Explainable Quantum Learning for Robust Interference Classification in Next-Gen O-RAN

Yared Abera Ergu, *Graduate Student Member, IEEE*, Po-Ching Lin, *Member IEEE*, Ren-Hung Hwang, *Senior Member, IEEE*, Trung Q. Duong, *Fellow, IEEE*, Van-Linh Nguyen, *Senior Member, IEEE*

***Abstract*—As open radio access networks (O-RAN) progress into an intelligence-driven era, artificial intelligence (AI)-powered core functions and applications have recently attracted much attention. The AI-driven techniques enable near-real-time tasks such as signal interference classification for large-scale networks. Accordingly, accurate identification of interference types enables more effective successive interference cancellation, reducing performance degradation and strengthening reliability in O-RAN-based mobile networks. Since quantum computing remains energy-intensive and limited by current hardware capabilities, hybrid quantum–classical machine learning offers an efficient alternative. However, adversarial robustness and explainability remain critical challenges for such hybrid models, as emerging perturbation-based attacks can easily distort their decision boundaries. This work presents *Q-SHIELD*, a quantum-assisted defense framework that enhances robustness and explainability via trusted-manifold-guided quantum state transformations and fidelity-based boundary reinforcement. Q-SHIELD is evaluated under various poisoning attacks and adversarial attacks. Using a public dataset from real O-RAN testbeds, Q-SHIELD attains 98.2% accuracy in common cases and maintains 84.7% accuracy in worst-case conditions, even under strong adversarial attacks, e.g., quantum projected gradient descent. For explainability, Q-SHIELD emphasizes causal compactness, achieving a drift correlation of 0.988 and sufficiency/comprehensiveness scores of 0.871/0.007 on the top 20% features. These results outperform adversarial-training baselines by up to 35% in defense performance. The results show that Q-SHIELD can deliver robust quantum-assisted services for the intelligent O-RAN.**

***Index Terms*—Secure quantum machine learning, robust signal classification, adversarial attacks, quantum learning explainability**

I. Introduction

THE evolution toward intelligent, software-defined radio access networks has accelerated the adoption of artificial intelligence (AI) in open radio access networks (O-RAN) microservices, such as radio signal classification for interference-aware xApp services [1], [2]. Recent efforts have further explored the integration of hybrid quantum machine learning (QML) models within wireless communication networks, where network data are encoded into high-dimensional Hilbert spaces and processed via variational quantum circuits or quantum kernels [3], [4]. While this paradigm enables expressive decision functions under stringent latency constraints such as in O-RAN, it simultaneously introduces new vulnerabilities that are not present in purely classical learning. In particular, the interaction between classical feature extractors and quantum encoders creates a cross-layer interface where adversarial perturbations can be amplified through quantum state evolution, raising critical concerns for reliable deployment in interference classification tasks. As a result, this work focuses on hybrid quantum-classical interference classification in O-RAN and studies how such models behave under adversarial conditions that affect input representations and quantum state dynamics.

Two key technical challenges arise in this context. First, hybrid QML models remain difficult to interpret, as the combined effects of classical feature extraction, quantum encoding, and measurement collapse obscure the relationship between input perturbations and final decisions [5]–[7]. Second, the disaggregated and software-driven nature of O-RAN introduces new adversarial surfaces, where attackers can exploit exposed analytics interfaces or manipulate xApp deployment pipelines [8]–[14]. These threats manifest both at the system level (e.g., Y1-consumer abuse of radio access analytics [15]) and at the model level (e.g., supply-chain compromise of QML-enabled xApps [10]), highlighting the need for an efficient framework that jointly considers robustness and interpretability in hybrid quantum-classical pipelines. Despite recent advances, existing studies either focus on adversarial robustness in classical models without accounting for quantum encoding effects, or analyze QML vulnerabilities in isolated settings without considering deployment constraints such as near-RT RIC latency and E2-based data flows. This leaves a critical gap in understanding and securing the classical-quantum interface in practical O-RAN deployments. To address this gap, this paper proposes a hybrid quantum-classical interference classification framework and evaluates its robustness under adversarial conditions that explicitly target both classical inputs and quantum representations. Furthermore, the study incorporates fidelity-based constraints and interpretable circuit diagnostics, enabling traceability of decision behavior through

This work was supported by the National Science and Technology Council (NSTC) of Taiwan under Grants 114-2221-E-194-022-MY3, NSTC 114-2923-E-194 -001 -MY3, CHIST-ERA Project SHIELD, and in part by the Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

Y. A. Ergu, V.-L. Nguyen, and P.-C. Lin are with the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan (e-mail: {yared111p, nvlinh, pclin}@cs.ccu.edu.tw).

R.-H. Hwang is with Asia University, and National Yang Ming Chiao Tung University, Taiwan (e-mail: rhhwang@nycu.edu.tw).

T. Q. Duong is with Memorial University, Canada (e-mail: tduong@mun.ca).

Corresponding author: V. L. Nguyen.

Hilbert-space sensitivity analysis.

The remainder of this paper is structured as follows. Section II related studies. Section III outlines the system model. Section IV details the threat model assumption, and Section V presents the proposed defense method. Section VI delves into the details of the model performance evaluation. Finally, Section VII conclusions and points out possible future work.

II. Related Work

This section reviews prior work relevant to adversarial robustness and explainability in hybrid quantum-classical systems in O-RAN environments. We focus on three complementary dimensions: (i) system-level attack surfaces in O-RAN, (ii) adversarial robustness techniques in quantum and hybrid learning models, and (iii) explainability methods for QML.

A. Adversarial Vulnerabilities Across O-RAN Interfaces

The modular and AI-native architecture of O-RAN significantly broadens the adversarial attack surface across both data and control planes [16], [17]. As illustrated in Fig. 1, six key attack surfaces emerge across the O-RAN architecture under practical deployment constraints: ① OTA attacks: the UE–O-RU wireless link is susceptible to jamming and spoofing that introduce stealthy perturbations into raw I/Q signals and degrade PHY-layer fidelity [11]; ② O-FH attacks: adversaries may manipulate CSI or telemetry exchanged between the O-RU and O-DU [18]; ③ F1/E1 attacks: compromised intermediate data streams can propagate manipulated measurements toward higher-layer network functions [19]; ④ E2 attacks: performance metrics and spectrograms exposed to the near-RT RIC can be manipulated at inference time to mislead ML-driven xApps [20]; ⑤ O1 attacks: adversaries may exploit configuration and management paths to inject stealthy triggers or manipulate model updates at the non-RT RIC [2]; and ⑥ O2/CI-CD attacks: backend compromise may enable white-box model access, trojanized xApps, or orchestrator backdoors through software supply-chain attacks [21], [22].

Generally, adversarial vulnerabilities in O-RAN are often aligned with ingress ①, inference ④, and orchestration ⑥, where QML-powered xApps are most exposed. Recent studies reinforce these architectural concerns. For example, inference-time attacks via telemetry spoofing on the E2 interface are shown to degrade xApp decision integrity [11], [20]. Controlled experiments on O-DU and O-CU units validate how timestamp-aware perturbations induce CSI misalignment [17], while Rimedo Labs reports highlight the risk of spoofed KPMs [18]. Additionally, CI/CD compromise and poisoned AI updates are demonstrated as practical risks through supply-chain poisoning attacks in [2], [21]. Recent analyses by [22] further reveal the orchestration risks posed by misconfigured plugins and third-party packages, especially in open RAN marketplaces.

While prior works have analyzed classical xApp poisoning and OTA signal perturbations [11], [16], [20], inference-time threats against hybrid xApps, particularly those manipulating

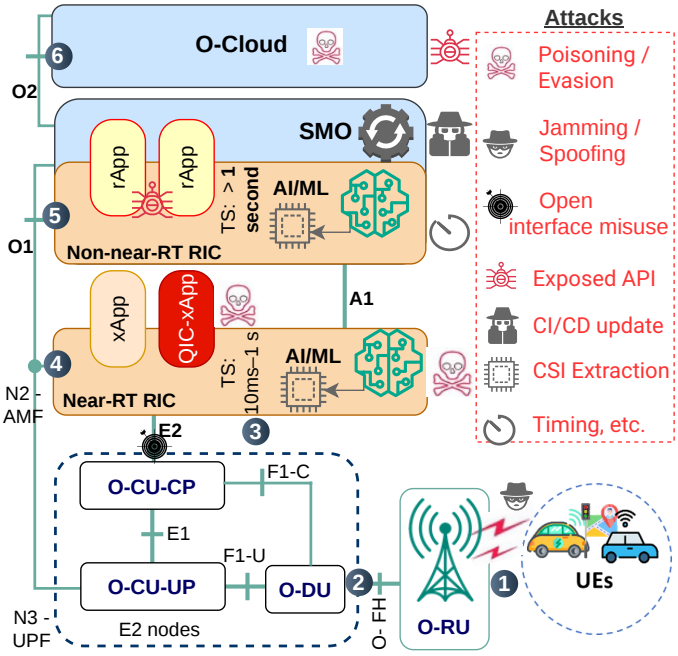

Fig. 1: Attack surfaces in O-RAN with step-wise service flow.

quantum encodings via telemetry ingress or trojaned deployment pipelines, remain largely unexplored. The threat model (Sec. IV) inherits key attacks (QC-FGSM, QC-PGD, QC-Poison) in [23] that directly map to surfaces ①, ④, and ⑥.

B. State-of-the-art Quantum Adversarial Threats Defenses

QML is progressing from theoretical foundations toward practical hybrid quantum-classical implementations, with fully quantum systems remaining out of reach due to the absence of scalable commercial quantum hardware [24]. At this early stage, ensuring robustness against adversarial threats has emerged as a critical research frontier [25]. As summarized in **Table I**, the strategies can be categorized into three paradigms: (i) stochastic encoding and quantum noise-based defenses, (ii) quantum-aware optimization and regularization techniques, and (iii) collaborative or transfer-based learning.

The first category of defenses mitigates adversarial gradients by introducing stochastic input transformations and noise handling. Techniques such as random basis encoding and quantum error correction reduce the success of gradient-based attacks, e.g., quantum fast gradient sign method (QC-FGSM), quantum projected gradient descent (QC-PGD), by injecting controlled randomness during quantum state preparation and also the common adversarial training technique [21], [30]. The second category of defenses focuses on architectural or training-time circuit adaptations aimed at improving robustness by reshaping the quantum model's optimization. Gradient-free optimization and barren plateau avoidance schemes have been explored to mitigate adversarial sensitivity and vanishing gradients issues [27], but their high computational overhead

TABLE I: State-of-the-art quantum adversarial defenses compared to Q-SHIELD

Category	Study	Threat	Defense Strategy	Methods	Timing	Perturb. Scaling	Limitations
Stochastic / Noise (Adversarial Training)	[21]	FGSM, PGD	Random Enco + QEC	Gradient obfuscation	Training	×	Requires shared keys[a]
	[26]	I-FGSM, L_2	Depolarization noise	Quantum diff. privacy	Inference	×	Accuracy-noise trade-off
Optimization / Reg.	[27]	Gradient-based	Grad-free optimize	Barren-plateau resistance	Training	×	High computation cost[b]
	[25]	Gate noise	Circuit regularization	Hilbert-space smoothing	Training	×	Not tested on hybrid
	[28]	Universal pert.	Metric alignment	Circuit-metric optim.	Training	×	Depth-sensitive[c]
Collaborative / Transfer	[12]	Universal pert., noise	Q-transfer learning	TQ-Learn	Both	×	Noisy-QPU instability
	[29]	Evasion attacks	Fed-adversarial train	Client aggregation	Training	×	Focused on detection only
	Ours	Poisoning, drift	QST + Q-DBR	Fidelity based filtering	Both	✓	Fidelity estimation cost

[a] Key sharing difficult in federated/hybrid settings. [b] Gradient-free optimization costly on large circuits. [c] Circuit depth tuning fragile to hardware noise.

and limited generalizability to hybrid pipelines present unseen challenges. Other techniques leverage Hilbert space regularization or circuit metric alignment to increase robustness [25], [28], but require fine-tuning of quantum circuit depth and hyperparameters, making them sensitive to hardware noise [31].

The last category of defenses incorporates collaborative training settings such as federated learning [32], [33] or transfer learning. In these frameworks, robustness is pursued through distributed aggregation or knowledge reuse across clients [29], detecting certain evasive attacks, e.g., QC-FGSM, quantum iterative gradient sign method (QC-IGSM), QC-PGD, and a hybrid of QC-IGSM and QC-PGD. These methods, while promising in decentralized environments, often neglect the interaction between input perturbations and quantum encoding fidelity in hybrid models. Meanwhile, approaches such as noise saturation [12], [34] introduce stochasticity during training to enhance generalization, but lack explicit mechanisms to address adversarial gradient flow at the quantum gate level.

Generally, hybrid models are vulnerable to training-time poisoning and gradient-based attacks. For example, the attacker can manipulate parameterized encodings and variational decision layers [35]. Existing defenses often freeze or randomize the encoder, yielding brittle margins and worst-case risk under random and uncontrolled poison. To address the limitations, we propose a layered defense that jointly constrains the encoder and the variational head via encoder-aware sanitization, contrastive boundary reinforcement, and adversarial training.

C. Explainability in Quantum Machine Learning

In QML, robustness alone is insufficient for deployment: models must also be explainable and trustworthy to support verification, debugging, and accountability. This is especially important because QML models can inherit fragility from high-dimensional Hilbert-space geometry (e.g., concentration-of-measure effects), where small perturbations may induce disproportionately large changes in the learned representation, making failures hard to diagnose without principled attribution mechanisms [36], [37]. Moreover, the immaturity of quantum software engineering (limited debuggers, static analysis, and standardized testing workflows) further complicates validation of complex quantum programs and can conceal subtle failure modes or malicious behaviors [38].

Existing explainability methods for QML primarily focus on interpreting variational quantum circuits and hybrid quantum-classical models through attribution, visualization, or sensitivity analysis. Gradient-based attribution identifies influential circuit parameters, while Shapley-value methods quantify qubit-level feature importance, as demonstrated by SVQXs [39]. Frameworks such as ExQUAL [6] and QRLaXAI [40] further extend these ideas to hybrid models by combining classical explainability techniques with quantum representations. Complementary visualization systems such as VIOLET [7] expose encoder and measurement behavior, whereas tensor-network approaches provide interpretable low-rank representations of quantum operators [41]. Collectively, these studies have established a foundation for explainable QML, although they differ substantially in their interpretation granularity and target applications. Within 6G and O-RAN research, explainability is increasingly recognized as an essential requirement for transparent and accountable AI operation [42], [43]. However, as summarized in Table II, existing QML explainability methods primarily emphasize attribution or visualization of isolated quantum components. Key research gap is that they do not explicitly relate quantum-state perturbations to decision robustness in hybrid inference pipelines.

TABLE II: Summary of SOTA Explainable QML

Scope	Ref.	Method	Model	A*	V	G
Full QML	[5]	Q-LIME	Theoretical-QNN	×	✓	×
	[7]	Vis. (multi-view)	Vis. system	×	✓	×
	[39]	Shapley (VQC)	SVQXs	✓	×	✓
Hybrid	[40]	LIME+SHAP–QRLaXAI	Vision (QRL)	✓	✓	×
	[41]	LIME	QSVC	✓	×	×
	Ours	Fidelity-Weighted Attr.	HQML	✓	✓	×

* A: Attribution; V: Visualization; G: Gate-level granularity explanation

Generally, building upon existing attribution methods such as SHAP and LIME [44], this work addresses the aforementioned gap by introducing a fidelity-weighted attribution scheme that scales feature importance according to the quantum-state fidelity between clean and perturbed encodings. This establishes an explicit connection between classical feature attribution and quantum-state sensitivity, providing more informative explanations of hybrid model behavior under adversarial perturbations.

D. Contributions

Unlike randomized encoding obfuscation [21], this work jointly hardens the quantum encoder and the variational decision head. The approach maintains a baseline snapshot of the quantum-fusion parameters and clean-state statistics so that drift can be measured and reinforced resets triggered whenever a poisoning threshold is crossed. Key contributions are summarized as follows.

- We propose a novel defense technique, namely Q-SHIELD, which transforms inputs via quantum state transformation (QST), blends repairs when states drift from the class manifold, and keeps quantum feature centers separated via decision-boundary reinforcement. This aims to tighten margins against evasion attacks and quantum poisoning attacks.
- We propose a min-max adversarial training framework that defends against QC-FGSM and QC-PGD attacks by jointly perturbing inputs and quantum encodings during the inner maximization step, while injecting stochastic circuit noise to enhance stability. We conduct a comprehensive evaluation across multiple robustness metrics, demonstrating that *Q-SHIELD* preserves near-clean accuracy ($0.982 \rightarrow 0.974$), significantly improves resistance to QC-FGSM ($0.558 \rightarrow 0.847$), QC-PGD ($0.160 \rightarrow 0.736$), and outperforms standard adversarial training baselines by 9–35% under worst-case conditions.
- We introduce a fidelity-weighted explainability mechanism for *Q-SHIELD*, enabling layer-wise attribution tracking across classical and quantum paths. By scaling Shapley contributions with quantum-state fidelity between clean and perturbed encodings, the method projects attribution into Hilbert-space drift, revealing how quantum encoders sustain markedly lower attribution divergence under adversarial perturbations compared to classical components. This quantifies an explainable quantum advantage in interference-based classification by prioritizing features with the highest predictive contribution. To support reproducibility, our reference implementation is publicly available on GitHub[1].

Note that the objective of this work is not to establish formal quantum advantage in terms of complexity, hardware efficiency, or throughput. While existing studies and industrial efforts have reported promising QML capabilities at small scales [45], [46], further validation is still required before large-scale commercial deployment. Nevertheless, this does not diminish the importance of investigating the security and reliability challenges associated with hybrid QML architectures, particularly as such models can already be deployed on GPU-based infrastructures and are increasingly viewed as a potential paradigm for future AI systems. Accordingly, we treat hybrid QML as an emerging option for O-RAN and focus on analyzing vulnerabilities at the classical-quantum interface. Furthermore, the explainability component is not intended as a standalone XAI contribution; rather, it is tightly integrated with the proposed defense mechanism to interpret robustness behavior and characterize adversarially induced decision drift within the hybrid quantum-classical pipeline, thereby improving the trustworthiness of hybrid QML systems under attack.

[1]https://github.com/Jaredabera/Q-Poisoning-Adversarial-Attack

III. System Model and Problem Formulation

We consider an intelligent interference classification xApp deployed in the near-RT RIC of O-RAN. The xApp integrates a classical convolutional neural network (ResNet-18) backbone for feature extraction with a hybrid variational quantum classifier (VQC) for decision-making. Its objective is to distinguish the signal of interest (SOI) from continuous-wave interference (CWI) for real-time jammer detection The end-to-end pipeline comprises (1) UE→RAN sensing: the UE transmits uplink radio signals that are measured at the RAN (RU/DU/CU), producing KPMs and IQ samples converted into spectrograms; (2) convolutional neural network (CNN) feature extraction: a convolutional backbone processes spectrograms to extract feature vectors; (3) data-processing microservice: features are normalized and mapped into a latent representation $\mathbf{z}$, with internal messaging to RIC services and optional persistence in the database; (4) Q-Encoding: a quantum encoder $\psi(\boldsymbol{\theta}_q)$ embeds $k\mathbf{z}$ into a quantum state for a VQC, $V(\boldsymbol{\theta}_q)$, whose readout yields the classification decision; (5) O-RAN E2 control loop: closed-loop policy enforcement and actuation toward the RAN; and (6) E2 agent: request-reply messaging between the near-RT RIC and the RAN.

To emulate adversarial conditions, a jammer injects CWI over the same carrier frequency as the SOI during over-the-air (OTA) transmission. This interference alters the IQ samples at the measurement stage (Step 1), propagating through the feature extraction and classification pipeline. The attack surface includes both physical-layer perturbations (via CWI) and logical-layer threats such as input evasion and poisoning, which are modeled within the xApp to evaluate robustness against encoder drift and adversarial manipulation. The received complex baseband signal over a time t is expressed as

$$r(t) = s(t) + \sum_{k=1}^{K} j_k(t) + n(t), \qquad n(t) \sim \mathcal{CN}(0, \sigma^2), \quad (1)$$

where $s(t)$ denotes the SOI, $j_k(t)$ is the k^{th} jammer, and $n(t)$ is complex Gaussian noise with variance σ^2. Each jammed signal instance is expressed as

$$j_k(t) = A_k e^{j(2\pi f_k t + \phi_k + \psi_k(t))}, \quad (2)$$

with amplitude A_k, carrier offset f_k, initial phase ϕ_k, and slow phase noise $\psi_k(t)$. The raw IQ samples are transformed into a time-frequency representation using short-time Fourier transform (STFT) [2]

$$\mathbf{S}(f, \tau) = \left| \int r(t) w(t - \tau) e^{-j2\pi f t} dt \right|^2 \in \mathbb{R}^{H \times W}, \quad (3)$$

where $w(\cdot)$ is an analysis window, H is the number of frequency bins, and W is the number of time frames. The

TABLE III: Table of Notations

Notation	Description	Notation	Description
	III System Mode and Problem Formulation		
$r(t)$	Received baseband signal	$j_k(t)$	k-th jammer
$s(t)$	Signal-of-interest (SOI)	$n(t)$	Complex AWGN noise
$\mathbf{S}(f,\tau)$	STFT spectrogram	$\mathcal{F}(\rho,\sigma)$	Quantum fidelity
H, W	Frequency/time bins	$\mathbf{z}$	CNN latent feature
U_{enc}	Quantum feature encoder	q	Number of qubits
$y_i \in \{0,1\}$	Class label SOI or CWI	$\mathbb{E}[\cdot]$	Expectation operator
$\rho(\mathbf{S})$	Encoded quantum state	$V(\theta_q)$	Variational circuit
$\mathcal{M}(\cdot)$	Quantum measurement	$f_\theta(\cdot)$	Hybrid classifier output
	IV Threat Model and Attack Assumptions		
$\mathcal{C}_{\text{poison}}$	Complexity of poisoning attack	$\mathcal{C}_{\text{evasion}}$	Complexity of evasion attack
p	Number of variational parameters	m	Number of optimization iterations
d	Circuit depth	$\epsilon\lvert\mathcal{D}_{\text{train}}\rvert$	Fraction of poisoned samples
$\Phi(\mathbf{x})$	Classical-quantum encoder	$\sigma(\mathbf{x})$	Adversarial encoded state
$d_{\sigma\rho_i}$	Hilbert-space distance $\lVert\sigma-\rho_i\rVert_F$	δ_{q}	Quantum-side perturbation
$\mathcal{L}(\cdot,\cdot)$	Task loss (e.g., CE)	$\Delta=\{\delta : \lVert\delta\rVert_p \le \epsilon\}$	Bounded input perturbation set
ϵ	Perturbation radius	Ξ	Set of quantum noise channels
	V Q-SHIELD: Proposed Explainable Quantum xApp for Robust Interference Classification		
$\mathcal{D}^{(c)}$	Class-c quantum anchors	$p^{(c)}$	Class prototype
$\bar{\rho}^{(c)}$	Class centroid state	δ_i	Drift to centroid
$\hat{\rho}$	The normalized quantum state	$\hat{p}^{(c)}$	The stored class anchor
$\mathcal{L}_{\text{QST}}$	QST loss	$\mathcal{L}_{\text{DBR}}$	Q-DBR loss
$\lambda_{\text{rec}}, \lambda_{\text{proto}}, \lambda_{\text{drift}}$	QST weights	$\lambda_{\text{intra}}, \lambda_{\text{inter}}$	Q-DBR weights
τ_i	Drift-gated update weight	f_{Robust}	Final defended model
	V-D Enhancing Q-SHIELD's Explainability		
D_{h}	Global Hilbert drift $1-\mathcal{F}$	F_i	Local fidelity (tile i)
s_i	Baseline attribution	$s_i^\star$	Fidelity-drift reweighted attribution
z	Fused feature vector	μ_y	Class centroid (feature space)
$D_{\cos}$	Cosine drift $1-\langle z,\mu_y\rangle$	d_i	Drift of feature i

resulting spectrogram $\mathbf{S}(f,\tau)$ is normalized and optionally augmented with cyclostationary features for modulated interference separation.

The normalized spectrogram $\mathbf{S}$ is first embedded into a latent feature vector $\mathbf{z}$ via a lightweight CNN (ResNet-18 backbone with projection layer). The QAE then maps $\mathbf{z}$ into a q-qubit vector

$$|\psi(\mathbf{z})\rangle = U_{\text{enc}}(\mathbf{z})\,|0\rangle^{\otimes q}, \tag{4}$$

where U_{enc} is a data-reuploading feature map. The encoded state is evolved through a parameterized variational circuit

$$\big|\Psi_{\theta_q}(\mathbf{z})\big\rangle = V(\theta_q)\,|\psi(\mathbf{z})\rangle, \tag{5}$$

with θ_q denoting trainable quantum parameters. Measurements $\mathcal{M}(\cdot)$ yield quantum features, which are fused with a classical projection head $\mathcal{C}_{\theta_c}(\cdot)$

$$f_\theta(\mathbf{S}) = \mathcal{C}_{\theta_c}\big(\mathcal{M}(\big|\Psi_{\theta_q}(\mathbf{z})\big\rangle)\big). \tag{6}$$

The final decision function $f_\theta(\mathbf{S}) \in \{0,1\}$ corresponds to SOI (0) vs. CWI (1). The hybrid workflow leverages quantum superposition and entanglement for expressive feature extraction, while classical layers ensure scalable training and optimization. This structure circumvents the challenges of fully quantum pipelines in near-term devices, enabling deployable interference detection in O-RAN edge controllers (e.g., real-time processing close to the source) without incurring the thousands of error-corrected qubits required by fully quantum algorithms in the current noisy intermediate-scale quantum (NISQ) era [47].

A. RF Classifier Problem Formulation

Let $\mathcal{D}_{\text{train}} = \{(\mathbf{S}_i, y_i)\}_{i=1}^N$ and $\mathcal{D}_{\text{test}}$ denote spectrogram-label pairs with $y_i \in \{0,1\}$. The learnable $\theta = (\theta_c, \theta_q)$ parameters to minimize worst-case risk over admissible classical perturbations and quantum noise

$$\min_\theta \; \mathbb{E}_{(\mathbf{S},y)\sim\mathcal{D}_{\text{train}}}\left[\max_{\delta\in\Delta,\,\xi\in\Xi} \mathcal{L}\big(f_\theta(\mathbf{S}+\delta), y\big)\right], \tag{7}$$

where $\Delta = \{\delta : \lVert\delta\rVert_p \le \epsilon\}$ models bounded state perturbations [27], [48] (e.g., QC-FGSM/PGD in the classical interface [35]), and Ξ parameterizes stochastic quantum channels applied during the forward pass. To constrain quantum-state drift, we impose a fidelity guard

$$\mathcal{F}\big(\rho(\mathbf{S}), \rho(\mathbf{S}+\delta)\big) \ge \tau, \tag{8}$$

$$\mathcal{F}(\rho,\sigma) = \left(\mathrm{Tr}\sqrt{\sqrt{\rho}\,\sigma\,\sqrt{\rho}}\right)^2, \tag{9}$$

with $\rho(\mathbf{S})$ the encoded state and $\tau \in [0,1]$ a minimum similarity threshold [49]. This objective is designed to integrate with the defensive stack: encoder-aware sanitization constrains Δ, contrastive reinforcement regularizes quantum embeddings, and quantum-native adversarial training realizes the inner robustness maximization with realistic channel noise. The notations used in this paper are presented in **Table III**.

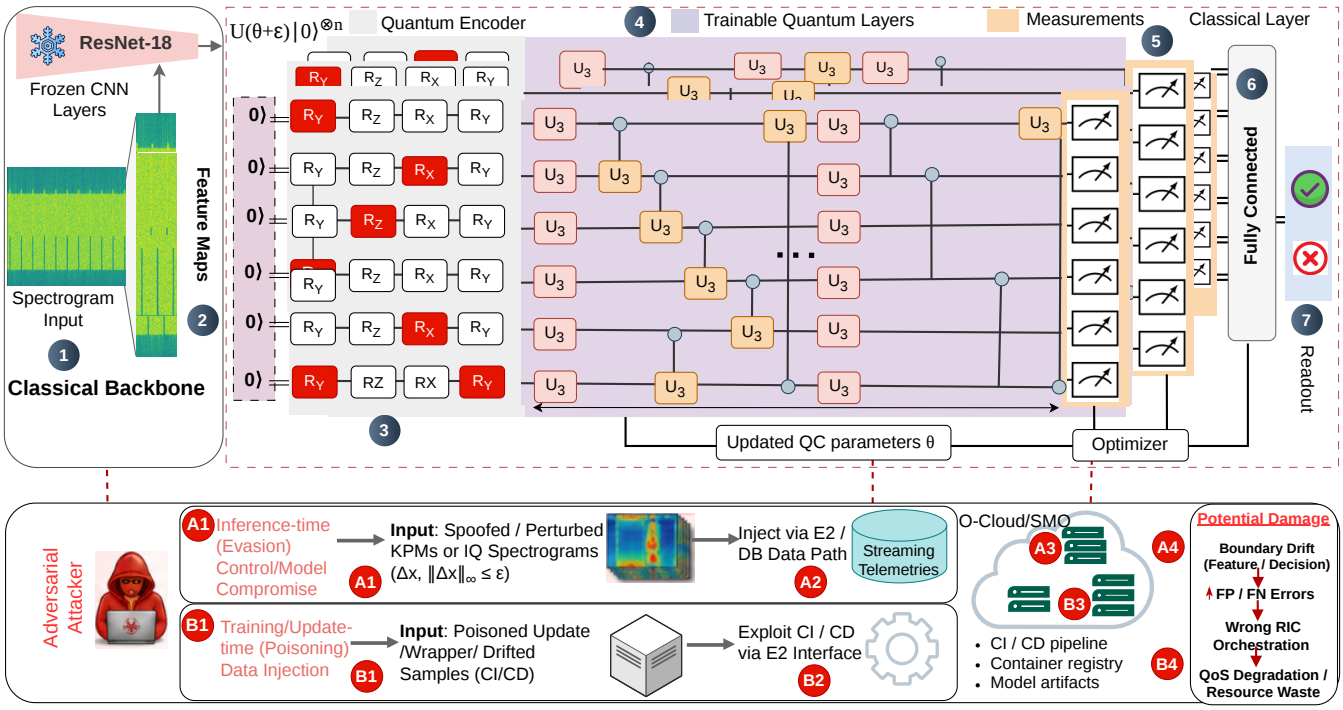

Fig. 2: Attack surface of the hybrid ResNet-18+VQC interference-classification xApp in near-RT RIC. Steps ①–⑦ depict the inference pipeline: ① an input I/Q spectrogram is ingested by the classical backbone; ②–③ the frozen ResNet-18 extracts feature maps and the quantum encoder maps the projected feature vector into a 6-qubit state $U(\boldsymbol{\theta}+\epsilon)\,|0\rangle^{\otimes n}$ via parameterized single-qubit rotations; ④ trainable variational layers (entangling + U_3 blocks) transform the encoded state; ⑤ measurement produces an expectation-value embedding; ⑥–⑦ a classical fully-connected head performs readout and outputs the final decision. The lower panel summarizes attacker ingress: (A) inference-time evasion via spoofed/perturbed KPMs or I/Q spectrograms (A1) injected through the E2/DB data path (A2); (B) training/update-time poisoning via drifted samples or malicious wrappers (B1) delivered through the CI/CD interface (B2); (C) the O-Cloud/SMO supply-chain surface (CI/CD pipeline, container registry, model artifacts) enabling white-box exposure; and (D) the resulting impact as decision-boundary drift, increased FP/FN errors, and incorrect RIC orchestration decisions leading to QoS degradation/resource waste.

IV. Threat Model and Attack Assumptions

In the considered O-RAN deployment, an xApp performs spectrogram-based interference classification using a hybrid quantum-classical model, where classical encoders and VQCs jointly process baseband IQ data from O-DUs via the E2 interface. The adversary aims to compromise this pipeline either during training or through post-deployment updates The primary attack vectors include supply-chain infiltration of unverified xApp containers, malicious telemetry replay via O1/E2 interfaces, and compromised quantum-cloud training environments [50], [51]. In another scenario, an attacker can register an unintended RMR (RIC Message Router) message type during xApp registration, potentially disrupting other service components [52]. This vulnerability was identified in `appmgr` within the O-RAN Near-RT RIC I-Release.

Building on the modular threat taxonomy outlined in Fig. 1 and [23], we now present the core inference pipeline of our QIC-xApp in the near-RT RIC. The system integrates classical and quantum modules to support real-time interference detection across O-RAN interfaces vulnerable to adversarial manipulation. Fig. 2 depicts the step-wise inference workflow and highlights attacker entry points aligned with the surface model in Fig. 1. This architecture serves as the basis for our threat formulation and motivates the layered defense strategy illustrated in Fig. 3 below.

Fig. 2 expands the QIC-xApp block detailing the hybrid classifier architecture and its adversarial hooks. Spectrogram features extracted by the frozen ResNet-18 backbone are encoded into quantum states via parameterized rotations (R_y, R_z, R_x), processed through variational layers of U_3 gates, and measured to produce quantum features for a classical projection head. The diagram also highlights attack surfaces: input-level perturbations affecting spectrograms, encoding-level poisoning through parameter shifts $U_3(\boldsymbol{\theta}+\delta)$, and consequential resultant effect on the optimizer during parameter updates. This connection illustrates how physical-layer interference (SOI vs. CWI) and cloud-side adversarial actions impact the decision pipeline.

Following the quantum-cloud adversary model, where adversaries will use quantum resources to support their strategies [53], the attacker could also operate within a third-party quantum infrastructure where the hybrid model is trained, possessing partial visibility into the victim's preprocessing and data encoding. Specifically, the adversary can access the labeled training dataset $\mathcal{D}_{\text{train}}$ and the classical-to-quantum encoder $\Phi(\mathbf{x})$ that maps spectrogram features into quantum states [10]. Here, $\mathbf{x}$ denotes the latent feature vector obtained

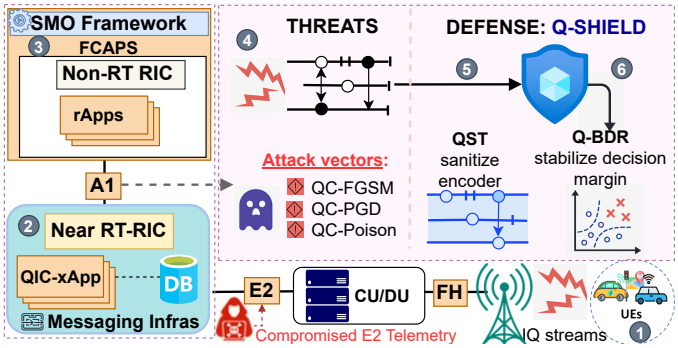

Fig. 3: Q-SHIELD defense architecture for hybrid ResNet-VQC interference classification in the near-RT RIC. The system augments the hybrid inference pipeline (cf. Fig.2) and the attack surfaces outlined in Fig. 1 by introducing two key defense stages: (i) *QST* at the classical-to-quantum encoding interface, mitigating encoding drift and quantum decoherence, and (ii) *Q-DBR* post-measurement to stabilize classification against adversarial decision shift. The threat vectors(QC-FGSM, QC-PGD), and QC-Poison(target telemetry inputs, circuit encoding, and CI/CD pathways) respectively, and are counteracted within this modular, xApp-based deployment.

after STFT-based spectrogram generation and embedding (**see Eq. 3**), not raw IQ samples. Within this constrained view, the adversary performs an encoding-level poisoning attack on a fraction $\epsilon|\mathcal{D}_{\text{train}}|$ of samples by recomputing their quantum encodings and constructing density matrices

$$\sigma(\mathbf{x}) = \Phi(\mathbf{x})|0\rangle\langle 0|\Phi(\mathbf{x})^{\dagger}, \tag{10}$$

and then computing quantum distances $d_{\sigma\rho_i} = \|\sigma(\mathbf{x}) - \rho_i\|$ for label reassignment (quantum gate configuration-structure level). Since each density matrix has dimension $2^q \times 2^q$, the cost of computing pairwise distances scales as $(2^q)^2 = 4^q$, resulting in $\mathcal{C}_{\text{poison}} = O\big(\epsilon|\mathcal{D}_{\text{train}}| \cdot 4^q\big)$. For instance, $q = 6$ requires $4^6 = 4096$ operations per sample, which is feasible under current O-RAN constraints (small q), whereas larger q values quickly exceed practical limits, making large-scale corruption computationally costly.

The same adversary can perform test-time evasion by crafting fidelity-aware perturbations δ_{q} that alter encoded states $\rho(\mathbf{x} + \delta_{\text{q}})$ toward adversarial regions of the variational landscape. Using the parameter-shift rule

$$\nabla_\theta f(\theta) = \frac{f(\theta + \pi/2) - f(\theta - \pi/2)}{2}, \tag{11}$$

the complexity becomes $\mathcal{C}_{\text{evasion}} = O(2pm \cdot d \cdot q^2)$, where p is the number of variational parameters, m the optimization steps, and d the circuit depth. This polynomial scaling makes evasion moderately feasible for small-scale hybrid models but increasingly expensive for deeper circuits and larger parameter spaces given the current NISQ-era hardware limitations.

V. Q-SHIELD: Proposed Explainable Quantum xApp for Robust Interference Classification

This section overviews the detail of Q-SHIELD, a layered defense that augments a hybrid quantum-classical classifier with two modules: quantum state transformation and decision-boundary reinforcement. Adversarial attacks such as QC-FGSM and QC-PGD exploit weak or flat regions of the decision boundary, while QC-Poison [35] targets the autoencoder level, manipulating latent representations and corrupting fusion weights. Q-SHIELD mitigates these threats by applying QST at the quantum and fusion layers to reduce decoherence and bias, and decision-boundary reinforcement at the classifier level to detect and correct margin shifts using parameter-shift diagnostics (detect shifts in model behavior). **Fig. 3** illustrates attack vectors at the autoencoder and decision-boundary layers and shows where Q-SHIELD modules intervene to restore quantum state integrity and stabilize hybrid weights.

Building on the hybrid inference pipeline in **Eq. 4**, Q-SHIELD augments the same ResNet→VQC→head workflow with two tightly coupled defense modules: quantum state transformation (QST) at the classical-quantum interface and quantum decision-boundary reinforcement (Q-DBR) at the hybrid decision layer. QST reduces vulnerability to encoder drift and hybrid perturbations by stabilizing the encoded state (or measurement embedding) against small input shifts, while Q-DBR enforces a robustness margin so that adversarial perturbations require larger drift to flip the decision.

A. Module I: Quantum state transformation

From a trusted seed set, we form class-conditional anchors $\mathcal{D}^{(c)} = \{\rho_j^{(c)}\}$ and maintain two clean descriptors per class: the normalized prototype $p^{(c)}$ (running, momentum-averaged anchor) and the centroid $\bar{\rho}^{(c)}$ (mean quantum direction). For each encoded sample $\rho(\mathbf{S})$, minimize the following objective

$$\mathcal{L}_{\text{QST}} = \lambda_{\text{rec}}\|x - x_{\text{ae}}\|_1 + \lambda_{\text{proto}}\big(1 - \langle\hat{\rho}, \hat{p}^{(c)}\rangle\big) + \lambda_{\text{drift}}\big(1 - \langle\hat{\rho}, \hat{\bar{\rho}}^{(c)}\rangle\big). \tag{12}$$

This combines (i) reconstruction fidelity, (ii) prototype alignment (cosine similarity of the normalized quantum state $\hat{\rho}$ to the stored class anchor $\hat{p}^{(c)}$), and (iii) a drift penalty relative to the centroid $\hat{\bar{\rho}}^{(c)}$. Points whose peak fidelity $\mathcal{F}_{\max}^{(c)} = \max_{\rho_j \in \mathcal{R}^{(c)}} |\langle\rho_j|\rho\rangle|^2$ is greater than or equal to the threshold τ are accepted; otherwise, QST blends the auto-encoded state toward the nearest anchor using a drift-gated weight $\tau_i = \sigma(\tau + \beta\delta_i)$, where β is the drift-scale hyperparameter controlling sensitivity to the measured drift δ_i. The function $\sigma(\cdot)$ maps the gate argument into $[0, 1]$ for blending. The term $\beta\delta_i$ adjusts the gate based on the drift of the current sample from its class anchor: a larger drift increases the gate argument, making $\sigma(\tau + \beta\delta_i)$ closer to 1, resulting in stronger blending toward the nearest anchor. Conversely, if the drift is small, the gate remains near $\sigma(\tau)$, preserving the original auto-encoded state. Thus, τ_i acts as a dynamic weight in the convex combination

$$\rho' = (1 - \tau_i)\,\rho_{\text{ae}} + \tau_i\,\rho_{\text{anchor}}, \tag{13}$$

where ρ_{ae} is the auto-encoded state and ρ_{anchor} is the nearest class anchor. This adaptive correction improves robustness by reducing distortions caused by adversarial attacks.

B. Module II: Quantum decision-boundary reinforcement

To harden the variational classifier, we enforce large quantum margins with

$$\begin{aligned}\mathcal{L}_{\mathrm{DBR}} = \lambda_{\mathrm{intra}}\mathbb{E}_{(i,j)\in\mathcal{S}}[-\log\mathcal{F}(\rho_i,\rho_j)] \\ +\lambda_{\mathrm{inter}}\mathbb{E}_{(i,k)\in\mathcal{D}}[\log\mathcal{F}(\rho_i,\rho_k)]\,,\end{aligned} \tag{14}$$

where $\mathcal{S}$ and $\mathcal{D}$ denote same-class and different-class pairs, respectively, and $\mathcal{F}(\cdot,\cdot)$ is quantum fidelity. The first term tightens intra-class similarity, while the second penalizes cross-class overlap. This step is named quantum decision-boundary reinforcement (Q-DBR). Q-SHIELD maintains a baseline snapshot that (i) stores the quantum-fusion parameters defining the anchored circuit and (ii) records the clean-state feature mean along with drift statistics captured immediately after training. At the end of each epoch, this snapshot is updated by saving fusion weights, class prototypes, and drift estimates. During evaluation, cosine drift is monitored, and whenever it exceeds the learned threshold, the circuit is redirected toward the stored anchor before scoring. This layered reset mechanism enables the model to recover dynamically when a poisoning attempt perturbs circuit configurations.

Algorithm 1 couples adversarial training filtering with Q-DBR reinforcement. Lines 1–4 build per-class quantum anchor manifolds and centroids from $\mathcal{D}_{\text{seed}}$. Lines 5–13 score each encoded sample by peak fidelity and centroid drift; points above the fidelity gate are kept, while others are drift-weighted blends toward their nearest anchor before entering $\mathcal{D}_{\text{clean}}$. Training on $\mathcal{D}_{\text{clean}}$ with the fidelity-aware and Q-DBR loss yields the final hybrid robust model.

For quantum adversarial training (QAT), we defend against quantum-native test-time adversaries by solving a min-max optimization that perturbs inputs and injects quantum noise channels Ξ (e.g., depolarizing, amplitude-damping) during training. The inner maximization is implemented via QC-PGD using parameter-shift gradients under near-real-time constraints, as shown in **Eq. 7**. We compare our approach with existing QAT baselines. At each update, QC-FGSM/PGD perturbations are applied to spectrogram inputs and propagated through the variational circuit with stochastic quantum noise to ensure stable convergence. The resulting objective

$$\min_{\theta}\max_{\|\delta\|\le\epsilon}\mathcal{L}\big(f_\theta(x+\delta),y\big) \tag{15}$$

encourages the model to learn quantum features that remain predictive under bounded RF manipulations.

Algorithm 1 Multi-layer defense algorithm in Q-SHIELD

Require: Hybrid model f_θ, Training set $\mathcal{D}_{\text{train}} = \{(x_i,y_i)\}_{i=1}^{N}$, seed subset $\mathcal{D}_{\text{seed}} \subset \mathcal{D}_{\text{train}}$, encoding circuit U_{encode}, $\mathcal{F}$ threshold τ
Ensure: Filtered/repairable dataset $\mathcal{D}_{\text{clean}}$

1: **for** each class c **do**
2: $\quad\mathcal{M}_{\text{ref}}^{(c)} \leftarrow \{U_{\text{encode}}(x_j) \mid (x_j,y_j=c)\in\mathcal{D}_{\text{seed}}\}$
3: $\quad\big|\bar{\psi}^{(c)}\big\rangle \leftarrow \frac{1}{|\mathcal{M}_{\text{ref}}^{(c)}|}\sum_{|\theta\rangle\in\mathcal{M}_{\text{ref}}^{(c)}}\frac{|\theta\rangle}{\||\theta\rangle\|}$ ▷ class centroid
4: **end for**
5: $\mathcal{D}_{\text{clean}} \leftarrow \emptyset$
6: **for** each $(x_i,y_i)\in\mathcal{D}_{\text{train}}$ **do**
7: $\quad|\psi_i\rangle \leftarrow U_{\text{encode}}(x_i)$
8: $\quad\mathcal{F}_{\max}(x_i) \leftarrow \max_{|\theta\rangle\in\mathcal{M}_{\text{ref}}^{(y_i)}}|\langle\theta|\psi_i\rangle|^2$ ▷ $\mathcal{F}$ to seed
9: $\quad\delta_i \leftarrow 1-|\langle\bar{\psi}^{(y_i)}|\psi_i\rangle|$ ▷ cosine drift
10: $\quad\tau_i \leftarrow \sigma(\tau+\beta\delta_i)$ ▷ learned gating on reconst
11: $\quad$**if** $\mathcal{F}_{\max}(x_i)\ge\tau$ **then**
12: $\quad\quad\mathcal{D}_{\text{clean}} \leftarrow \mathcal{D}_{\text{clean}}\cup\{(x_i,y_i)\}$ ▷ retain clean sample
13: $\quad$**else**
14: $\quad\quad\hat{\psi}_i \leftarrow \arg\max_{|\theta\rangle\in\mathcal{M}_{\text{ref}}^{(y_i)}}|\langle\theta|\psi_i\rangle|^2$
15: $\quad\quad$Replace encoded state: $|\psi_i\rangle \leftarrow (1-\tau_i)|\psi_i\rangle+\tau_i\hat{\psi}_i$
16: $\quad\quad$Add repaired pair (x_i,y_i) to $\mathcal{D}_{\text{clean}}$
17: $\quad$**end if**
18: **end for**
19: **return** Train f_θ on $\mathcal{D}_{\text{clean}} \rightarrow f_{Robust}$

C. Explainability for signal classifier in Q-SHIELD

Spectrograms are high-dimensional time-frequency objects with structured correlations that can challenge conventional feature extractors. Quantum feature maps and variational circuits offer an alternative representation by embedding such data into Hilbert space, where distances and overlaps between encoded states can capture discriminative structure [3]. Recent hybrid quantum-classical models have reported accuracy and generalization gains on selected radio classification tasks, while remaining compatible with modular O-RAN deployment through xApps on the near-RT RIC [35], [54].

The opacity of variational quantum models motivates explainability methods adapted from classical XAI and extended to quantum settings. Current approaches include circuit and feature-level attribution and visualization for VQCs and hybrid pipelines [5]–[7], [39], [40], [55]. In this work, the Q-SHIELD training loop, quantum state transformation (QST) with class prototypes and quantum decision boundary reinforcement (Q-DBR) with margin regularization, are leveraged to define and visualize Hilbert-space drift per instance, and to couple this drift with fidelity-weighted attribution for hybrid inference. Furthermore, explainability supports adversarial robustness by exposing circuit-level vulnerabilities and anomalous decision paths, ensuring secure deployment in mission-critical O-RAN environments. By combining quantum computational advantages with interpretability frameworks, this work advances the development of robust radio interference services.

D. Enhancing Q-SHIELD's explainability

The Q-SHIELD framework integrates quantum state transformation with decision boundary reinforcement to maintain class prototypes and track feature drift in hybrid quantum-classical xApps. During training, drift is computed as one minus the cosine similarity between normalized fused quantum features and their class prototypes, while momentum updates preserve reference anchors for post-hoc correction. To extend this mechanism toward interpretability, the present

Algorithm 2 Drift mapping and attribution algorithm

Require: Input $\mathbf{x}$, tiles $\{i\}_1^d$, encoder U_{enc}, shots S, baseline $\{s_i\}$, class center μ_y, gains γ, λ

1: $\rho_0 \leftarrow U_{\text{enc}}(\mathbf{x})|0\rangle\langle 0|U_{\text{enc}}^\dagger(\mathbf{x})$
2: $z \leftarrow \mathsf{FuseFeatures}(\rho_0)$; $D_{\cos} \leftarrow 1 - \langle z, \mu_y \rangle$
3: **for** $i = 1$ to d **do**
4: $\quad \mathbf{x}^{(i)} \leftarrow \mathsf{PerturbTile}(\mathbf{x}, i)$
5: $\quad \rho_i \leftarrow U_{\text{enc}}(\mathbf{x}^{(i)})|0\rangle\langle 0|U_{\text{enc}}^\dagger(\mathbf{x}^{(i)})$
6: $\quad \mathcal{F}_i \leftarrow \mathsf{FidelityEstimate}(\rho_0, \rho_i; S)$
7: $\quad \mathsf{d}_i \leftarrow 1 - \mathcal{F}_i$ as in (**Eq. 9**);
8: $\quad s_i^\star \leftarrow s_i(1 - \mathcal{F}_i)(1 + \gamma D_{\cos})(1 - \lambda\langle z, \mu_y \rangle)$
9: **end for**
10: Normalize $\{s_i^\star\}$;
11: **return** A (drift map), $S^\star$ (attribution)

work reformulates drift in the Hilbert space of quantum states and links it with fidelity-weighted attribution for explainable inference within O-RAN controllers. Hilbert drift is defined for a spectrogram input $\mathbf{x}$ encoded into a q-qubit density matrix $\rho(\mathbf{x})$ with class anchor $\sigma^{(c)}$

$$\mathsf{D}_{\text{h}}(\mathbf{x}; c) = 1 - \mathcal{F}\big(\rho(\mathbf{x}), \sigma^{(c)}\big), \tag{16}$$

where $\mathcal{F}(\rho, \rho(\mathbf{x}+\delta_i))$ denotes the quantum fidelity between the normal and perturbed states as in **Eq. 9**. Localized perturbations δ_i on spectrogram tiles produce $\rho(\mathbf{x}+\delta_i)$ and fidelities $F_i = \mathcal{F}(\rho(\mathbf{x}), \rho(\mathbf{x}+\delta_i))$, yielding drift values $\mathsf{d}_i = 1 - F_i$ that generate per-instance heatmaps visualizing Hilbert-space sensitivity to adversarial perturbations.

Let s_i denote baseline attributions derived from SHAP/LIME-style methods adapted to QML [39]. The Shapley value for feature i is

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \big[f(S \cup \{i\}) - f(S)\big], \tag{17}$$

with N the feature set and $f(\cdot)$ the restricted model. We define the fidelity-weighted attribution

$$s_i' = s_i(1 - F_i), \qquad 0 \le 1 - F_i \le 1, \tag{18}$$

which preserves the sign of s_i and bounds its magnitude by local quantum sensitivity. To incorporate class geometry from Q-DBR, let z be the normalized fused quantum feature and μ_y the class centroid; a drift-aware variant is then

$$s_i^\star = s_i(1 - F_i)\big(1 + \gamma(1 - \langle z, \mu_y \rangle)\big)\big(1 - \lambda\langle z, \mu_y \rangle\big), \tag{19}$$

where $\gamma, \lambda \in [0, 1]$ modulate global drift gain and prototype-proximity gating. **Eqs. (18)–(19)** preserve attribution polarity and magnitude bounds since $F_i \in [0, 1]$. For small perturbations $\rho(\mathbf{x}+\delta_i) = \rho(\mathbf{x}) + \Delta\rho_i$ with $\|\Delta\rho_i\| \ll 1$, the approximation

$$1 - \mathcal{F}(\rho(\mathbf{x}), \rho(\mathbf{x}+\delta_i)) \approx \tfrac{1}{4}\|\Delta\psi_i^\perp\|_2^2 \tag{20}$$

relates fidelity change to the Bures metric, confirming its role as a valid quantum sensitivity measure. Alternatively, defining the Shapley-drift value function $v(S) = -\mathsf{D}_{\text{h}}(\mathbf{x}^{(S)}; y)$ [39] yields gate- and feature-level attributions that quantify drift reduction, complementing prediction-based explanations.

To further align Q-SHIELD with prior interpretability frameworks, an attribution stability term is introduced [56], analogous to the gate-level variance $\sigma[\phi_i]$ used to assess Shapley consistency. In Q-SHIELD, the corresponding stability of fidelity-weighted attributions is expressed as

$$\sigma[s_i] = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\big(s_i^{(r)} - \bar{s_i}\big)^2}, \quad \bar{s_i} = \frac{1}{R}\sum_{r=1}^{R} s_i^{(r)}, \tag{21}$$

where $s_i^{(r)}$ denotes the attribution value obtained from run r under randomized encoder perturbations or shot noise, and R is the number of trials. This Monte-Carlo estimate provides a directly comparable scalar to $\sigma[\phi_i]$ in [39]. It can be used to quantify Q-SHIELD attributions under stochastic quantum effects, establishing a bridge between gate-level Shapley consistency and inference-level attribution robustness in hybrid quantum-classical O-RAN models. **Algorithm 2** Lines 1–2 encode the clean quantum state and compute global cosine drift to the class center. Lines 3–6 iterate over spectrogram tiles: the sample input is perturbed, re-encoded, and its local fidelity to the clean state is then estimated. Then local drift d_i is defined as $1-\mathcal{F}_i$. The same fidelity gates the baseline attribution, further modulated by the global drift and prototype proximity to reflect class geometry. Line 7 normalizes the re-weighted scores and arranges both attribution and drift into aligned heatmaps for analysis and interpretation.

VI. Evaluation and Discussion

The evaluation directly follows the threat model and attack mechanisms defined in Section IV and Section V. Specifically, QC-FGSM and QC-PGD are instantiated as input-space evasion attacks, where bounded perturbations δ are applied to spectrogram inputs $\mathbf{S}$ and propagated through the hybrid pipeline, including the quantum encoder $U_{\text{enc}}(\cdot)$ and variational circuit. QC-FGSM uses a single-step gradient update, while QC-PGD applies iterative updates under an L_∞ constraint, as described in Section V. In contrast, QC-Poison is evaluated as a parameter-space attack aligned with the encoding-level and fusion-level perturbations described in Section IV, where adversarial drift is induced in the quantum representation and measured via the deviation $\|\Delta\theta\|_\infty$ relative to the clean anchor. The QAT baseline follows the min-max formulation in Section V, where QC-PGD perturbations are generated during training and injected into the hybrid pipeline. Unlike Q-SHIELD, QAT does not include quantum state transformation (QST) or decision-boundary reinforcement (Q-DBR), allowing us to isolate the contribution of these modules. All reported robustness metrics, including worst-case accuracy $\text{Acc}_{\min}$ and attack success rate (ASR), are computed over the perturbation range $\epsilon_q \in \{0.01, 0.03, 0.04, 0.05 - 0.1\}$ defined in Table IV.

A. Dataset Description

We used the publicly available wireless interference dataset introduced in [2]. The dataset consists of 10,000 spectrogram

TABLE IV: Experimental setup for hybrid quantum–classical model.

Aspect	Setting
Hardware	HPC cluster; NVIDIA RTX A6000 GPU and CUDA V-12.2
Software	`PennyLane` 0.30.0 + `PyTorch` 1.13.1
Backbone	ResNet-18 $\rightarrow$ 512-d $\rightarrow$ 6-d (quantum embed.)
Q-simul.	`default.qubit`; 6 qubits, mean of 3 seeds
Circuit	$3 \times [U_3^{\otimes 6} + \mathrm{ZZ}_{\mathrm{dense}}]$; readout $\langle Z \rangle^{\otimes 6}$
Hyper-par	`Optuna` 3.2; 50 trials; lr $\in [10^{-5}, 10^{-2}]$; batch $\in \{32, 64\}$
Training	AdamW (wd 10^{-4}); dropout 0.3; BN; cosine anneal
Perturb.	QC-FGSM/PGD/poison; $\epsilon_q \in \{0.01, 0.03, 0.05, 0.1\}$– pert.scale

images of size 128×128 (grayscale), divided into two classes: 5,000 representing uplink UE signals of interest without interference and 5,000 with CWI. Each spectrogram corresponds to a 10 ms LTE frame, where the x-axis represents time and the y-axis represents frequency. SOI spectrograms were generated from uplink transmissions at a carrier frequency of 2.56 GHz using 25 physical resource blocks (PRBs), corresponding to approximately 5 MHz of bandwidth, and sampled at 7.68 MS/s. Uplink TCP traffic was created using `iperf3` between the UE and the RAN. For interference scenarios, CWI signals were transmitted at high gain values (30–40 dB) on the same carrier frequency as the SOI to evaluate the robustness of near-RT RIC components. The signals were transmitted over-the-air using the open-source `srsRAN` stack and USRPs, with MATLAB scripts generating baseband I/Q samples for CWI. After OTA transmission, spectrograms were resized and converted to grayscale for model training. The hardware, software, and training configurations of the hybrid learning framework are summarized in **Table IV**.

All experiments are conducted using a state-vector simulation backend (PennyLane `default.qubit`) integrated with PyTorch and executed on GPU hardware. This enables controlled evaluation of hybrid quantum-classical models under adversarial perturbations. While this setting avoids hardware-induced noise, real quantum devices introduce stochasticity due to finite measurement shots and decoherence. In such cases, fidelity must be estimated from measurement statistics, resulting in bounded variance. The proposed framework relies on relative fidelity variations, which reduces sensitivity to such stochastic effects and supports extension to noise-aware quantum implementations. Importantly, the proposed attack/defense mechanisms are expressed at the circuit-encoding and measurement levels using hardware-compatible primitives (e.g., single-qubit rotations and two-qubit entanglers), and are implemented through PennyLane's device abstraction so the same QNode can be executed on a simulator or QPU backends without changing the formulation. However, hardware-calibrated evaluation and end-to-end validation under realistic quantum device noise are left for future work when commercial quantum computing systems become more widely accessible.

For measurement, we analyze the baseline models with metrics on accuracy and training dynamics. Next, we evaluate its robustness against the three considered adversarial attacks and examine the effects of two defense strategies: QAT (adversarial training only) and Q-SHIELD (full protection mode) on latent stability, particularly in suppressing quantum drift and enhancing recovery performance. Furthermore, we report the runtime robustness gain, denoted as Δ_{ACC}, where a positive value of Δ_{ACC} signifies improved robustness attributable to quantum-enhanced defenses. Model size is also assessed to determine its suitability for near-real-time deployment in RIC systems. Finally, we introduce an interpretability analysis using a fidelity-weighted attribution method in Q-SHIELD, applied to radio spectrogram data, highlighting the contribution of salient features. Detailed results and discussions are presented below.

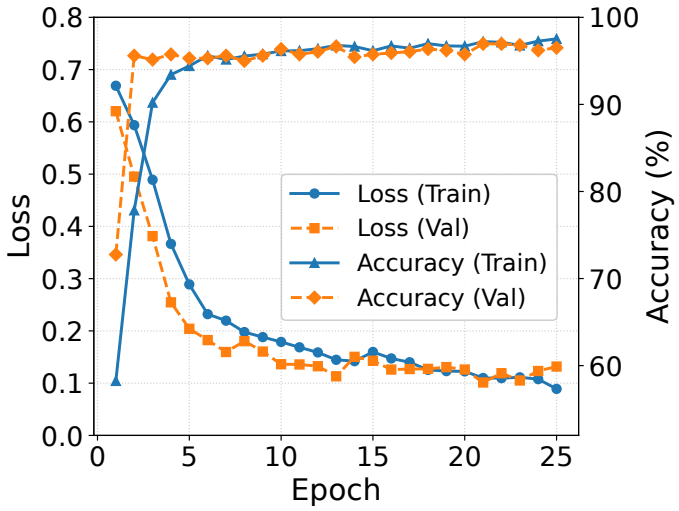

Fig. 4: Training and validation loss/acc, under no-attack scenario.

B. Evaluation results on various scenarios

Under standard training, the hybrid quantum–classical xApp attains 0.982(98.2%) accuracy, matching CNN/DNN xApps for O-RAN interference classification and exhibiting only a negligible generalization gap, as shown in **Fig. 4**. When exposed to gradient-based RF adversaries as seen in **Table V** shows the vulnerability of the baseline model under adversarial conditions, where accuracy drops sharply across QC-FGSM, QC-PGD, and poisoning attacks (e.g., PGD worst-case accuracy falls to 0.160(16%) at $\epsilon = 0.10$). The accuracy vs various ϵ profiles in **Fig. 5** reveal that the undefended model collapses rapidly. In contrast, under defense, both QAT and layered Q-SHIELD in **Fig. 6** maintain a wide quantum decision margin and preserve high accuracy, even against strong evasion attacks and transient poisoning-drift attacks with minimal training overhead.

Table VI compares clean and worst-case robustness for the fully QNN, an undefended baseline model, QAT, and the proposed Q-SHIELD stack. The undefended baseline drops to 0.558/0.160(55.8%/16%) at $\epsilon = 0.1$ with a QC-Poison residual of $(0.239/3.1 \times 10^{-2})$ under attack. QAT recovers much of the QC-FGSM/PGD margin worst-case 0.804/0.692(80.4%/69.2%), but still leaks accuracy when weights are poisoned 0.611(61.1%). Q-SHIELD achieves the

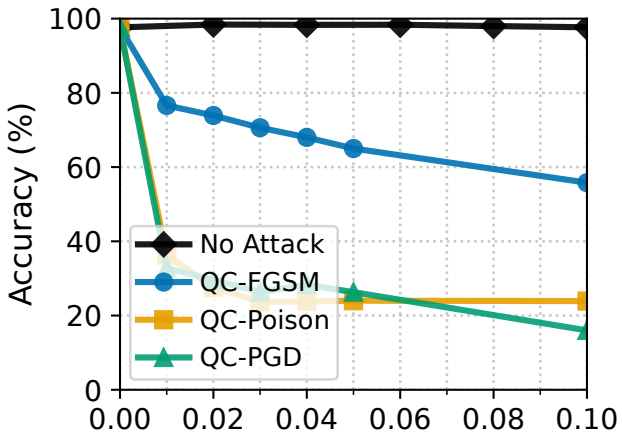

Fig. 5: Attack scenarios: accuracy vs. perturbation scale ϵ.

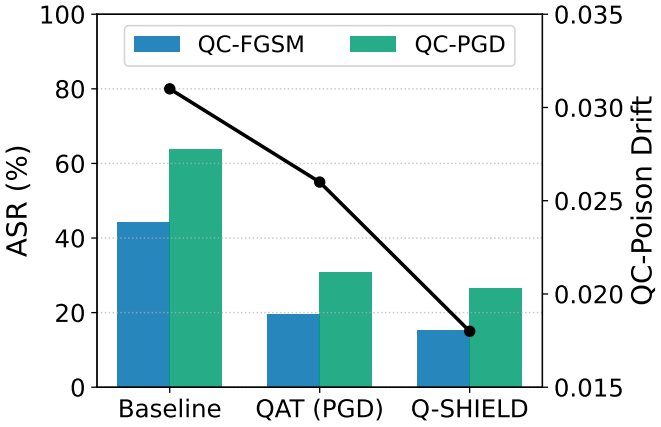

Fig. 6: ASR: Baseline vs. Q-SHIELD poison drift reduction.

TABLE V: Performance metrics under various attack scenarios

Scenario	Scale (ϵ)	ACC	PREC	REC	F1	ROC
Baseline	No attack	0.982	0.980	0.858	0.914	0.999
Q-FGSM [48]	0.05	0.430	0.530	0.240	0.330	0.635
	0.10	0.500	0.50	0.100	0.670	0.585
	0.15	0.280	0.170	0.040	0.600	0.520
	0.20	0.720	0.710	0.870	0.780	0.890
	0.25	0.620	0.660	0.840	0.740	0.835
	0.30	0.480	0.470	0.890	0.620	0.780
QC-FGSM	0.01	0.766	0.632	0.319	0.387	0.723
	0.02	0.739	0.535	0.258	0.324	0.695
	0.03	0.706	0.416	0.215	0.235	0.658
	0.04	0.680	0.330	0.179	0.194	0.627
	0.05	0.650	0.253	0.151	0.160	0.592
	0.10	**0.558**	**0.028**	**0.029**	**0.014**	**0.510**
QC-Poison	0.01	0.363	0.275	0.826	0.413	0.508
	0.02	0.276	0.254	0.866	0.393	0.461
	0.03	0.237	0.244	0.866	0.381	0.435
	0.04	0.238	0.253	0.866	0.381	0.435
	0.05	0.240	0.245	0.866	0.382	0.437
	0.10	**0.239**	**0.245**	**0.866**	**0.382**	**0.436**
QC-PGD	0.01	0.328	0.106	0.214	0.142	0.203
	0.02	0.296	0.093	0.194	0.126	0.177
	0.03	0.263	0.076	0.153	0.101	0.163
	0.04	0.283	0.069	0.142	0.093	0.147
	0.05	0.263	0.064	0.128	0.085	0.138
	0.10	**0.160**	**0.049**	**0.090**	**0.064**	**0.133**

strongest robustness overall, with baseline worst-case accuracies of 0.847/0.736(84.7%/73.6%), the lowest attack success rates, and QC-Poison accuracy of 0.708(70.8%) while limiting parameter drift to 0.018 less than half that of the undefended model. This shows Q-SHIELD preserves decision accuracy and quantum parameters under strong adversaries.

To contextualize the observed robustness improvement, the recent work by Naik *et al.* [2] reported that a conventional CNN-based interference classifier deployed as an O-RAN xApp achieved a clean accuracy of approximately 98.5% on the real RF dataset. However, under adversarial perturbations, its performance degraded drastically, dropping to nearly 25% under an FGSM attack ($\epsilon \approx 0.04$) and collapsing completely to 0% accuracy under PGD perturbations. These results confirm the inherent fragility of classical ML models. In contrast, the proposed Q-SHIELD framework preserves a substantially higher degree of resilience, maintaining 84.7% accuracy under FGSM and 73.6% under PGD attacks while retaining 97.4% clean accuracy. This corresponds to an average relative robustness gain of more than $3\times$ compared with the classical CNN baseline and a clear margin over both the QNN baseline [48] and the PGD-trained QAT defense. Such results underscore the comparative advantage of quantum-enhanced feature encoding and decision boundary reinforcement for robust xApps.

Table VIII reports the training cost of the three models. As expected, the baseline is the most lightweight, requiring only $1128\,\mathrm{s}$ per epoch and reaching optimal accuracy within $4.7\,\mathrm{h}$; however, this efficiency comes at the expense of robustness. Both QAT and Q-SHIELD incur substantially higher cost, with total training times of $45.8\,\mathrm{h}$ and $42.4\,\mathrm{h}$, respectively. Although Q-SHIELD is only moderately faster than QAT, it achieves noticeably smoother convergence and better accuracy. Given that training is typically performed offline and updates are periodic in practical O-RAN deployments, this cost is acceptable, making Q-SHIELD a viable and robust alternative.

C. Discussion on Q-SHIELD's advantages

Table VI shows that directly adapting conventional adversarial training, such as PGD-based QAT, is insufficient for the considered hybrid quantum-classical threat model. Although QAT improves resistance to QC-FGSM and QC-PGD, it remains vulnerable to QC-Poison-induced parameter and Hilbert-space drift. Q-SHIELD addresses this limitation by jointly combining manifold-guided QST with reliability-based Q-DBR, thereby protecting both the encoded quantum representation and the hybrid decision boundary.

Across the evaluated attacks, Q-SHIELD reduces QC-FGSM attack success rate-ASR from 0.442 to 0.153 by 65.4% and QC-PGD ASR from 0.639 to 0.264 by 58.7%, while reducing poisoning-induced parameter drift $\|\Delta\theta\|_\infty$ from 3.1×10^{-2} to 1.8×10^{-2} by 41.9%. Compared to conventional adversarial training (QAT), Q-SHIELD also achieves a 22.9% reduction in total training time and 15.6% faster convergence to its best validation accuracy (see Table VIII), demonstrating computational efficiency in addition to robustness gains. As further shown in

TABLE VI: Q-SHIELD robustness report: clean accuracy, worst-case accuracy (max. $\epsilon = 0.1$ grid), attack success rate–ASR, and QC-Poison $\rightarrow$ mean L_∞ parameter deviation relative to clean anchor.

Scenario	Pre-attk checks	FGSM $\text{Acc}_{\min}$	FGSM ASR	PGD $\text{Acc}_{\min}$	PGD ASR	QC-Poison $\text{Acc}_{\min}$/ $\|\Delta\theta\|_\infty$
CNN [2]	0.985	0.250	0.750	0.000	1.000	NA
QNN [48]	0.520	0.500	0.500	NA	NA	NA
Baseline	0.982	0.558	0.442	**0.160**	0.639	**0.708** / 3.1×10^{-2}
QAT (PGD)	0.948	0.804	0.196	0.692	0.308	0.611 / 2.6×10^{-2}
Q-SHIELD	**0.974**	**0.847**	**0.153**	**0.736**	**0.264**	**0.239** / 1.8×10^{-2}

Fig. 6, these gains indicate that state alignment and parameter regularization jointly limit single-step perturbations, iterative gradient accumulation, and parameter-space poisoning. Thus, the primary advantage of Q-SHIELD is its layered protection across both input and parameter-level attack surfaces, providing greater robustness than QAT alone without requiring a separate defense for each attack mechanism.

D. Discussion on model explainability

To interpret how the hybrid QC classifier forms decisions, we apply the fidelity-weighted attribution mechanism derived in **Section V-D**. For a spectrogram input x, local tile perturbations δ_i produce perturbed quantum states $\rho(x+\delta_i)$ and fidelity scores $\mathcal{F}_i = \mathcal{F}(\rho(x), \rho(x+\delta_i))$. The resulting drift values $\mathsf{d}_i \leftarrow 1 - \mathcal{F}_i$ indicate each tile's quantum sensitivity. **Fig. 7** visualizes these effects for SOI (a) and CWI (d) samples. Panels (b),(e) show baseline Shapley-style attributions s_i, which mainly follow broad spectral energy. Panels (c),(f) show fidelity-weighted scores $s'_i = s_i(1-F_i)$, which highlight tiles that produce meaningful Hilbert-space responses. For the SOI, Q-SHIELD isolates harmonic rows that drive quantum-state change. For the CWI, it emphasizes periodic interference ridges rather than diffuse background energy. These regions correspond to tiles most capable of shifting the quantum encoder, making them operationally relevant for attack analysis.

To quantify alignment with Hilbert drift, **Table VII** reports the Spearman correlation $\rho(s_i, d_i)$ between baseline or fidelity-weighted attributions and their corresponding drift scores. Baseline attributions yield $\rho \approx 1$, dominated by global energy structure. Fidelity weighting reduces the correlation ($\rho \approx 0.86$–0.99) because it suppresses low-impact tiles and concentrates mass on a smaller, quantum-sensitive region. We further evaluate a locality score, defined as the fraction of total attribution mass contained in the top 20% drift-responsive tiles, $L_{20} = \frac{\sum_{i\in\Omega_{20}} |s'_i|}{\sum_j |s'_j|}$, where Ω_{20} denotes the highest-drift tiles. L_{20} measures compactness: higher values indicate that the explanation is concentrated on a small, meaningful subset of tiles. For Q-SHIELD, L_{20} for the SOI sample exceeds 0.80, forming a tight attribution cluster aligned with **Fig. 7**. For CWI, L_{20} is lower (≈ 0.61), consistent with the wider and brighter interference ridge structure visible in the spectrogram.

Finally, attribution stability is assessed through the variance $\sigma[s'_i]$ defined in **Eq. (21)**. Stability values below 0.01 indicate minimal fluctuation under random encoder perturbations, confirming consistency across inference runs. Together, $\rho(s_i, d_i)$, L_{20}, and $\sigma[s'_i]$ provide a domain-appropriate interpretability

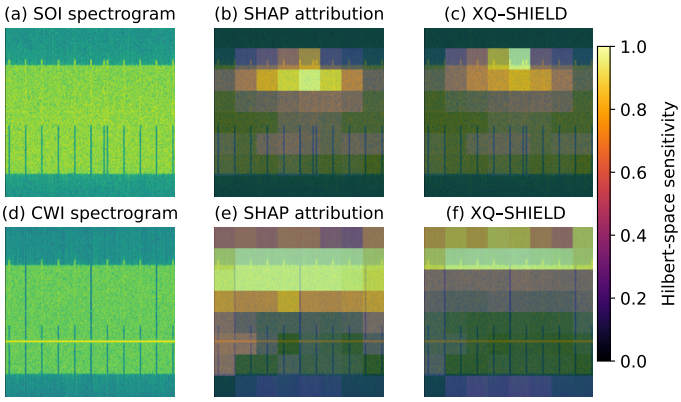

Fig. 7: Visualization of fidelity-weighted attribution in hybrid QC-RF spectrogram classifier. (a)–(c) show the SOI and (d)–(f) the CWI. Baseline attributions-SHAP explainer (b),(e) highlight general spectral regions, while Q-SHIELD tracing (c),(f) emphasize quantum-sensitive time-frequency tiles where local perturbations cause Hilbert-space drift, showing physically meaningful decision regions.

TABLE VII: Drift alignment, locality, and stability.

Class	ρ (Drift)		Locality (L_{20})		Stability	
	Baseline	Q-SHIELD	Baseline	Q-SHIELD	Baseline	Q-SHIELD
SOI	1.000	0.988	0.943	0.871	0.001	0.000
CWI	1.000	0.860	0.637	0.612	0.006	0.007

ρ: Spearman correlation between attrib. scores and tile-wise Hilbert drift. *Locality*$_{20}$: fraction of attribution mass in the top 20% drift-responsive tiles. *Stability*: variance of fidelity-weighted attribution across random encoder perturbations.

profile. Correlation indicates Hilbert-space alignment; locality quantifies the compactness of quantum-sensitive tiles and stability measures robustness under stochastic effects. In the security context, these metrics reveal the model's attack-relevant regions. These regions are tiles whose perturbations induce the largest quantum state changes. This supports O-RAN requirements for auditable and trustworthy xApp behavior.

Note that the attribution analysis is introduced as an interpretability mechanism to interpret the robustness behavior of the proposed defense framework, rather than as an independent explainable AI component. Specifically, it traces robustness degradation back to the time-frequency regions whose perturbations most strongly induce Hilbert-space drift at the quantum-classical interface, reflected by significant changes in the VQC measurement embedding and the associated fidelity score. The proposed fidelity-weighted attribution is directly derived from the same drift formulation used in the defense mechanism, ensuring methodological consistency between the robustness and explainability analyses. Consequently, the attribution results provide insight into how adversarial perturbations affect the hybrid quantum-classical decision process and help validate that the observed performance degradation is driven by interface-level drift rather than benign input variations. This coupling enables a more transparent understanding of the defense behavior while preserving a unified framework for robustness analysis and interpretation.

TABLE VIII: Training efficiency and near-RT deployment cost for the evaluated models.

Scenario	Best val. acc.	Mean epoch time [s]	Total train time [hr]	Time-to-best acc. [hr]
Baseline	0.982	1,128	7.83	4.70
QAT (PGD)	0.948	6,600	45.8	42.2
Q-SHIELD	0.974	6,100	42.4	35.6

E. Discussion on Q-SHIELD vs. Classical Models

The classical CNN baseline(without defense) achieves 98.5% clean accuracy but degrades to nearly zero under PGD, indicating substantial sensitivity to adversarial perturbations. Under the corresponding hybrid attacks, Q-SHIELD retains 84.7% and 73.6% accuracy against QC-FGSM and QC-PGD, respectively, while limiting quantum-parameter drift to 0.018 without post-attack retraining. This robustness incurs a higher offline training cost, with convergence requiring 25 epochs (within a maximum of 500 epochs) compared with 15 for the classical CNN (out of 300 total epochs). The comparison is conducted under early-stopping criteria for fair convergence tracking. This difference concerns training complexity rather than near-RT inference latency, which depends on the quantum backend and deployment platform.

The results in Table VI and VIII position hybrid quantum-classical models as complements to, rather than replacements for, classical xApps. Classical models remain preferable in Fig. 4: ① settings with stationary interference statistics, where CNN kernels efficiently capture local time-frequency patterns, and ② deployments with strict near-RT latency constraints, where GPU-optimized inference provides predictable and low-cost execution. Q-SHIELD is more relevant in ③ security-critical settings where adversarial perturbations or representation drift destabilize classical decision boundaries. Its 6-qubit VQC, comprising three repetitions of $[U_3^{\otimes 6} + \mathrm{ZZ}_{\mathrm{dense}}]$ and 612 trainable quantum parameters, keeps the quantum component compact. However, the end-to-end model also includes the frozen ResNet-18 backbone; thus, compactness of the VQC alone does not establish a reduced deployment footprint.

Accordingly, the comparison demonstrates an empirical robustness advantage under the evaluated dataset, attacks, and model configurations, rather than a universal quantum advantage over classical learning. Selecting Q-SHIELD for a practical xApp therefore requires balancing its improved adversarial stability against quantum-execution overhead. End-to-end latency, memory, energy consumption, and hardware-dependent reliability must be quantified before confirming its suitability for near-RT RIC deployment.

VII. Conclusion

This work presented Q-SHIELD, a security defense model for hybrid quantum-classical O-RAN xApps exposed to adversarial input perturbations and model-level tampering. Q-SHIELD combines quantum state transformation with quantum decision-boundary reinforcement to preserve encoded-state geometry and stabilize classification margins under QC-FGSM, QC-PGD, and QC-Poison attacks. It achieves 98.2% clean accuracy and retains 73.6% accuracy under the strongest evaluated attack, using a frozen ResNet-18 backbone and 612 trainable quantum parameters only. The proposed fidelity-weighted attribution mechanism further links time–frequency feature importance to Hilbert-space drift. It achieves attribution–drift correlations of $\rho \approx 0.86$–0.99 and concentrates more than 60–87% of the explanatory mass within the most responsive 20% of spectrogram regions. An attribution variance below 0.01 further demonstrates explanation stability under encoder perturbations. Collectively, these results show that Q-SHIELD provides adversarial robustness together with stable, measurement-driven explanations for hybrid interference classification.

Future research will extend hybrid-model robustness beyond signal inference. First, life-cycle security mechanisms are needed to detect and quarantine drifted, tampered, or mismatched encoders before their activation in the RIC. Second, adaptive quantum anchoring should allow class prototypes to track non-stationary interference without accumulating geometric bias, which remains an open challenge for hybrid quantum models. Third, Hilbert-space drift can be exposed as cross-layer telemetry so that scheduling, resource-allocation, and slice-control policies respond to detected signal anomalies. Finally, evaluating Q-SHIELD with hardware-efficient circuits and noisy quantum backends is essential for assessing its transition from simulation to deployable O-RAN microservices.

References

[1] M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding o-ran: Architecture, interfaces, algorithms, security, and research challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, pp. 1376–1411, 2023.

[2] N. N. Sapavath, B. Kim, K. Chowdhury, and V. K. Shah, "Experimental study of adversarial attacks on ml-based xapps in o-ran," in *2023 IEEE Global Commun Conf.*, 2023, pp. 6352–6357.

[3] N. Q. Thoong, A. A. Cheema, B. Canberk, D. T. Tran, O. A. Dobre, and T. Q. Duong, "Channel estimation for reconfigurable intelligent surface-aided 6g noma systems: A quantum machine learning approach," *IEEE Trans. Netw. Sci. Eng.*, 2025.

[4] T. Hur, L. Kim, and D. K. Park, "Quantum cnn for classical data classification," *Quantum Mach. Intell.*, vol. 4, p. 3, 2022.

[5] L. Pira and C. Ferrie, "On the interpretability of quantum neural networks," *Quantum Mach. Intell.*, vol. 6, no. 2, p. 52, 2024.

[6] K. Kadian, S. Garhwal, and A. Kumar, "Exqual: an explainable quantum machine learning classifier," *Appl. Intell.*, vol. 55, no. 13, p. 940, 2025.

[7] S. Ruan, Z. Liang, Q. Guan, P. Griffin, X. Wen, Y. Lin, and Y. Wang, "Violet: Visual analytics for explainable quantum neural networks," *IEEE Trans. Vis. Comput. Graph*, vol. 30, no. 6, pp. 2862–2874, 2024.

[8] B. Li, T. Alpcan, C. Thapa, and U. Parampalli, "Computable model-independent bounds for adversarial quantum machine learning," *IEEE Trans. Quantum Eng*, vol. 6, pp. 1–22, 2025.

[9] J. Guo, W. Jiang, R. Zhang, W. Fan, J. Li, G. Lu, and H. Li, "Backdoor attacks against hybrid classical-quantum neural networks," *Neural Networks*, vol. 191, p. 107776, 2025.

[10] S. Kundu and S. Ghosh, "Adversarial data poisoning attack on quantum machine learning in the nisq era," in *Proc. Great Lakes VLSI (GLSVLSI 2025)*, ser. GLSVLSI '25. New York, NY, USA: ACM, 2025.

[11] V.-L. Nguyen, L.-H. Nguyen, R.-H. Hwang, B. Canberk, and T. Q. Duong, "Quantum machine learning for 6g network intelligence and adversarial threats," *IEEE Commun. Stand. Mag.*, pp. 1–1, 2025.

[12] A. Khatun and M. Usman, "Quantum transfer learning with adversarial robustness for classification of high-resolution image datasets," *Adv. Quantum Technol.*, vol. 8, no. 1, p. 2400268, 2025.

[13] O-RAN Next Generation Research Group (nGRG), "Research report on quantum security," O-RAN Alliance, Technical Report RR-2023-04, Sep. 2023, contributors: HCLSoftware, CICT, Dell, Nokia, Rakuten Mobile, Reliance Jio, Qualcomm.

[14] J. Groen, S. D'Oro, U. Demir, L. Bonati, M. Polese, T. Melodia, and K. Chowdhury, "Implementing and evaluating security in o-ran: Interfaces, intelligence, and platforms," *IEEE Netw.*, vol. 39, no. 1, pp. 227–234, 2025.

[15] S. B. Mallikarjun, M. Asif Habibi, M. Dixit, X. Costa-Pérez, M. Debbah, and H. D. Schotten, "Exploring y1 communications and services in o-ran: Background, privacy, and security," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 4638–4666, 2025.

[16] Y. Abera Ergu and V.-L. Nguyen, "Radar: Robust drl-based resource allocation against adversarial attacks in intelligent o-ran," *IEEE Trans Green Commun Netw*, vol. 9, no. 4, pp. 2305–2318, 2025.

[17] A. S. Abdalla and V. Marojevic, "End-to-end o-ran security architecture, threat surface, coverage, and the case of the open fronthaul," *IEEE Commun. Mag.*, vol. 8, no. 1, pp. 36–43, 2024.

[18] H. Bogucka, P. Kryszkiewicz, M. Hoffmann, and M. Wasilewska, "The o-ran whitepaper 2023: Security in o-ran," Rimedo Labs, Tech. Rep., August 2023, whitepaper on security in Open RAN (O-RAN) networks.

[19] Y. A. Ergu, V.-L. Nguyen, R.-H. Hwang, Y.-D. Lin, C.-Y. Cho, and H.-K. Yang, "Unmasking vulnerabilities: Adversarial attacks against drl-based resource allocation in o-ran," in *ICC 2024 - IEEE Int. Conf. Commun.*, 2024, pp. 2378–2383.

[20] C.-F. Hung, Y.-R. Chen, C.-H. Tseng, and S.-M. Cheng, "Security threats to xapps access control and e2 interface in o-ran," *IEEE Open J. Commun. Soc*, vol. 5, pp. 1197–1203, 2024.

[21] W. Gong, D. Yuan, W. Li, and D.-L. Deng, "Enhancing quantum adversarial robustness by randomized encodings," *Phys. Rev. Res.*, vol. 6, no. 2, p. 023020, 2024.

[22] M. Karbalaee Motalleb, C. Benzaid, T. Taleb, M. Katz, V. Shah-Mansouri, and J. Kim, "Towards secure intelligent o-ran architecture: vulnerabilities, threats and promising technical solutions using llms," *Digit. Commun. Netw.*, 2025.

[23] V.-L. Nguyen and Y. A. Ergu, "Adversarial attacks on hybrid quantum-classical interference classifier in intelligent o-ran," *IEEE Trans. Netw. Serv. Manag.*, vol. 23, pp. 6493–6506, 2026.

[24] M. R. Uddin, S. Shaon, R. Rahman, D. C. Nguyen, O. Dobre, and D. Niyato, "Quantum federated learning: A comprehensive survey," *IEEE Commun. Surv. Tutor*, pp. 1–1, 2025.

[25] S. Lu, L.-M. Duan, and D.-L. Deng, "Quantum adversarial machine learning," *Phys. Rev. Res.*, vol. 2, no. 3, Aug. 2020.

[26] Y. Du, M.-H. Hsieh, T. Liu, D. Tao, and N. Liu, "Quantum noise protects quantum classifiers against adversaries," *Phys. Rev. Res.*, vol. 3, no. 2, p. 023153, 2021.

[27] H. Liao, I. Convy, W. J. Huggins, and K. B. Whaley, "Robust in practice: Adversarial attacks on quantum machine learning," *Physical Review A*, vol. 103, no. 4, p. 042427, 2021.

[28] W. El Maouaki, A. Marchisio, T. Said, M. Shafique, and M. Bennai, "Designing robust quantum neural networks via optimized circuit metrics," *Adv. Quantum Technol.*, p. 2400601, 2025.

[29] F. Ullah, N. Mohammad, L. Mostarda, D. Cacciagrano, and Y. Zhao, "Q-p2fl: Quantum-enhanced federated edge intelligence for privacy-preserving adversarial attack detection on consumer edge devices," *IEEE Trans. Consumer Electron*, 2025.

[30] M. T. West, S.-L. Tsang, J. S. Low, C. D. Hill, C. Leckie, L. C. Hollenberg, S. M. Erfani, and M. Usman, "Towards quantum enhanced adversarial robustness in machine learning," *Nature Machine Intelligence*, vol. 5, no. 6, pp. 581–589, 2023.

[31] Y. Huang, L. Wang, and J. Xu, "Quantum entanglement path selection and qubit allocation via adversarial group neural bandits," *IEEE Trans. Netw*, vol. 33, no. 2, pp. 583–594, 2025.

[32] F. Ullah, N. Mohammad, L. Mostarda, D. Cacciagrano, and Y. Zhao, "Q-p2fl: Quantum-enhanced federated edge intelligence for privacy-preserving adversarial attack detection on consumer edge devices," *IEEE Trans. Consum. Electron*, vol. 71, no. 2, pp. 4914–4924, 2025.

[33] W. Yamany, N. Moustafa, and B. Turnbull, "Oqfl: An optimized quantum-based federated learning framework for defending against adversarial attacks in intelligent transportation systems," *IEEE Trans. Intell. Transp. Syst*, vol. 24, no. 1, pp. 893–903, 2023.

[34] C. Lin, F. Cheng, P. Yang, T. He, J. Liang, and Q. Wang, "Qans: Toward quantized neural network adversarial noise suppression," *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst*, vol. 42, pp. 4858–4870, 2023.

[35] Y. A. Ergu, V.-L. Nguyen, P.-C. Lin, and R.-H. Hwang, "Q-poison: Quantum adversarial attacks against qml-driven interference classification in o-ran," in *Proc. IEEE Int. Conf. Mach. Learn. Commun. Netw.*, 2025.

[36] R. Razavi-Far, M. Meymani, E. Mahmoudinia, D. Vazirzade, P. Paknezhad, F. Ghasemi, S. Saravani, S. Nikkhoo, and K. Haghjooei, "Quantum adversarial machine learning: from classical adaptations to quantum-native methods," *Artificial Intelligence Review*, 2026.

[37] G. Montalbano and L. Banchi, "Quantum adversarial learning for kernel methods," *Quantum Machine Intelligence*, vol. 7, no. 1, p. 15, 2025.

[38] D. Arias, I. G. R. de Guzman, M. Rodriguez, E. B. Terres, B. Sanz, J. G. de la Puerta, I. Pastor, A. Zubillaga, and P. G. Bringas, "Let's do it right the first time: Survey on security concerns in the way to quantum software engineering," *Neurocomputing*, vol. 538, p. 126199, 2023.

[39] R. Heese, T. Gerlach, S. Mücke, S. Müller, M. Jakobs, and N. Piatkowski, "Explaining quantum circuits with shapley values: Towards explainable quantum machine learning," *Quantum Mach. Intell.*, vol. 7, pp. 1–33, 2025.

[40] A. K. K. Don and I. Khalil, "Qrlaxai: Quantum representation learning and explainable ai," *Quantum Mach. Intell.*, vol. 7, no. 1, p. 24, 2025.

[41] A. Routh, A. Bhattacharjee, and S. Roy, "An empirical study on enhancing explainability through quantum machine learning for credit risk analysis," in *Proc. of IEEE Int. Conf. Front. Mach. Learn. Data Sci. (FMLDS 2024)*, 2024, pp. 1–8.

[42] C. Fiandrino, G. Attanasio, M. Fiore, and J. Widmer, "Toward native explainable and robust ai in 6g networks: Current state, challenges and road ahead," *Comput. Commun.*, vol. 193, pp. 47–52, 2022.

[43] B. Brik, H. Chergui, L. Zanzi, F. Devoti, A. Ksentini, M. S. Siddiqui, X. Costa-Pèrez, and C. Verikoukis, "Explainable ai in 6g o-ran: A tutorial and survey on architecture, use cases, challenges, and future research," *IEEE Commun. Surveys Tuts.*, 2024.

[44] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst. (NeurIPS'17)*. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[45] F. A. et al., "Quantum supremacy using a programmable superconducting processor," *Nature*, vol. 574, no. 574, p. 505–510, 2019.

[46] U. Mahadev, "Classical verification of quantum computations," in *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018, pp. 259–267.

[47] D. Heimann, H. Hohenfeld, G. Schönhoff, E. Mounzer, and F. Kirchner, "Learning fourier series with parametrized quantum circuits," *Phys. Rev. Res.*, vol. 7, p. 023151, May 2025.

[48] M. S. Akter, H. Shahriar, A. Cuzzocrea, and F. Wu, "Quantum adversarial attacks: Developing quantum fgsm algorithm," in *Proc. of IEEE Comput. Softw. Appl. Conf. 2024 (COMPSAC 2024)*, 2024, pp. 1073–1079.

[49] J. Hou and X. Qi, "Fidelity of states in infinite-dimensional quantum systems," *Sci. China Phys. Mech. Astron.*, vol. 55, no. 10, pp. 1820–1827, 2012.

[50] W. Zhang, M. Krunz, and G. Ditzler, "Stealthy adversarial attacks on machine learning-based classifiers of wireless signals," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 261–279, 2024.

[51] P. Baguer, G. M. Yilma, E. Municio, G. Garcia-Aviles, A. Garcia-Saavedra, M. Liebsch, and X. Costa-Pérez, "Attacking o-ran interfaces: Threat modeling, analysis and practical experimentation," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 4559–4577, 2024.

[52] National Institute of Standards and Technology, "Cve-2024-34473: Improper input validation in o-ran near-rt ric," 2024. [Online]. Available: https://nvd.nist.gov/vuln/detail/CVE-2024-34473

[53] M. Barbeau and J. Garcia-Alfaro, "Cyber-physical defense in the quantum era," *Sci. Rep.*, vol. 12, no. 1, p. 1905, 2022.

[54] T. T. An, S. L. Cotton, J. Zhang, Y. Ding, and T. Q. Duong, "Lora radio frequency fingerprinting identification using a hybrid quantum-classical neural network," in *Proc. of IEEE Veh. Technol. Conf.(VTC2024-Fall)*, 2024, pp. 1–6.

[55] P. de Schoulepnikoff, G. Muñoz Gil, H. P. Nautrup, and H. J. Briegel, "Interpretable representation learning of quantum data enabled by probabilistic variational autoencoders," *Phys. Rev. A*, p. 062423, 2025.

[56] M. Pawlicki, A. Pawlicka, F. Uccello, S. Szelest, S. D'Antonio, R. Kozik, and M. Choraś, "Evaluating the necessity of the multiple metrics for assessing explainable ai: A critical examination," *Neurocomputing*, vol. 602, p. 128282, 2024.