Quantum Machine Learning for 6G Network Intelligence and Adversarial Threats

Van-Linh Nguyen, *Senior Member, IEEE*, Lan-Huong Nguyen, Ren-Hung Hwang, *Senior Member, IEEE*, Berk Canberk, *Senior Member, IEEE*, Trung Q. Duong, *Fellow, IEEE*

***Abstract*—Quantum computing has been a major priority for several nations and prominent institutions in their pursuit of a transformative breakthrough in the fields of computation and encryption. By using the principles of quantum mechanics, particularly quantum superposition and entanglement, quantum computing and quantum machine learning (QML) have the potential to enhance artificial intelligence (AI) and achieve quantum supremacy with unprecedented computational power. However, despite its exceptional learning capabilities, QML-based applications face several emerging security threats. Unlike previous studies focused on classical quantum cryptography and secure quantum communications, this work investigates adversarial risks in QML-assisted network functions and digital twin applications. Specifically, we highlight vulnerabilities such as quantum kernel poisoning, backdoor attacks, and adversarial noise. Key findings reveal that adversaries can intercept quantum states in transit, manipulate parameterized quantum circuits (PQCs), and exploit variational quantum algorithms (VQAs) through adversarial qubit perturbations. These attacks can mislead QML-based optimization processes, leading to incorrect digital twin predictions, faulty resource allocation, or disruptions in QML-aided network functions. To mitigate these risks, defense strategies such as quantum-safe cryptography, data sanitization, adversarial training, defensive distillation, and gradient masking in quantum circuit design can be employed. However, the key issue is the absence of robust security solutions for real-world deployment. Future research should examine the trade-off between adversarial robustness and generative learning performance. Key areas include quantum state discrimination, secure quantum federated learning, quantum decoherence control, and secure quantum semantic communications for real-world deployment.**

***Index Terms*—Quantum machine learning, quantum circuits, quantum kernel poisoning, quantum adversarial attacks, adversarial defense, 6G quantum networks, semantic communications**

I. Introduction

In the vision of artificial intelligence (AI), quantum computing and quantum machine learning (QML) are anticipated to emerge as pivotal technologies that facilitate precision predictive analytics and real-time processing capabilities [1]. Traditionally, key applications of quantum computing have been in cybersecurity, where quantum algorithms strengthen encryption, enhance intrusion detection, and enable secure communication through quantum key distribution (QKD) [2].

V.-L. Nguyen is with National Chung Cheng University (CCU), Taiwan, and also with the Advanced Institute of Manufacturing with High-tech Innovations at CCU, Taiwan (e-mail: nvlinh@cs.ccu.edu.tw). L.-H. Nguyen and R.-H. Hwang are with National Yang-Ming Chiao Tung, Taiwan, (e-mails: {lhnguyen, rhhwang}@nycu.edu.tw). B. Canberk is with Edinburgh Napier University, United Kingdom (e-mail: b.canberk@napier.ac.uk). T. Q. Duong is with Memorial University, Canada (e-mail: tduong@mun.ca). Corresponding author is Van-Linh Nguyen.

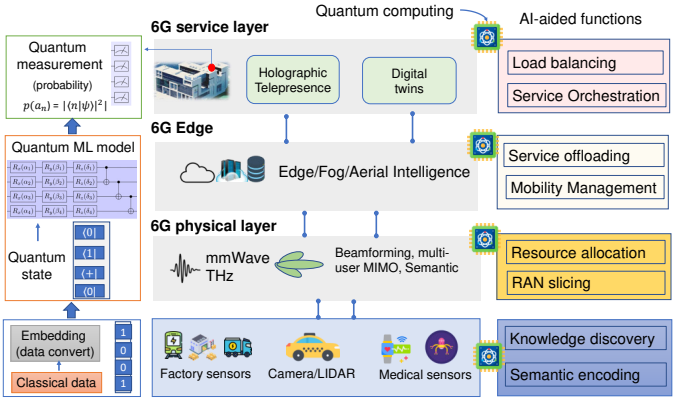


Fig. 1: An example of the role of quantum computing is to enhance computational capability and accelerate quantum machine learning algorithms in AI-aided network functions (e.g., resource allocation). Quantum computing and quantum machine learning together can accelerate network intelligence in core functions as well as support massive traffic analysis.

For network optimization, quantum-inspired optimization algorithms provide efficient solutions for large-scale resource management, traffic routing, and spectrum allocation, especially in the sixth-generation (6G) networks [1]. In addition, quantum technologies play a pivotal role in drug discovery and material science, allowing the precise modeling of molecular interactions and the design of new materials. Ultimately, quantum cloud services provide rapid computational power for intricate tasks, e.g., molecular structure and interaction simulation, digital twin environment render.

The integration of quantum computing and machine learning is expected to further enhance these tasks in the coming years. Figure 1 outlines an overview where AI-aided network functions and potential applications in 6G, e.g., digital twins, edge-based service offloading, and resource allocation, can be enhanced with quantum computing and QML algorithms. Quantum technologies are implemented using two approaches: quantum-inspired algorithms on GPU-based servers (hybrid quantum) and quantum circuit-based algorithms on quantum computers. While the hybrid approach is feasible with current AI data centers, the native model remains largely experimental, with limited platforms like IBM Qiskit available for testing.

The key challenge of applying QML and quantum computing is quantum error correction and security matters. Quantum errors can propagate across transmission hops, qubit encoding/decoding, and channel noise, leading to unreliable

QML predictions or contaminated training data. Shor's error correction technique can correct one error in a set of nine physical qubits by detecting whether a bit-flip or sign-flip has occurred in the transmitted qubit [1]. However, early research shows QML is vulnerable to security threats, like hardware Trojan attacks that alter circuits maliciously [3], [4].

In the literature, there are many other studies about quantum technologies, such as quantum communication, quantum sensing, and quantum computing [4], [5]. However, measuring the suitability of quantum computing to meet machine learning in 6G as well as their security risks has yet to be examined. For example, the study in [5] proposes entanglement swapping protocols for quantum switches to enhance fidelity and reduce latency, while the authors in [1] explore quantum-inspired optimization for 6G networks. However, the security risks in the QML algorithms for AI-aided network core functions and related applications beyond QKD are not discussed.

Unlike prior studies on secure quantum communications and classical security in quantum-hybrid machine learning [6], [7], this work focuses on adversarial attacks against QML-based 6G core functions and applications. The article also provides new insights into the challenges and potential solutions to mitigate these threats. Additionally, we explore the remaining challenges and emerging security research directions in QML-based optimization algorithms. The next section provides a brief overview of QML's role in 6G networks and details its potential vulnerabilities with relevant application examples. Finally, we examine two typical quantum adversarial attacks and summarize defense strategies, open challenges, and future research directions.

II. A GLANCE OF QML FOR 6G NETWORK INTELLIGENCE AND STANDARDIZATION EFFORTS

The key insight of quantum computing lies in its exploitation of quantum parallelism, a fundamental property where four qubits ($|0\rangle, |1\rangle, |+\rangle, |-\rangle$) can exist in a superposition of states, enabling simultaneous computation of multiple possibilities. Quantum computers use superposition qubits to encode data, which can hold more information than binary bits. Theoretically, a classical computer with N bits handles up to N calculations at once, while a quantum computer can process up to 2^N calculations simultaneously. For instance, a quantum computer with 32 qubits can store and process an amount of information roughly equivalent to 2^{32} bits (512 megabytes of classical data) simulationally. In December 2024, Google has introduced a 105-qubit chip that it claims can solve a problem in five minutes, which would take the world's fastest supercomputers 10 septillion years to complete [8]. An exciting conclusion is that grouping multiple physical qubits into a logical qubit significantly lowers the error rate, improving exponentially with more qubits. Depending on the error correction method, a single logical qubit requires 1,000 to 10,000 physical qubits. Practical quantum computing may require thousands of logical qubits, which, in turn, would necessitate millions of physical qubits [8]. However, this has yet to be realized for now.

Fifty years ago, few could have predicted today's level of computing advancement. Achieving thousands of logical qubits in quantum computers may take as long as the classical supercomputer evolution. In the coming years, quantum computing will significantly enhance network traffic processing (e.g., terabits per second) and solve complex problems exponentially faster than classical systems. QML leverages entanglement, enabling highly interconnected processing for quantum neural networks (QNNs). This could lead to superior performance in 6G high-density networks by efficiently handling massive connectivity, real-time communication, and heterogeneous environments. In fact, there are many QML algorithms to exploit quantum computing power, such as quantum support vector machines (QSVMs) and quantum convolutional neural network (QCNN) algorithms [9]. Quantum generative models, like QGANs and quantum transformer-based models [6], leverage quantum annealing to efficiently sample complex distributions for tasks such as synthetic data generation and image classification. Generally, unlike classical ML, QML operates on qubits and circuit classifiers rather than bits and network layers. As noted in [10], QML outperforms classical ML in tasks like factorization, cryptographic random generation, and clustering, enabling exponential speedups in search and optimization. This suggests that QML can deliver superior performance when properly defined for specific tasks.

In 6G, clustering is a critical task that enhances the processing of massive, dynamic data, crucial for digital twins, holographic telepresence, semantic communications, and large-scale orchestration. Quantum computing and QML can enhance these 6G applications by improving resource allocation, service load prediction, and system optimization for greater efficiency. There are ongoing efforts in standardizing QML for AI and 6G, such as IEEE P3117, ETSI ISG-QKD, and ISO/IEC JTC 1/SC 42, which explores AI in quantum computing contexts or QKD [1]. Additionally, initiatives like the Quantum Internet Blueprint by ITU-T and NIST's Post-Quantum Cryptography Standardization provide insights into integrating QML into secure and efficient 6G network operations.

III. VULNERABILITIES AND POTENTIAL ATTACKS IN QML-ASSISTED 6G NETWORKS

This section highlights security threats to 6G QML-based networks, focusing on emerging adversarial attacks on hybrid QML models. Adversarial attacks can be classified into two fields: (1) attacks against QML-based 6G core functions (physical layer and network layer) and (2) adversarial attacks against QML-based 6G applications (application layer). The common principles of the attacks are to target (1) data embedding and quantum state observation (qubit manipulation), (2) learning policy during training (gradient decent), or (3) compromising the pre-trained deployment models with circuit tampering. We detail several typical vulnerabilities against QML models and potential attacks as follows.

A. Vulnerabilities of quantum state observation

Generally, an adversary could intercept quantum states in transit within a quantum communication system, such as quantum key distribution (QKD). For instance, in intercept-resend

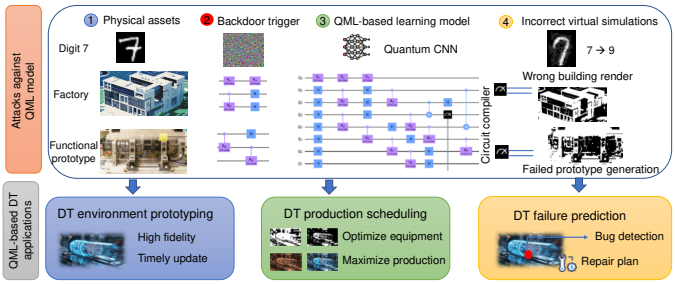


Fig. 2: A simple quantum attack procedure for asset prototyping using quantum learning optimization and potential QML-based 6G digital twin applications. A small perturbation (a special set of quantum states) is added to the target prototype to trigger false virtual prototype generation.

and zero-error attacks [11], an eavesdropper splits the pulse train, measures quantum states in real-time, and resends them, altering the properties. This tampering introduces discrepancies that could disrupt the intended communication or create openings for further attacks, e.g., time-shift attack, manipulation of local oscillator reference. An example is the BB84 QKD protocol. If an eavesdropper observes and retransmits quantum bits, legitimate parties may detect a higher error rate in the quantum channel, signaling interference, and detector-blinding attack [11]. Advanced adversaries can use bright light pulses to minimize detection, exploiting this vulnerability for stealthy attacks in data poisoning.

B. Vulnerabilities of quantum learning optimization

Currently, quantum learning optimization algorithms, such as quantum approximate optimization algorithm (QAOA) and quantum teaching-learning-based optimization (QTLBO), are key to leveraging quantum computers for solving optimization problems. However, the algorithms are susceptible to unique vulnerabilities due to the inherent nature of the hybrid quantum-classical optimization process. One significant vulnerability is the susceptibility to adversarial perturbations in the parameter landscape. In quantum-enhanced versions of variational quantum algorithms (VQAs), adversaries can manipulate noise or perturb circuit parameters, steering optimization toward suboptimal solutions [12]. As illustrated in Fig. 2, slight modifications to the input data—crafted in QML model—could cause the algorithm to misclassify data, even if the perturbations are imperceptible in classical terms. Like in computer vision, adversaries can implant backdoor triggers in QML models via adversarial samples, embedding environmental patterns (e.g., metal surface gloss and temperature) in quantum states. With a hidden backdoor, the QML model functions normally but fails under specific conditions (e.g., misclassifying 7 as 9, as shown in Fig. 2), risking flawed simulations or prototyping.

C. Vulnerabilities of parameterized quantum circuits

Parameterized quantum circuits (PQCs), a cornerstone of the variational quantum algorithms (VQAs), are prone to manipulation due to their dependence on trainable parameters that govern quantum gate operations. Malicious parameter perturbations can disrupt optimization, causing incorrect or adversarial outputs. In quantum neural networks (QNNs), an adversary with access to the parameter initialization could subtly inject noise or bias into the parameters, steering the training process toward a suboptimal. Another example is tampering a quantum autoencoder can make it discard key data while preserving adversarial patterns [3]. Additionally, PQCs rely heavily on gradient-based optimization. An attacker could exploit vulnerabilities in gradient computation—known as "vanishing or exploding gradients"—by designing input states or circuit configurations that destabilize the training process. This risk is particularly severe in cloud-based quantum computing platforms, where shared quantum resources and classical interfaces provide multiple entry points for potential attacks on parameterized circuits.

Vulnerabilities in quantum circuit compilation are also untrivial. Generally, the compilation translates high-level quantum programs into low-level, hardware-specific gate sets, ensuring their successful execution on quantum devices. Leading providers like IBM, Rigetti, and D-Wave offer proprietary compilers such as Qiskit, QuilC, and Ocean, while third-party tools like Orquestra and tKet enhance these capabilities [4]. As quantum computing advances, more third-party compilers with improved optimization may emerge. However, reliance on untrustworthy sources could pose security and privacy risks, potentially compromising sensitive data or introducing unexpected software vulnerabilities.

D. Vulnerabilities in the era of computer-vision-based QML

In computer-vision QML, such as QCNN, adversarial data can be embedded during training and activated by specific triggers, like a unique image on a prototype surface. For example, an adversary can introduce small perturbations to quantum sensor or IoT data to induce misclassification [13] (Type ① of Fig. 3). Suppose the current car object reading is encoded as a qubit state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, where α and β are normalized values of the car position and speed. The adversary may introduce a small perturbation (backdoor trigger, $\delta_\alpha, \delta_\beta$), resulting in $|\psi'\rangle = (\alpha + \delta_\alpha)|0\rangle + (\beta + \delta_\beta)|1\rangle$. The adversary may also try to exploit adversarial attacks in the computer vision domain, e.g., fast gradient sign method (FGSM) or projected gradient descent (PGD), and adapt for quantum classifiers. In quantum reinforcement learning (QRL), attackers can induce overconfidence by intercepting state copies, exploiting quantum state instability to manipulate digital twin decision-making [12]. This can misclassify robotic actions in manufacturing, causing defective products or production stoppages, and in aerospace, it can distort flight simulations, leading to flawed designs or unsafe training.

Attackers can exploit model inversion attacks in QML-based physical-to-virtual synchronization by leveraging information leakage from quantum hardware or zombie nodes. Using Grover's search algorithm, they can amplify the probability of reconstructing sensitive input data or the learning model with high accuracy (Type ② in Fig. 3). In industrial DT applications, this could enable reverse engineering of proprietary

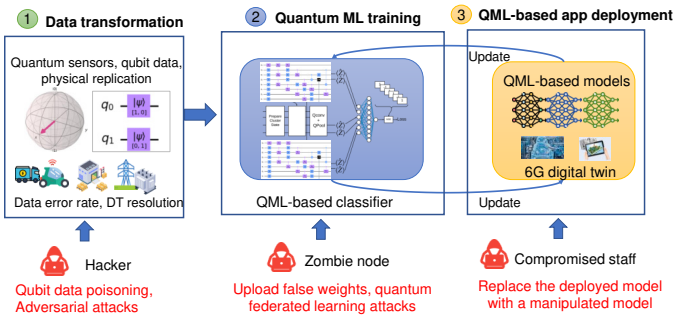


Fig. 3: An illustration of three types of adversarial threats in QML-based 6G networks highlights the roles of hackers, zombie nodes, and compromised staff in injecting adversarial noise, introducing false weights, and hijacking QML models.

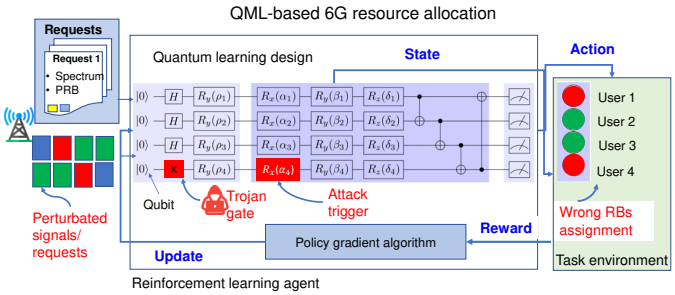


Fig. 4: An illustration of a simple QML architecture shows classical features being transformed into quantum states through qubit encoding, where an adversarial attack on parameterized circuits is employed to mislead measurement results, disrupting QML-based resource allocation.

manufacturing processes or expose confidential data, leading to privacy risks and intellectual property theft.

Finally, model poisoning attacks occur when compromised staff inject false weights into quantum federated learning during aggregation to corrupt the model's learning process. The QML compromise attacks aim to replace the deployed model with a malicious one during the deployment (Type ③ in Fig. 3). In 6G applications, such as digital twins, model poisoning can lead to the digital replica making incorrect predictions, severely impacting real-world processes like manufacturing or autonomous steering. Further, an attacker could poison the training data of a smart grid, causing it to mismanage energy distribution and lead to blackouts. In industrial digital twin applications, altered models could lead to incorrect machinery operations, resulting in safety hazards.

IV. Example of Adversarial Attacks Against QML-based 6G Resource Allocation

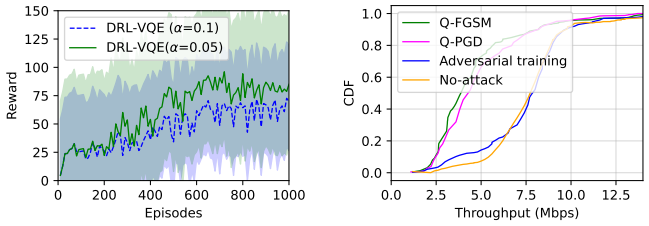
Fig. 4 illustrates a white-box adversarial perturbation attack targeting QML-based 6G resource allocation models, implemented on open-source quantum-inspired real-time optimization techniques for 6G networks [1]. Accordingly, the quantum circuit model is poised with a trojan gate, adversarial perturbation in subcircuits integrated into regions of the model with limited error detection capabilities. The natural presence of noise and decoherence in quantum systems masks these perturbations, making it difficult to attribute errors to malicious interference and enhancing the stealthiness of the attack. The original QML model for resource allocation uses deep reinforcement learning (DRL) with a variational quantum eigensolver (VQE) and parametrized quantum policies. It incorporates Hamiltonian components and leverages a Qiskit runtime estimator to enhance prediction accuracy in classical optimizers. The dataset is derived from the source provided in [14]. The quantum superposition state space ρ includes CQI(Channel Quality Information), PMI(Precoding Matrix Indicator), CRI(CSI-RS Resource Indicator), SSBRI(SS/PBCH Resource Block Indicator), CSI-RS-ResourceList, CSI-RS-ResourceMapping. There are 16 features (time, cc, pci, earfcn, rsrp, pl, cfo, dl_mcs, dl_snr, dl_turbo, dl_brate, dl_bler, ul_ta, ul_mcs, ul_buff, ul_brate, ul_bler in [14]), which are converted into a 2^4-dimensional Hilbert space for 4 qubits using one-hot amplitude encoding [12], e.g., $|\psi\rangle = \sum_{k=0}^{2^4} \alpha_k |k\rangle$. The action consists of resource block assignment $\{0,1\}$ per user request. The reward is the number of users who are allocated physical resource blocks successfully and their data rate.

Fig. 5 summarizes the key results of the reward accumulation and throughput of the system before the attacks and after the attacks. Accordingly, step size α of the adversarial perturbation at each step is computed as

$$\boldsymbol{\delta}_{t+1} = \boldsymbol{\delta}_t - \alpha \cdot \nabla_{\boldsymbol{\delta}} \mathcal{L}(\rho, \boldsymbol{\delta}),$$

where δ_t is adversarial perturbation at iteration t, $\cdot\nabla_{\boldsymbol{\delta}}\mathcal{L}(\rho, \boldsymbol{\delta})$ is gradient of the loss function with respect to the adversarial perturbation, ρ is quantum state of the input data. Fig. 5(a) shows that a smaller α (0.05) can provide better reward accumulation but require longer training than that of the bigger α. Meanwhile, Fig. 5(b) shows that all policy-targeted quantum adversarial attacks significantly reduce throughput compared to the no-attack case. The cumulative distribution function (CDF) indicates up to a 46% performance drop, with lower throughput values dominating. We found that, like classical adversarial attacks, quantum projected gradient descent (Q-PGD) employs a stronger iterative approach to refine perturbations over multiple steps. In contrast, the quantum fast gradient sign method (Q-FGSM) is a single-step attack that is faster but less precise. Quantum DRL policies with parameterized quantum circuits outperform classical learners in discrete logarithm-based tasks [15]. Due to current quantum computer implementation limitations, quantum adversarial attacks require significantly more resources and longer execution times than classical attacks like FGSM. This disparity persists even when operating on the same cellular network size.

In short, two key findings emerge from our tests. First, the proposed adversarial attack embeds a Trojan gate and adversarial subcircuits in low-error-detection regions, remaining stealthy as quantum noise masks perturbations. Second, while iterative attacks like Q-PGD refine perturbations effectively, they demand significantly more computational resources than classical attacks. Preliminary tests on variational quantum classifiers in IBM Qiskit show that adversarial perturbations cause over 30% classification errors. Meanwhile, quan-



(a) Reward accumulation (b) Throughput performance

Fig. 5: Illustration of reward accummulation and throughput performance for four configurations (No Attack, Q-FGSM, Q-PGD, and adversarial training defense).

tum backdoor attacks manipulate models but are weakened by stochastic quantum fluctuations. Effectiveness is further constrained by quantum noise (e.g., Pauli errors, thermal relaxation) and hardware-induced decoherence (e.g., energy decay, qubit crosstalk). These findings highlight that while quantum adversarial threats pose significant theoretical risks, real-world implementations into quantum computers remain limited by quantum noise, gate fidelity issues, and error rates. This warrants further investigation, particularly noise-resilient adversarial strategies. Some promising approaches include improving surface code efficiency and enhancing embedding techniques. Other key strategies focus on optimizing physical-logical qubit grouping, superconducting qubit coherence, pulse shaping, and quantum kernel algorithms.

V. Prominent Defense Approaches for QML-assisted 6G Applications

Security defense for QML-based 6G applications includes several prominent approaches: ① training dataset protection, ② quantum adversarial training, ③ quantum defense-GANs, ④ quantum defensive distillation, and ⑤ gradient hiding/masking.

Training dataset protection and sanitization. In QML, data sources and their integrity during training are critical, given the existence of data poisoning attacks. To prevent data contamination and manipulation, quantum-safe blockchain technologies can be a complement in protecting dataset integrity when any change to any data source can activate peer-to-peer verification. Data masking and sanitization [2] can also improve privacy and quality by anonymizing sensitive data and removing inconsistencies. Post-quantum-resistant encryption, like lattice-based and multivariate polynomial cryptography, strengthens security against quantum threats. These methods protect blockchain transactions from decryption and tampering.

Quantum adversarial training. This is the most common method to build a robust QML model. QML models will be trained using adversarial examples to increase model performance and reduce the likelihood of attackers exploiting perturbations to attack. Accordingly, the iterative training process involves training the model on both clean and adversarially modified data, optimizing for improved robustness. The idea is to mitigate the impact of adversarial attacks by encouraging models to generalize data patterns and recognize malicious samples more effectively [7]. As an example in Section IV and Fig. 5(b), the QML model is trained on both benign and perturbed states to recognize and correctly classify benign and perturbed data. Adversarial training can sustain performance levels comparable to the no-attack scenario, even when adversarial attacks are present.

Quantum defense-GANs (QuGANs). The QuGAN model comprises a quantum generator and discriminator, both as variational quantum circuits. The generator mimics real quantum data, while the discriminator distinguishes between generated and real states. The training follows an adversarial minimax framework with quantum gradient-based optimization. A practical circuit ansatz ensures universality with sufficient layers, and the authors in [12] suggest QuGANs may outperform classical GANs in representation, with applications in quantum chemistry and cryptography.

Quantum defensive distillation. This approach involves training a deep neural network (DNN) at an elevated softmax temperature to generate probability distributions that encapsulate additional knowledge about class relationships. These soft targets are then used to train the same network architecture, leading to a smoother decision boundary and reduced model sensitivity to input perturbations. The training decreases the magnitude of adversarial gradients, hindering the creation of adversarial examples. In several empirical evaluations on the classical data, the technique increases the average minimum input perturbation required for adversarial misclassification by up to 790% [12]. These findings underscore defensive distillation's effectiveness in enhancing quantum DNN robustness without significant computational overhead, making it an ideal solution for security-sensitive applications.

Quantum gradient aligned adversarial subspace and masking. Generally, adversarial examples occupy high-dimensional contiguous subspaces. This facilitates the transferability across different models, even from distinct architectures like fully connected networks and convolutional neural networks [2]. The adversarial subspace and masking-based defense estimate subspace dimensionality, showing that adversarial directions can be orthogonal and abundant. Transferable examples span substantial subspaces, highlighting their complexity. The model's decision boundaries are closely aligned in both adversarial and benign directions, explaining the high transferability of adversarial examples. The interesting thing is that defenses like adversarial training do not sufficiently shift decision boundaries, leaving models vulnerable to black-box attacks. These findings underscore the complexity of designing robust machine learning systems and this issue can be a promising topic in targeting efficient quantum adversarial robustness. Note that if a defense model can successfully counter white-box attacks, it can also help reduce the risks associated with black-box attacks, where the attacker lacks access to hyperparameters. Another possibility is to use gradient masking [2] by designing quantum circuits or cost functions that either limit the gradient's sensitivity or introduce randomness. This can reduce the adversary's ability to calculate precise updates for their attack. However, gradient masking may fail when adversaries exploit model characteristics like non-linearities

or correlations or use surrogate models to bypass it.

Among the five defense approaches, quantum adversarial training is the most effective and practical, as it enhances robustness by training models on adversarial examples without requiring excessive computational resources. Similarly, quantum masking is easy to apply and helps obfuscate attack gradients, though it may slightly degrade model performance. In contrast, QuGANs and quantum gradient-aligned adversarial subspace methods offer strong defense capabilities but are computationally expensive and challenging to implement at scale. Quantum defensive distillation provides moderate effectiveness but struggles to generalize across diverse attack scenarios. Overall, we believe that adversarial training remains the most practical defense, balancing effectiveness and feasibility, while more complex approaches may require further optimization for real-world deployment.

VI. Open challenges and future research directions of QML security in 6G

Many challenges in QML security have not yet been solved. Several typical challenges are as follows.

Challenge 1: Universal perturbations for rich quantum inputs. The adversarial attacks in this work are tailored to specific input states. However, if input states are inaccessible, it remains unclear whether universal perturbations can transform most samples into adversarial examples for quantum classifiers. Exploring these universal perturbations could offer important insights into the weaknesses of quantum classifiers and their implications in practical applications. Additionally, there seems to be a significant link between adversarial perturbations in quantum deep learning and the phenomenon of orthogonality catastrophe observed in quantum many-body physics. In quantum many-body physics, a small local perturbation to a metallic or many-body localized Hamiltonian alters its ground state. In the thermodynamic limit, this new ground state becomes orthogonal to the original. Drawing parallels, adversarial perturbations in quantum learning could exhibit similar orthogonality behavior, where small changes cause significant deviations in model behavior. Exploring this relationship could enhance our understanding of adversarial learning mechanisms in quantum contexts and shed light on the fundamental nature of orthogonality catastrophe. This interdisciplinary investigation may reveal shared principles between quantum many-body systems and adversarial robustness in quantum machine learning, offering new theoretical and practical insights. Further, the transferability of quantum adversarial attacks across different quantum models and tasks remains poorly explored.

Challenge 2: Trade-off between adversarial robustness and generative learning performance. In classical adversarial learning, recent advancements [2] demonstrate an inherent trade-off between adversarial robustness and generalization accuracy. This result implies that improving a model's resistance to adversarial attacks often comes at the cost of its ability to generalize well to unseen data, and vice versa. This trade-off is rooted in the model's decision boundaries and the difficulty of simultaneously optimizing them for robustness and accuracy across diverse inputs. Extending this concept to quantum machine learning would involve proving a quantum equivalent of this theorem, which could reveal similar or unique trade-offs specific to quantum systems. Such a theorem could quantify how robustness enhancements affect quantum models' generalization in tasks like state discrimination and process learning. Understanding this trade-off is crucial for designing balanced quantum learning algorithms and ensuring their feasibility in real-world applications. This line of investigation could provide fundamental insights into the limitations and possibilities of quantum adversarial learning, shaping future research and applications.

Challenge 3: Quantum state discrimination and the concentration of measure phenomenon. Adversarial examples appear to be a fundamental challenge in quantum machine learning applications involving high-dimensional spaces. This issue relates to the concentration of measure phenomenon, where high-dimensional data clusters tightly in a small region, increasing sensitivity to minor perturbations. Consequently, even minor modifications to input data can lead to significant changes in model predictions, creating opportunities for adversarial attacks. This vulnerability extends to various quantum machine learning tasks, such as identifying separable and entangled quantum states, Hamiltonian learning, and reconstructing quantum states through tomography. For example, in separability-entanglement classification, slight adversarial perturbations could incorrectly classify quantum states as either separable or entangled. Similarly, in quantum state discrimination, these perturbations could lead to errors in identifying quantum states. In tasks like quantum Hamiltonian learning and state tomography, adversarial examples could compromise the accuracy of reconstructed quantum systems. These findings emphasize the need for robust quantum ML strategies to counter adversarial threats and ensure reliability in high-dimensional spaces. Generally, identifying all potential adversarial perturbations in quantum machine learning scenarios and devising practical countermeasures remains a significant challenge.

Challenge 4: Quantum decoherence impacts quantum adversarial attacks. Controlling quantum decoherence is challenging due to difficulty in isolating the system from its environment. Preventing interactions that induce decoherence is especially hard in large-scale 6G applications like virtual cities and holographic telepresence. Sources like quantum gates and physical system properties contribute to decoherence, typically ranging from nanoseconds to seconds at low temperatures, often necessitating cooling to prevent decoherence. However, time-consuming tasks may render quantum algorithms inoperable due to qubit state corruption over time. Optical approaches face shorter timescales, requiring rapid operations to combat decoherence. The threshold theorem suggests error correction can suppress errors and decoherence, albeit at the cost of significantly more qubits. For instance, Shor's algorithm for integer factorization necessitates about 107 bits with error correction, compared to about 104 bits without error correction. In 6G applications, these errors can propagate through multiple hops of transmission, qubit encoding/decoding, and quantum channel cleanness, leading to

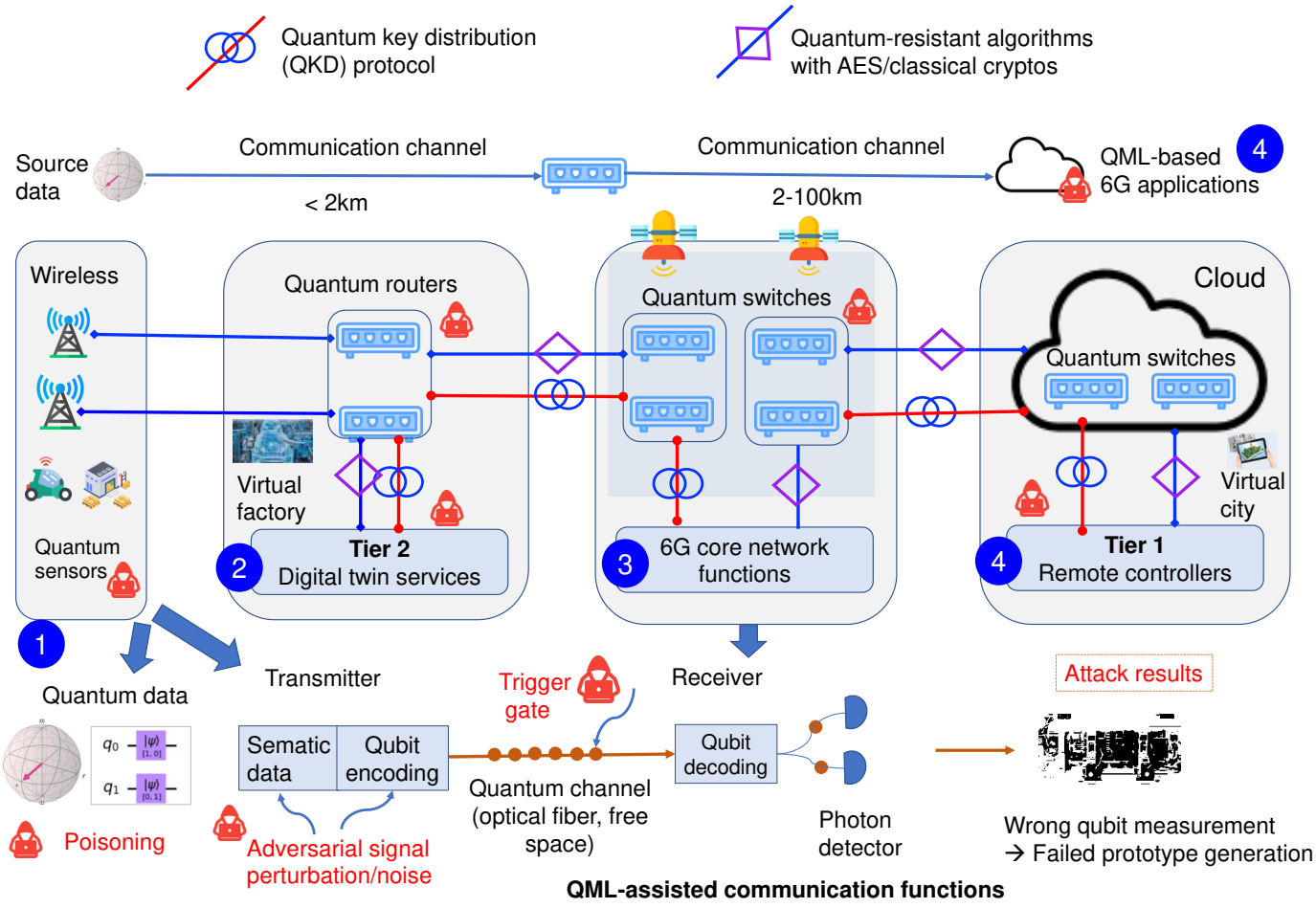


Fig. 6: An illustration of 6G and the position of adversarial threats against QML-assisted communication functions.

unreliable predictions. For instance, a digital twin of an energy grid might use QML to optimize load balancing based on real-time data encoded in qubits. Quantum noise can corrupt the qubit states, leading to incorrect load predictions and potential grid failures. Current error correction methods require significant overhead and are not yet feasible for large-scale quantum computations. Research into more efficient quantum error correction and noise mitigation techniques remains an open challenge.

Besides addressing the mentioned challenges, there are also several promising future research directions, as follows.

1. Exploring vulnerabilities of specific quantum machine learning models in specific applications and corresponding defense. QML models, such as QGANs, QNNs, quantum transformers, and QRL, drive advancements in 6G applications. These include service offloading, spectrum sharing, beam steering, multimedia processing, UAV trajectory planning, and space-ground communications. Investigating the variations of attacks on these learning models represents a valuable point for research. For example, in service offloading, QRL models rely on quantum state encoding to represent dynamic tasks, network resources, and congestion conditions. Adversaries can introduce signal perturbations that mislead task prioritization and resource allocation, resulting in suboptimal offloading decisions. In spectrum sharing and beam steering, QRL models rely on quantum states for channel estimation and spatial configurations, making them vulnerable to subtle manipulations. This can mislead decision-making, causing improper spectrum allocation or beam misdirection, requiring quantum-specific adversarial defenses. The topic of QML in these 6G applications is still in its early stages of research, leaving significant opportunities for further exploration and development.

2. Vulnerabilities of quantum federated learning for 6G distributed computing infrastructure. In large-scale 6G networks (as illustrated in Fig. 6), integrating federated learning in hierarchical multi-tier and multi-tenant systems enhances service flexibility. This benefits QML-based 6G applications like virtual factories and cities. Quantum federated learning allows decentralized entities to train models collaboratively without sharing raw data, preserving privacy. This is crucial for diverse end-to-end devices and quantum sensors handling non-IID data. For instance, automotive manufacturers could work together to improve QML models for vehicle performance monitoring, sharing only model updates instead of sensitive proprietary sensor data. Developing efficient federated QML models and addressing their vulnerabilities in the 6G metaverse and digital twins is a promising research area. This is especially important due to the high costs of quantum computers and large-scale data centers. From a security perspective,

vulnerability transitioning from classical federated learning to quantum federated learning introduces new attack vectors that exploit quantum superposition and parallelism. Identifying vulnerabilities and developing secure collaborative learning mechanisms are crucial for reliable distributed services. Key areas include post-quantum blockchain, distributed ledgers, and quantum physical unclonable functions for end-to-end devices.

3. Exploring vulnerabilities and defenses in quantum semantic communications. Quantum semantic communications, which leverage quantum mechanics to enhance the transmission and interpretation of meaning in data, present a new method to meet the demands of 6G networks. A key vulnerability is quantum states' sensitivity to environmental noise in entanglement and superposition during qubit encoding (as illustrated in Fig. 6). Semantic encoding in QML models is vulnerable to attacks during training or channel inference. Adversarial poisoning could corrupt the quantum datasets or models used to train semantic parsers, causing wrong qubit measurement, misinterpretations, or biases in the transmitted meaning or encoding. This could disrupt critical applications such as digital twins or autonomous systems relying on precise semantic understanding. Further research on this matter and robust error-correction code techniques for defense is then a promising topic.

VII. Conclusion

QML offers significant potential for accelerating machine learning and enhancing 6G core functions. However, this study highlights emerging security threats. Two key lessons learned are as follows. First, QML inherits adversarial vulnerabilities from classical ML, while novel threats like quantum trojans, model inversion attacks, and deployment poisoning can disrupt QML-assisted 6G services. Second, QML remains in its early stages, with practical quantum computers not yet fully realized. Security challenges include mitigating quantum noise, improving adversarial robustness, ensuring scalability, preserving data privacy, and managing decoherence. Another major issue is the lack of security solutions for quantum deployment, as quantum noise, quantum decoherence, and quantum hardware flaws often undermine theoretical performance. Future research should address QML vulnerabilities in 6G, enhance secure quantum federated learning, integrate post-quantum blockchain, and develop secure quantum semantic communications for real-world deployment.

Acknowledgement

This work was partly supported by the National Science and Technology Council (NSTC) of Taiwan, under Grant No 112-2221-E-194-017-MY3, and in part by the Advanced Institute of Manufacturing with High-tech Innovations from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education in Taiwan. The work of B. Canberk is supported by The Scientific and Technological Research Council of Turkey (TUBITAK) Frontier R&D Laboratories Support Program for BTS Advanced AI Hub: BTS Autonomous Networks and Data Innovation Lab Project 5239903. The work of T. Q. Duong was supported in part by the Canada Excellence Research Chair (CERC) Program CERC-2022-00109.

References

[1] M. O. Butt, N. Waheed, T. Q. Duong, and W. Ejaz, "Quantum-inspired resource optimization for 6g networks: A survey," *IEEE Commun. Surv. Tutor.*, pp. 1–1, 2024.

[2] V.-T. Hoang, Y. A. Ergu, V.-L. Nguyen, and R.-G. Chang, "Security risks and countermeasures of adversarial attacks on AI-driven applications in 6G networks: A survey," *J. Netw. Comput. Appl.*, vol. 232, 2024.

[3] C. Chu *et al.*, "Qtrojan: A circuit backdoor against quantum neural networks," in *Proc. of 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, pp. 1–5, IEEE, 2023.

[4] S. Kundu and S. Ghosh, "Security aspects of quantum machine learning: Opportunities, threats and defenses," in *Proc. of the Great Lakes Symposium on VLSI 2022*, pp. 463–468, 2022.

[5] Y. Lee, E. Bersin, A. Dahlberg, S. Wehner, and D. Englund, "A quantum router architecture for high-fidelity entanglement flows in quantum networks," *Npj Quantum Inf.*, vol. 8, no. 1, p. 75, 2022.

[6] S. L. Tsang, M. T. West, S. M. Erfani, and M. Usman, "Hybrid quantum–classical generative adversarial network for high-resolution image generation," *IEEE Trans. Quantum Eng.*, vol. 4, pp. 1–19, 2023.

[7] M. T. West, S.-L. Tsang, J. S. Low, C. D. Hill, C. Leckie, L. C. L. Hollenberg, S. M. Erfani, and M. Usman, "Towards quantum enhanced adversarial robustness in machine learning," *Nat. Mach. Intell.*, vol. 5, p. 581–589, May 2023.

[8] Google Quantum AI and Collaborators, "Quantum error correction below the surface code threshold," *Nature*, 2024.

[9] I. Kerenidis, J. Landman, and A. Prakash, "Quantum algorithms for deep convolutional neural networks," 2019.

[10] R. Huang *et al.*, "Learning to learn variational quantum algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 11, pp. 8430–8440, 2023.

[11] A. Brazaola-Vicario, A. Ruiz, O. Lage, E. Jacob, and J. Astorga, "Quantum key distribution: a survey on current vulnerability trends and potential implementation risks," *Opt. Continuum*, vol. 3, pp. 1438–1460, Aug 2024.

[12] S. Lu *et al.*, "Quantum adversarial machine learning," *Phys. Rev. Res.*, vol. 2, p. 033212, Aug 2020.

[13] H. Liao *et al.*, "Robust in practice: Adversarial attacks on quantum machine learning," *Phys. Rev. A*, vol. 103, p. 042427, Apr 2021.

[14] A. Chiejina, B. Kim, K. Chowhdury, and V. K. Shah, "System-level analysis of adversarial attacks and defenses on intelligence in o-ran based cellular networks," in *Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '24, (New York, NY, USA), p. 237–247, ACM, 2024.

[15] S. Jerbi, C. Gyurik, S. C. Marshall, H. J. Briegel, and V. Dunjko, "Parametrized quantum policies for reinforcement learning," in *Proc. of 2021 Conference on Neural Information Processing Systems (NeurIPS 2021)*, 2021.

Van-Linh Nguyen (Senior Member of IEEE) is an assistant professor at the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan. His research interests include cybersecurity, wireless communications, and quantum machine learning.

Lan-Huong Nguyen is a postdoc fellow at the College of Artificial Intelligence, National Yang Ming Chiao Tung University, Taiwan. Her research interests include network optimization, traffic engineering, cybersecurity, and quantum AI.

Ren-Hung Hwang (Senior Member of IEEE) is the dean of the College of Artificial Intelligence, National Yang Ming Chiao Tung University, Taiwan. His research interests include deep learning, network security, wireless communications, Internet of Things, cloud and edge computing.

Berk Canberk (Senior Member, IEEE) is a Full Professor at Edinburgh Napier University, UK, leading research in AI-powered Digital Twins, IoT communication, and Smart Wireless Networks.

Trung Q. Duong (Fellow, IEEE) is a Canada Excellence Research Chair (CERC) at Memorial University, Canada. His current research interests include wireless communications, quantum machine learning optimization.