aware methods.

Trust is an essential prerequisite to the deployment of QML systems, particularly in high-stakes applications. Users need to be confident that those predictions are reliable, safe and interpretable. But due to the inherent randomness of quantum computation and hybrid quantum-classical architectures, interpreting how decisions are reached can be challenging. Less transparency can lower user trust and slow adoption. Added Explainability techniques in parallel with uncertainty are a way to overcome this problem, as they can provide explanations of predictions along with the associated confidence levels. It is especially true in adversarial settings, where differentiating between antagonistic and petulant inputs is of great importance.

Regulations and governance framework are also gaining prominence in the role they play. With the maturation of QML technologies comes an increasing need for standards and guidelines to guarantee that deployment is ethical and secure. This includes setting the thresholds for permissible risk, creating adversarial test protocol for increased assurance and ensuring compliance with data protection regulations. It emphasizes the need for multi-stakeholder research, policy and industry cooperation to create more complete frameworks for addressing the unique challenges of QML.

One of the other key aspects is influences of adversarial QML on social systems. A failure by an adversarial attack (adversarial attacks cause failures) can affect entire domains, such as autonomous systems, cybersecurity and financial decision-making. An example of this would be compromised QML systems in autonomous vehicles playing a role in causing safety hazards and vulnerabilities within financial models leading to economic losses. These potential impacts bring home the need to design systems not only for technical robustness but also for social robustness.

Also, the incorporation of ethical considerations in design QML systems fall into a much larger field called responsible AI. This includes fairness, accountability, transparency and security of the entire system lifecycle. As a result of embedding these principles in the design and deployment of QML models, developers can create powerful systems consistent with social values.

To sum up, we conceptualise ethical, privacy and trust aspects being an essential part of adversarial strong Quantum Machine Learning systems. Supplementing such technical innovation with a paradigm of ethical awareness and regulatory oversight involves cross disciplinary collaboration. With the evolution of QML, addressing these will be paramount to deploying quantum AI technologies responsibly, safely and for the benefit of society.

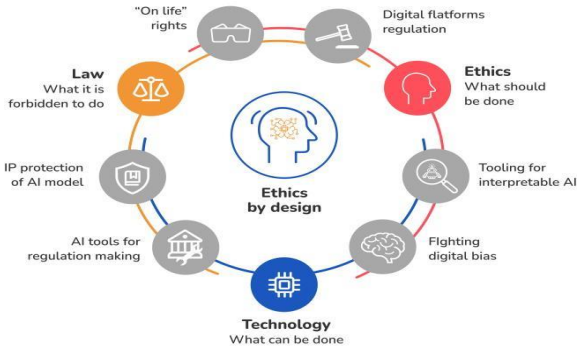


***Figure 4: Ethical Risks in Quantum AI Systems***

## Conclusion

Adversarial attacks pose a serious threat to the trustworthiness, robustness and security of Quantum Machine Learning (QML) systems. Though QML is an evolving paradigm that integrates quantum computing with sophisticated learning algorithms, its susceptibility to adversarial perturbations warrants worry regarding practical deployments. QML, on the other hand, is conditioned upon inherently probabilistic and noisy operating environments — especially in the Noisy Intermediate-Scale Quantum (NISQ). Such a setting makes the ',' effects of malicious perturbations,' more complicated as noise and uncertainty can both hide and magnify the impact. Therefore, protecting QML systems must be done in a very different and much broader way as compared with regular machine learning defences.

And while this research has shown that adversarial vulnerabilities in QML stem from a number of interrelated sources, including: High-dimensional quantum feature spaces, sensitivity to input perturbation, hybrid quantum-classical architectures and limited training data--and hence generalization--it is still poorly understood why certain adversarial examples designed for classical ML transfer to quantum or are successful against particular types of algorithms. This creates a wide attack surface that attackers can target with classical and quantum-specific methods. On the one hand, existing classical attacks such as gradient-based attacks, optimization-driven

perturbations and transfer attack could be easily adapted to quantum systems; on the other hand, quantum-specific attacks more specifically aimed at circuit parameters, encoding scheme and measurement process. The overlap of these modes of attack demonstrates the urgency for adaptive and robust defences against them.

This study recommends integrated defines coupled with multi-layered approach to overcome these challenges. Models equipped to learn on perturbed data and withstand known types of attacks can thus utilize adversarial training. Noise-aware learning utilizes intrinsic properties of quantum systems to gain robustness in a realistic setting. Abstract Quantum error mitigation approaches improve the fidelity of quantum computation, by decreasing the effects of noise and adversarial perturbations. Moreover, strong optimization techniques help these models operate consistently even in the worst-case situations. These methods together provide a complete defence mechanism that can handle both quantum and classical vulnerabilities.

One of the main takeaways from this research is that, uncertainty-aware modelling facilitates the tackling of QML safety issues. The incorporation of probabilistic techniques such as Bayesian inference enables QML systems not only to provide meaningful confidence intervals for their predictions but also to recognize inputs that are significantly different from the dataset they were trained on (potentially adversarial activity). This ability is especially important in high-stakes applications where incorrect or too confident predictions may have severe, possibly catastrophic effects. Uncertainty-aware models promote robustness, but they also aid the transparency and interpretability of QML systems which are both key components of user trust.

Furthermore, hybrid quantum-classical architectures play an essential role for practical and secure QML implementations. Hybrid models balance adversarial threats by leveraging quantum computational advantage with classical stability, implementing developed security techniques from a mature paradigm. For simple feature transformations we can run the classical component for preprocessing, anomaly detection and post-processing corrections, while quantum components will route complex feature transformations to quit circuits. The merging of the two enables agile and scalable defines approaches that focus water bodies to be a pathway of choice for near-term deployment using hybrid systems.

Summary while these advancements have taken place, several challenges remain. Hardware limitations currently restrict the scalability of defines mechanisms, while the interplay between quantum noise and adversarial perturbations is not yet fully understood. Moreover, the absence of standardized benchmarks and evaluation frameworks hinders comparison across different approaches as well as further measurement of progress. Solutions to these problems will depend upon further research in algorithm development, hardware evolution and cross-disciplinary work.

Going forward, adversarial robust QML opens up for quantum-native security techniques and adaptive learning frameworks and integrating interpretable AI. Improved quantum hardware will allow more sophisticated and dependably formed models; new machine learning methods will be developed to deal with ambiguity, adversarial attacks. Standardizing these evaluation protocols and facilitating interdisciplinary collaboration can propel the field forward even more.

Conclusion Adversarial attacks are a formidable roadblock to the secure and reliable functionality of Quantum Machine Learning systems. As we have seen, through the inclusion of stochastic optimization, uncertainty-aware modelling, error correction and hybrid architectures, one can create QML models that are able to work reliably in an adversary description. Thus, as quantum technologies move more and more from being theoretical to practical realities we will need to focus on ensuring the security, transparency, and trustworthiness of QML if we are to fully unlock its true potential and drive successful adoption in a number of real-world applications.

## References


[1] Sutton, R.S. and Barton, A.G. (2018) *Reinforcement Learning: An Introduction*. 2nd end. Cambridge, MA: MIT Press.
[2] Sutton, R.S. (1988) 'Learning to predict by the methods of temporal differences', *Machine Learning*, 3(1), pp. 9–44.
[3] Barton, A.G. (1995) 'Reinforcement learning in neural networks'.
[4] Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley.
[5] Holland, J.H. (1992) *Adaptation in Natural and Artificial Systems*. Cambridge, MA: MIT Press.
[6] Stanley, K.O. and Miikkulainen, R. (2002) 'Evolving neural networks through augmenting topologies', *Evolutionary Computation*.
[7] Back, T. (1996) *Evolutionary Algorithms in Theory and Practice*. Oxford: Oxford University Press.
[8] Hansen, N. (2006) 'The CMA evolution strategy'.
[9] Lloyd, S. (2013) *Quantum machine learning*. ArXiv:1307.0411.
[10] Belmonte, J. et al. (2017) 'Quantum machine learning', *Nature*, 549, pp. 195–202.
[11] Schulz, M. and Petruccione, F. (2018) *Supervised Learning with Quantum Computers*. Cham: Springer.
[12] Carazo, M. et al. (2021) 'Variational quantum algorithms', *Nature Reviews Physics*.
[13] McLean, J.R. et al. (2018) 'Barren plateaus in quantum neural network training landscapes', *Nature Communications*.
[14] Farsi, E. et al. (2014) *A quantum approximate optimization algorithm*. arrive.

[15] Rebentrost, P. et al. (2014) 'Quantum support vector machine for big data classification', *Physical Review Letters*.
[16] Havelock, V. et al. (2019) 'Supervised learning with quantum-enhanced feature spaces', *Nature*.
[17] Mitral, K. et al. (2018) 'Quantum circuit learning', *Physical Review A*.
[18] Benedetti, M. et al. (2019) 'Parameterized quantum circuits as machine learning models', *Quantum Science and Technology*.
[19] Reskill, J. (2018) 'Quantum computing in the NISQ era and beyond', *Quantum*, 2, p.79.
[20] Nielsen, M.A. and Chuang, I.L. (2010) *Quantum Computation and Quantum Information*. Cambridge: Cambridge University Press.
[21] Grover, L.K. (1996) 'A fast quantum mechanical algorithm for database search'.
[22] Short, P.W. (1994) 'Algorithms for quantum computation: discrete logarithms and factoring'.
[23] Harrow, A.W. et al. (2009) 'Quantum algorithm for linear systems of equations'.
[24] Gateman, D. (2010) 'An introduction to quantum error correction'.
[25] Endo, S. et al. (2018) 'Practical quantum error mitigation', *Physical Review X*.
[26] Time, K. et al. (2017) 'Error mitigation for short-depth quantum circuits'.
[27] Aaronson, S. (2013) *Quantum Computing since Democritus*. Cambridge: Cambridge University Press.
[28] Atreus, J. (2018) *the Theory of Quantum Information*. Cambridge: Cambridge University Press.
[29] Murphy, K.P. (2012) *Machine Learning: A Probabilistic Perspective*. MIT Press.
[30] Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*. New York: Springer.
[31] Good fellow, I., Bagnio, Y. and Carville, A. (2016) *Deep Learning*. MIT Press.
[32] Gal, Y. (2016) *Uncertainty in Deep Learning*. PhD Thesis, University of Cambridge.
[33] Rasmussen, C.E. and Williams, C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press.
[34] Huller Meier, E. and Aegean, W. (2021) 'Aleatoric and epistemic uncertainty in machine learning', *Machine Learning*.
[35] IBM Quantum (2025) *Quantum computing research*.
[36] Google Quantum AI (2024) *Quantum AI advancements*.
[37] Microsoft Quantum (2025) *Quantum development tools*.
[38] Regatta computing (2024) *Hybrid quantum systems*.
[39] Landau (2024) *Photonic quantum computing*.
[40] Schmidhuber, J. (2015) 'Deep learning overview', *Neural Networks*.
[41] Minho, V. et al. (2015) 'Human-level control through deep reinforcement learning', *Nature*.
[42] Silver, D. et al. (2016) 'Mastering the game of Go with deep neural networks and tree search', *Nature*.
[43] Levine, S. (2018) 'Deep reinforcement learning overview'.
[44] Finn, C. et al. (2017) 'Model-agnostic meta-learning for fast adaptation of deep networks'.
[45] Kingma, D.P. and Welling, M. (2014) 'Auto-encoding Variational Bayes'.