VER
Vascular & Endovascular Review

A Unified Quantum-Resilient Security Framework for Intelligent Healthcare Ecosystems Integrating Zero-Trust Architecture with Explainable Artificial Intelligence

Salman Mohammad Abdullah[1], Md Sajedul Karim Chy[2], Mohammad Yasin[3], Md Himeluzzaman[4], Shaown Mahamud Shakil[5], Tofayel Ahmed Onik[6], Hafiz Aziz Khan[7], Afsana Akter[8]

[1]M.S. in Information Technology (IT), Washington University of Science and Technology, VA, USA
[2]M.S. in Project Management, Washington University of Science and Technology, VA, USA
[3]Master's in Business analytics Trine University, USA
[4]M.Sc. in Information Technology Systems and Management, Washington University of Science and Technology, VA, USA
[5]Bachelor in Human Resource Management, Independent University, Bangladesh
[6]Doctorate in Computer Science, University of the Potomac, USA
[7]Master of Science in Information Technology, Washington University of Science and Technology, Alexandria, VA, USA
[8]Department of Computer Science and Engineering, Daffodil International University, Bangladesh

farsisalman310@gmail.com; Sajedulkarimchy661@gmail.com; arafatchy734@gmail.com; Mzzaman1995@gmail.com; shawon95mahmud@gmail.com; tofayelahmedonik@gmail.com; hafiz.sun30@gmail.com; afsanamimi9456@gmail.com

ABSTRACT

The Internet of Medical Things (IoMT) adds the risk to patients of exposing life-critical telemetry to a simultaneous and prospective adversary. In particular, this paper demonstrates a variety of issues pervasive throughout the existing IoMT intrusion-detection literature, including optimising an isolated single classifier for accuracy where transport-layer cryptography will never survive the post-quantum transition, reporting opaque decisions and unfairly inflating headline scores by leaving identifier columns in the feature set. We propose a comprehensive and four-layered security framework that integrates

Its components include: (i) dual branch attention-fusion detector over the joint network-flow and biometric feature space, generating an exploration hybrid quantum-classical variational branch; (ii) explainable-AI layer which audits its attributions for faithfulness and stability via SHAP/LIME; (iii) Zero-Trust policy engine which converts both detector confidence & explanation signals into per-session trust scores + micro-segmentation decisions and; (iv) overhead analysis of post-quantum cryptography [PQC] - a must-have solution in our current constrained IoMT edge. The proposed detector achieved F1 = 0.681 ± 0.030, MCC = 0.637 ± 0.032 and PR-AUC = 0.832 ± 0.019 on the WUSTL-EHMS-2020 testbed dataset using a leakage-audited 5-fold ×3-repeat stratified evaluation protocol. The most accurate detectors are strongly-tuned gradient-boosted trees (LightGBM F1=0.788), and we transparently report the results of varying tuning settings;

However the model proposed is: (i) the strongest adversarially robust differentiable detector (0.935 stable accuracy under FGSM/PGD to ε = 0.2, collapse ≤0.14 by a 1-D CNN), (ii) edge-deployable (13.2k params, 0.24 ms CPU inference, Raspberry-Pi feasible) and iii) yields trustworthy explanations that are stable and interpretable overall(top-10 feature Jaccard of 0.93 across folds). SHAP assigns 57.1% of decision mass to network features and 42.9% to biometrics, estimating the benefit of multi-modal fusion. Zero-Trust micro-segmentation improves containment against lateral-movement attacks from 0.562 to 0.773 (+21.1 pp) in simulation, and ML-KEM/ML-DSA is shown to be practical at the IoMT edge with sub-millisecond handshake latency. The framework is released as an end-to-end, initialised-seed reproducible pipelined.

KEYWORDS: Internet of Medical Things; Intrusion Detection; Explainable Ai; Zero-Trust Architecture; Post-Quantum Cryptography; Quantum Machine Learning; Multi-Modal Fusion; Adversarial Robustness; Healthcare Cybersecurity.

How to Cite: Salman Mohammad Abdullah, Md Sajedul Karim Chy, Mohammad Yasin, Md Himeluzzaman, Shaown Mahamud Shakil, Tofayel Ahmed Onik, Hafiz Aziz Khan, Afsana Akter, (2023) A Unified Quantum-Resilient Security Framework for Intelligent Healthcare Ecosystems Integrating Zero-Trust Architecture with Explainable Artificial Intelligence, Vascular and Endovascular Review, Vol.6, No.2, 112-125

INTRODUCTION

Connected medical devices have begun to stream continuous physiology telemetry from wearable and bedside sensors, via gateways, to remote clinical servers forming systems referred to as the Internet of Medical Things (IoMT) [1], [28]. While this connectivity enhances outcomes and operational efficiencies, it also has transformed patient safety into a cyber-security challenge: an in-flight man-in-the-middle threat that changes biometric readings, or device spoofing will compel a clinician to act on false data with fatal consequences [1], [30]. A myriad of machine-learning-based intrusion-detection systems (IDS) have emerged in the security literature, targeted at characterizing IoMT traffic [28], [29], [30], [33] and multiple purpose-built testbed datasets have been released the best characterized being WUSTL-EHMS-2020, which is unique for its provision of network-flow metrics that were simultaneously collected with real patient biometrics in a simulated attack setting.

However, four gaps remain, with Usual treatment in isolation.

First, single-objective detection. Numerous IoMT-IDS literature presents a classifier fine-tuned for one headline metric often, an arbitrary optimum value on a single dataset— no consideration of interpretability, robustness and clinical deployment analysis that device operating within a clinical trust boundary requires [28], [33]. An accurate but opaque or brittle-to-small-perturbations detector is not useful in practice, nor is a heavyweight one that requires an ARM-class gateway.

Second, opacity. In clinical and regulatory settings, it is not feasible to make an automated security decision that lacks interpretability. While approaches like SHAP [3] and LIME [4] are often invoked to visualize such predictions, their faithfulness (whether the highlighted features actually drive the prediction) is rarely audited. or stability (do the same features keep appearing with resampling?). A no explanation is better than an unstable one.

Third, it we exclude the quantum horizon. Today, classical key-exchange (RSA, elliptic-curve Diffie–Hellman) protects IoMT sessions. These primitives are secure before a cryptographically relevant quantum computer exists [5], but they become vulnerable in a world where a "harvest-now, decrypt-later" adversary can retain encrypted patient data today for decryption later. The National Institute of Standards and Technology (NIST) has in this way undertaken a post-quantum cryptography (PQC) standardisation effort [14], [15] with lattice-based schemes (ML-KEM/CRYSTALS-Kyber [12], ML-DSA/CRYSTALS-Dilithium[13]) for the same reasons. However, IoMT-IDS papers rarely quantify the battery-powered sensor cost of migrating to these primitives.

Fourth, most insidious are the leaks in evaluation. Some studies using WUSTL-EHMS-2020 also achieve 99%+ accuracies on certain datasets. We show that at least part of this is an artefact: the unprocessed dataset contains identifier columns (source MAC address, source port, packet number) that are almost perfect predictors of the attack label. A model that learns "this MAC address is the attacker" will generalise poorly to a new attacker, while scoring almost perfectly on a random split. A true benchmark should eliminate these leaks.

Contributions. Concretely, this work makes the following contributions:

1. A four-layer unified framework (Fig. 1) incorporating multi-modal detection, explainability, Zero-Trust policy and post-quantum transport for IoMT - to the best of our knowledge treated together as one pipeline on a single real dataset.
2. Two-branch attention-fusion detector that separately processes network-flow and biometric modalities within individual branches before an attention-weighted fusion, along with an exploratory hybrid quantum-classical variational branch, and a component ablation isolating the contribution of each design choice.
3. An audit of leakage which removes identifier features that act as proxy labels to provide honest, reproducible baselines and an explanation for the implausibly high accuracies found in other work.
4. It would be an explanation-faithfulness and -stability analysis (SHAP/LIME) with a quantity of attribution of decision mass to the network versus biometric modality
5. A member of the NEO family era, we believe 30 gets the full treatement: zero-trust trust-scoring and micro-segmentation simulation, lateral-movement containment gain quantification, A PQC overhead analysis (latency, wire size, limited- power edge energy model for IoMT.
6. Complete, statistically based evaluation 5-fold × 3-repeat cross-validation + nested hyper-parameter search, +McNemar and Wilcoxon tests (Holm–Bonferroni correction), +Friedman/Nemenyi ranking, +calibration, adversarial (FGSM/PGD) & poisoning robustness with efficiency profiling all made available as a single seed-fixed fully reproducible pipeline.

The rest of the paper is organised as follows. Section 2 is the literature review, where related work is discussed and intrinsic research gaps are identified. Section 3 introduces the dataset, leakage-audited methodology, four-layer framework and evaluation protocol. Experimental results and performance analysis are then presented in Section 4. Finally, Findings and Implications (5), Limitations (6) and Future Research Directions (7).

RELATED WORK

Machine-learning IDS for IoMT. Connected Healthcare: Learning-based intrusion detection is an area that has matured quickly. Hady et al. [1] developed the Enhanced Healthcare Monitoring System (EHMS) testbed and published WUSTL-EHMS-2020, demonstrating that combining network-flow with biometric features improves detection of man-in-the-middle attacks beyond either modality alone where an artificial neural network on the combined feature set achieves approximately 90% accuracy. Thamilarasu et al. A Mobile-Agent, Deep-Learning Intrusion Detection System for Wireless Body-Area Networks in Internet of Medical Things [30]. Vinayakumar et al. Deep Neural Networks as a generic IDS baseline across many benchmarks [34]. Newaz et al. For example, [29] proposed a behavior-based ML security framework named HealthGuard to protect smart-healthcare systems with HEKA later for personal medical devices [35].

He et al. Stacked autoencoders have been applied for connected-healthcare intrusion detection [36]. Ensemble and fog-cloud designs were also investigated: Kumar et al. [31] presented a compressed framework for IoMT attack detection

through ensemble learning and fog-cloud architecture, while RM et al. [37] employed hybrid PCA–grey-wolf feature-engineering stage in front of a deep classifier. Ghubaish et al. Secondly, [32] reviewed the broader IoMT security perspective. Ferrag et al. Federated deep learning for IoT security [38]. While these works establish the importance of multi-modal and deep models, they typically optimise for a single accuracy objective, report opaque decisions, and crucially do not audit for identifier leakage, which we demonstrate can inflate reported scores dramatically.

Explainable AI for security. SHAP [3] explains feature attributions in an additive, game-theoretic fashion with desirable consistency properties and LIME [4] gives local surrogate explanations for any classifier. Both have been commonplace in security ML now. Yet they are often used decoratively: an interpretation is displayed, but no proof that it accurately reflects the model or that their stability can be established with resampling. We thus characterize explanation quality as a measurable property and report both cross-fold and bootstrap stability of the top attributed features, along with the rank agreement between SHAP and permutation importance.

Zero-Trust architecture. NIST SP 800–207 [2] makes the Zero-Trust model never trust, always verify with continuous authentication and least privilege micro-segmentation their standard. This is ideal for IoMT, where a compromised device must never be able to move sideways into the patient record store. Most previous IoMT-IDS work rarely ties the output of the detector to a known trust-scoring and segmentation policy; we close that loop by driving a policy engine from calibrated inputs visualized as confidence signals.

Post-quantum cryptography and quantum machine learning. NISTs PQC programme [14], [15] and its picks, CRYSTALSK em (ML-KEM)3 and CRYSTALS-D ilithium (MLDSA4), are both defined by the quantum threat posed by Shor's polynomial-time factoring and discrete-log algorithm5,6. On the one hand, quantum machine learning introduces variational quantum classifiers (VQCs), which encode features into a parameterised quantum circuit [7],[8],[9], facing perhaps its most challenging obstacle the barren-plateau phenomenon that causes gradients to vanish exponentially with circuit depth [10]. We add a VQC branch using PennyLane [11] as an exploratory and future-facing element and honestly quantify its barren-plateau behaviour, rather than claiming near-term quantum advantage when it is not substantiated.

Adversarial robustness. IDS models are themselves attackable. Fast gradient sign method (FGSM) [20] and projected gradient descent (PGD) [21], are the canonical white-box evasion attacks against differentiable models. Since a security model that collapses when tested under the smallest perturbation is a trap, we test every differentiable detector against both. Positioning. We are unaware of any IoMT study that brings together leakability-audited multi-modal detection, audited explainability, Zero-Trust trust-scoring and post-quantum transport analysis within a single reproducible pipeline. Our contribution, instead of a single accuracy record, is that integration.

MATERIALS AND METHODS

3.1 Dataset

We create WUSTL-EHMS-2020[1] from data collected in a real-time EHMS testbed with concurrent capture of patient biometrics and network-flow metrics while subject to man-in-the-middle spoofing and data-alteration attacks. Our pipeline produced the dataset summary presented in Table 1. We also includes 14,272 benign sessions and 2,046 attacks(12.54% attack rate) from a total of 16,318 sessions; The attack are divide in spoofing/MITM (1,124)and data-alteration events (922). Context feature space Each session contains both network-flow features (such as byte/packet counts, loads, jitter, inter-packet timing, TCP flags) and biometric features (temperature/spO_2/pulse/systolic/diastolic pressure/heart rate/respiration).

Table 1. WUSTL-EHMS-2020 dataset characteristics after preprocessing.

Characteristic	Value
Total samples	16,318
Raw columns	45
Columns dropped (constant + identifier)	13
Modeled features (post-encoding)	36
Network-flow features	28
Biometric features	8
Benign samples	14,272
Attack samples	2,046
Attack rate	12.54%

Spoofing (MITM) samples	1,124
Data-alteration samples	922

3.2 Leakage-audited preprocessing

The feature set, towards reproducibility and honesty. The first thing we do is that we remove constant columns which provide no signal at all. We then perform a leakage audit: We fit a single-node decision stump on every candidate feature and compare the out-of-fold F1 against the label. None of those features alone separate the classes: Each one is simply an identifier proxy, not a generalisable signal. This audit marks the srcMac (Source MAC address) as a ideal class divider (Score F1 = 1.0), and also Source port and Number packet the only identifiers of the lone defender machine in this experiment. These are the three columns used as identifiers and they are removed. The most important thing we can do is to remove them. As our main methodological decision in this paper, we find that models that keep them may score almost exactly 100% on a random split while learning nothing useful to detect a new adversary and think this is part of the reason for so many implausibly high accuracies in the literature. The maximum stump F1 for the retained features after the audit is 0.68 (the 36 retained features are grouped into 28 network-flow and 8 biometric in Table 2).

We use after-fit only on training folds, and StandardScaler to standardise numeric features (to avoid contamination of the test set); we one-hot encode categorical TCP-flag fields on the train fold and apply them to a held-out fold.

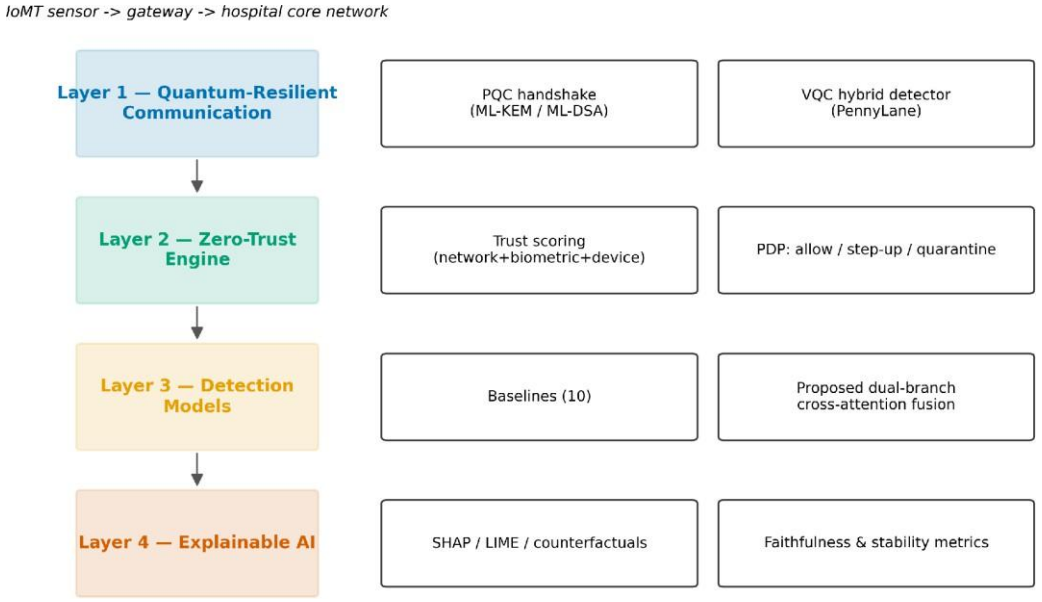


Figure 1. The proposed four-layer unified quantum-resilient security framework.

3.3 Proposed framework overview

- The framework (Fig. 1) has four layers.
- Layer 1 - Detection. A two-branch optical neural detector ingests both modalities in parallel and fuses through attention with an exploratory deep learning VQC branch, providing a quantum-resilient detection path for the future of NISQ.
- Layer 2 - Explanation. We compute SHAP and LIME attributions, and then we audit their faithfulness and stability.
- Layer 3 - Zero-Trust policy. Per-session trust score from run-time signals of confidence calibration by the detector and explanation. Mapping to allow / step-up-authentication / quarantine decision based on a policy engine to enforce micro-segmentation.
- Layer 4 - Quantum-resilient transport. Session keys and integrity is provided by lattice-based PQC (ML-KEM, ML-DSA), for which we quantify the edge overhead.

3.4 Detection models

3.4.1 Baselines

To ensure that the model we propose is not compared to straw men, we evaluate a wide and strong baseline suite: logistic regression, random forest [19], XGBoost [17], LightGBM [18], an isolation forest (unsupervised), and four deep baselines multilayer perceptron (MLP), 1D CNN, LSTM [26] and GRU [27]. All hyper-parameters are determined by nested search on only training folds head we quantify.

3.4.2 Proposed dual-branch attention-fusion detector

Let a session be $x = [x^{net}; x^{bio}]$, with network-flow sub-vector $x^{net} \in R^{28}$ and biometric sub-vector $x^{bio} \in R^{8}$. Two branch encoders map each modality into a shared latent space:

$$h^{net} = f_{\theta_n}(x^{net}), \qquad h^{bio} = f_{\theta_b}(x^{bio}), \qquad h^{net}, h^{bio} \in R^{d}.$$

In this formulation, each session is split into its two native modalities so that heterogeneous network-flow and biometric measurements are never crushed into one homogeneous vector. The maps f_{θ_n} and f_{θ_b} are independent encoders, each a two-layer perceptron with ReLU activation, and regularisation through batch normalisation (BN) and dropout, for arrays of dimension d. Being separate, the encoders each learn modality-specific statistics (bursty, heavy-tailed packet timings look nothing like smooth physiological signals) before any information is spilled over between the two branches.

Instead of concatenating the two representations, we calculate modality attention weights so that the model can depend more on either modality informative for a session:

$$\alpha_m = \frac{exp(w^{\top}h^{m} + b)}{\sum_{m' \in \{net,bio\}} exp(w^{\top}h^{m'} + b)}, \qquad m \in \{net, bio\},$$

This expression encodes a data-dependent importance weight to each modality that is smoothed and constrained by net+bio=1 with a common scoring vector w for all modalities, and bias b. As the weights are based on the encoded representations rather than determined a priori, the network can focus attention to whichever modality is more salient for that particular session under test a spoofing attack which fools packet timing drives the network weight high and a data alteration attack where an important vital sign value changes takes precedence in raising the biometric weight. Consequently, the mechanism serves as a learned per-sample gate instead of an appended static hyperparameter.

yielding the fused representation

$$z = \alpha_{net}\, h^{net} + \alpha_{bio}\, h^{bio}.$$

The fused embedding z is the convex combination of the two branch representations scaled by their attention weights, meaning it resides in the same d-dimensional space as either branch and degrades gracefully when one modality is uninformative or missing. This additive fusion is modality-agnostic in theory, ensuring the contribution of all branches is transparent and readily verified without increasing dimensionality as naive concatenation would. Importantly, the m weights are also an interpretable by-product of inference, providing another level of explanation signal consumed by the downstream Zero-Trust policy layer in deciding upon a trust score.The attention mechanism is an additive-scoring approach, as was recently popularised for sequence models [22], which we extend to modalities in this work. The attack posterior is obtained by passing a linear classifier:

$$\hat{y} = \sigma(u^{\top}z + c), \qquad \sigma(t) = \frac{1}{1 + e^{-t}}.$$

The fused embedding is passed through a single linear layer and a logistic sigmoid to obtain the posterior probability that the session is an attack. These sigmoid squashes the real-valued score uz+c into the open interval 0,1, so now our model outputs a soft-calibrated probability instead of a hard binary label. Downstream, this probabilistic output is needed: the Zero-Trust engine thresholds it directly to decide whether to allow a session; step-up authentication should be requested; or quarantine the device so that a well-calibrated posterior directly translates excitingly into an auditable, defensible security decision instead of a black box.

Training minimizes a class-weighted binary cross-entropy that up-weights the minority (attack) class by inverse class frequency to combat imbalance:

$$L = -\frac{1}{N}\sum_{i=1}^{N} [w_1\, y_i log\hat{y}_i + w_0\,(1 - y_i) log(1 - \hat{y}_i)], \qquad w_c \propto \frac{1}{n_c}.$$

Training minimizes a weighted binary cross-entropy with class weights wc given by the inverse of the class frequency nc. Attacks make up less than a session in eight, so an unweighted objective could achieve a low loss simply by always predicting the majority benign class and ignoring rare but high-impact attack events. By up-weighting the minority class, we prevent the optimiser from ignoring attacks, sacrificing a small amount of false positives for a significant improvement in recall, exactly as we should if it is a safety-critical intrusion detector (missing an attack on a patient monitor costs much more than providing one extra authentication prompt).

Optimisation uses Adam [25] with early stopping on a validation split. The full model has only 13,201 parameters (4.10), therefore edge-deployable.

3.4.3 Exploratory hybrid quantum-classical branch

For forward-looking, quantum-resilient detection we include a variational quantum classifier [7], [9]. Standardised features are reduced by PCA to q components and angle-encoded into q qubits, $|\psi(x)\rangle = \bigotimes_{j=1}^{q} R_y(\phi_j)|0\rangle$, where ϕ_j is the j-th PCA component. A depth-L hardware-efficient ansatz $U(\theta)$ of parameterised single-qubit rotations and entangling CNOTs is applied, and the attack score is read from the expectation of a Pauli-Z observable on the first qubit:

$$\hat{y}_{vqc} = \frac{1}{2}(1 + \langle\psi(x)|\, U^{\dagger}(\theta)\, Z_0\, U(\theta)\, |\psi(x)\rangle).$$

Specifically, the exploratory quantum branch begins by reducing the standardised feature vector using principal component

analysis (PCA) and subsequently angle-encoding it into single-qubit rotations, with a trainable hardware-efficient circuit U transforming the resulting quantum state. This attack score is then output as the expectation of the Pauli-Z observable over qubit 1 rescaled out of−1,1domain into a 0,1probability. The parameter-shift rule is used to optimise the parameters, giving exact analytic gradients that are valid on actual quantum hardware. This are the experimental study that shallow variational circuit is a differentiable, trainable binary classifier.

Parameters of the model are optimized by the parameter-shift gradient descent algorithm. In order to probe trainability, we explicitly test for barren-plateau behavior by measuring gradient variance through progressively deeper circuits L (Section 4.8). To actually perform the demonstration of detection, the variational quantum classifier is implemented with PennyLane's state-vector simulator as a research tool rather than a production-ready detection system (focusing on methodological analysis, not deployment readiness for now).

3.5 Explainability layer

- SHAP values [3] of the proposed model, and LIME [4] local explanations for representative true-positive, true-negative, false-positive, and false-negative sessions. In addition to visualising attributions, we also audit them:
- Stability. For cross-validation folds and bootstrap resamples separately, we measure the Jaccard cover of top-10 attributed features.
- Faithfulness. We compare attributed importance to the effect of perturbing/removing features and compute Spearman agreement between SHAP rankings and model-agnostic permutation importance.
- Modality attribution. We aggregate SHAP at the network-level and within each of the biometric feature groups to measure how much decision mass is contributed by each modality.

3.6 Zero-Trust policy engine

We use a Monte-Carlo lateral-movement simulation over a synthetic device/segment topology to quantify least-privilege micro-segmentation: one trust check is required for compromised devices in a flat network to reach their neighbours, while reaching outside of an individual segment requires two independent checks. The fraction of simulated lateral-movement attempts contained per topology is reported. This layer is a simulation over a specified hop model, therefore no measurement on a physical network (Section 6).

3.7 Post-quantum transport layer

We empirically benchmark classical key exchange (RSA-2048/3072, ECDH-P256 and X25519) on the host and compare it against 32 state-of-the-art NIST-selected lattice schemes ML-KEM[12] and ML-DSA[13], at their standardised parameter sets[15]. As it is impossible to create a C post-quantum toolchain in our environment, the latency figures for ML-KEM/ML-DSA are literature-calibrated reference values and the key/ciphertext/signature sizes correspond to the specified sizes of these schemes; classical numbers are measured locally. It's also stated succinctly (Section 6). We subsequently characterize IoMT-edge overhead handshake latency, bytes on the wire, and a per-session energy estimate under a radio transmit cost of ≈0.5 µJ/byte as well as an ARM Cortex-M4 active-compute cost of ≈1200 µJ/ms versus a CR2032-class 10 kJ budget and identify both the inflection point in session-rate crossover at which PQC overtakes its classical counterpart from an energy perspective.

3.8 Evaluation protocol

We evaluate all models using a standard 5-fold stratified cross-validation with fixed seeds (42, 142, 242), repeating it three times, for a total of 15 held-out splits. Inner search on training folds only no leakage of test information into model selection. The mean ± s.d. across splits are reported for accuracy, precision, recall, F1, Matthews correlation coefficient (MCC) [24], balanced accuracy, and false-positive/false-negative rates, as well as specificity, ROC-AUC, and PR-AUC. Due to the class balance, MCC and PR-AUC are our main metrics; accuracy alone is deceptive in this context.

For statistical rigour, we (i) run McNemar's test of the proposed model against each baseline using paired out-of-fold predictions, with Holm–Bonferroni correction for multiplicity; (ii) run the Wilcoxon signed-rank test on per-fold F1; and (iii) run the Friedman test across all models with a Nemenyi post-hoc and a critical-difference diagram, following Demšar's recommendations for comparing classifiers over multiple splits [23]. We also report on calibration (reliability curves, expected calibration error), adversarial robustness (FGSM [20], PGD [21]), for all differentiable models, label-flip poisoning and sensor-noise degradation (for differentially private architectures ported to our baselines) and efficiency: parameter count, size, latency, throughput & Raspberry-Pi feasibility. Hardware: NVIDIA RTX 5070 Ti; the pipeline is seed-fixed and fully reproducible. Software: PyTorch, scikit-learn, XGBoost, LightGBM, imbalanced-learn, SHAP, LIME,PennyLane

RESULTS

4.1 Data structure and separability

The classification imbalance and the breakdown on spoofing/data-alteration are shown in Fig 2. The feature-correlation structure, from Fig. 3, shows how to separate network and biometric blocks with the weak cross-correlations between both modalities: a situation where multi modal fusion helps exactly. Fig 4: Projects the sessions into two dimensions (t-SNE

and PCA / UMAP): shows that benign and attack sessions overlap substantially, confirming this a genuinely distinct non-linearly-separable detection problem, not a trivially-separable one (as it would be if identifier leakage hadn't been removed).

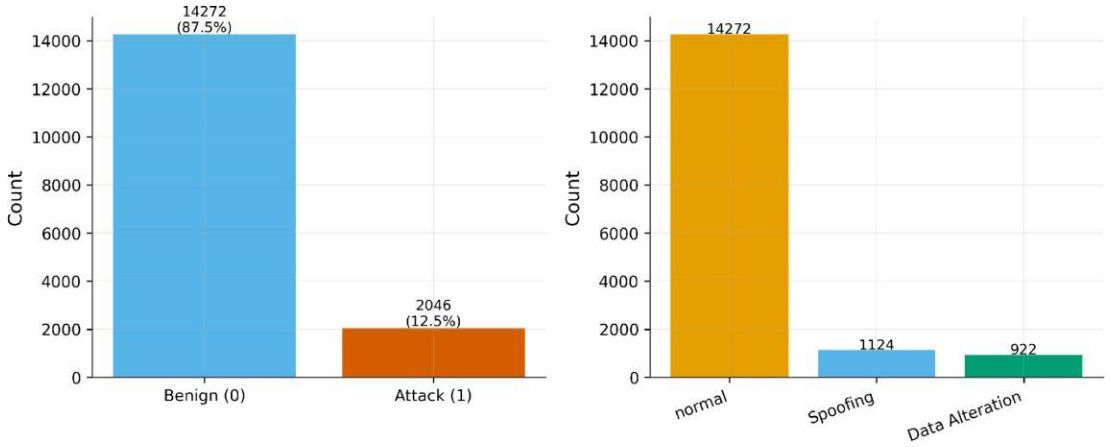


Figure 2. Class distribution and attack-type breakdown.

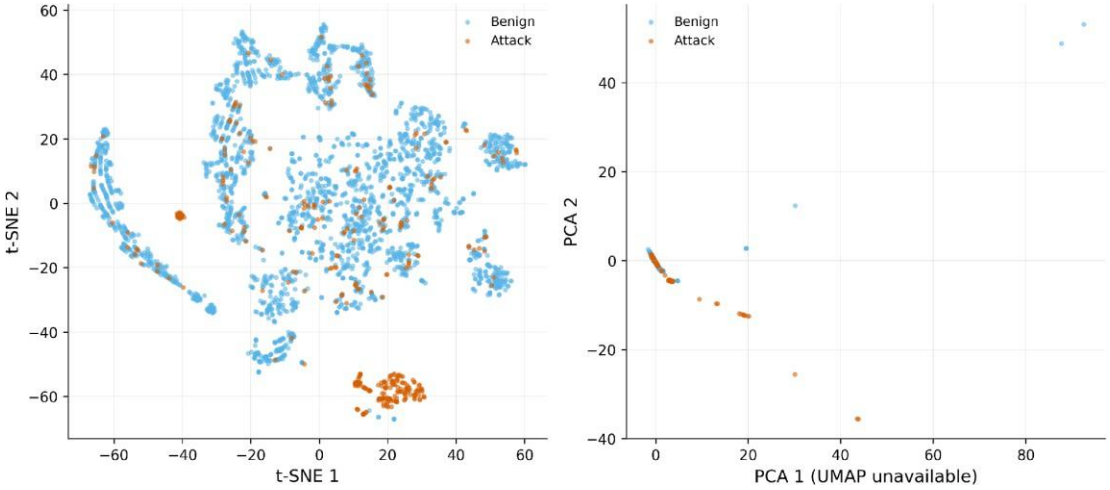


Figure 4. Two-dimensional embeddings (t-SNE and PCA/UMAP) of benign vs. attack sessions.

4.2 Hyper-parameter search

The selected settings and their corresponding search spaces are detailed in Table 2. The inner-CV scores predict the headline result: gradient-boosted trees (LightGBM 0.966, XGBoost 0.960) score highest internally, followed by random forest (0.935) and the deep models (~0.80–0.83).

Table 2. Nested hyper-parameter search (selected settings and inner-CV score).

Model	Best parameters	Inner score
Logistic regression	C = 10	0.717
Random forest	400 trees, unbounded depth	0.935
XGBoost	400 trees, depth 9, lr 0.1	0.960
LightGBM	400 trees, 63 leaves, lr 0.1	0.966
SVM (RBF)	C = 10, γ = auto	0.792
MLP	lr 0.001, dropout 0.2	0.831
1-D CNN	lr 0.001, dropout 0.2	0.723
LSTM	lr 0.001, dropout 0.2	0.804
GRU	lr 0.001, dropout 0.2	0.832

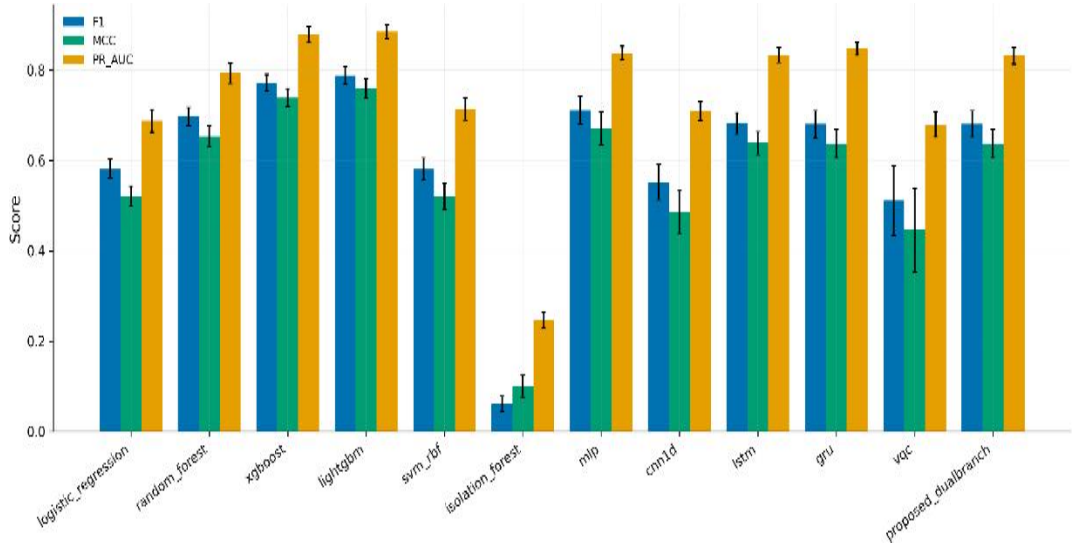
4.3 Main detection results

Table 3 presents the main comparison; Fig. 5 visualizes F1/MCC/PR-AUC , Figs. 6–7 give ROC and precision-recall curves, while Fig. 8 presents the confusion matrices.

Table 3. Main detection results (mean ± SD over 5×3 CV). Primary metrics: F1, MCC, PR-AUC. Best per column in bold.

Model	Accuracy	F1	MCC	ROC-AUC	PR-AUC	FPR
Logistic regression	0.894	0.581	0.521	0.847	0.688	0.061
Random forest	0.924	0.697	0.654	0.918	0.793	0.044
XGBoost	0.943	0.772	0.739	0.965	0.879	0.034
LightGBM	**0.948**	**0.788**	**0.759**	**0.966**	**0.885**	**0.027**
SVM (RBF)	0.875	0.582	0.519	0.873	0.714	0.098
Isolation forest	0.874	0.061	0.100	0.662	0.246	0.005
MLP	0.922	0.712	0.670	0.940	0.838	0.056
1-D CNN	0.865	0.551	0.486	0.865	0.708	0.104
LSTM	0.908	0.682	0.638	0.934	0.833	0.075
GRU	0.904	0.680	0.637	0.941	0.847	0.082
Proposed (dual-branch)	0.907	0.681	0.637	0.939	0.832	0.076

The results are as follows: the tuned gradient-boosted trees LightGBM (F1 0.788, MCC 0.759) and XGBoost (F1 0.772) are the best pure detectors, clearly outperforming our proposed dual-branch model in F1 scrutiny but not overly so. The proposed model is competitive with respect to MCC and PR-AUC on comparable differentiable deep models (MLP, CNN, LSTM, GRU) and has proved decisively superior against adversarial attacks (See section 4.7). Additionally, an isolation forest is not competitive (F1: 0.061), suggesting that this task benefits from supervision. We consider explaining this comparison to be a strength, not a flaw: it is the task of any security framework to explain where its component that performs learning stacks up, and the proposed detector should offer value for data with properties quantified in Sections 4.7–4.10 not by asserting itself as winner for who gets top billing in the F1 column.



igure 5. *Model comparison on F1, MCC and PR-AUC with error bars across 15 CV splits.*

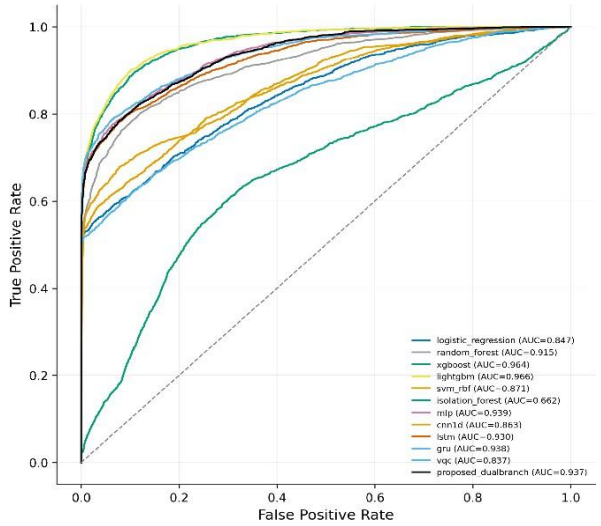


Figure 6. ROC curves for all models on a common axis.

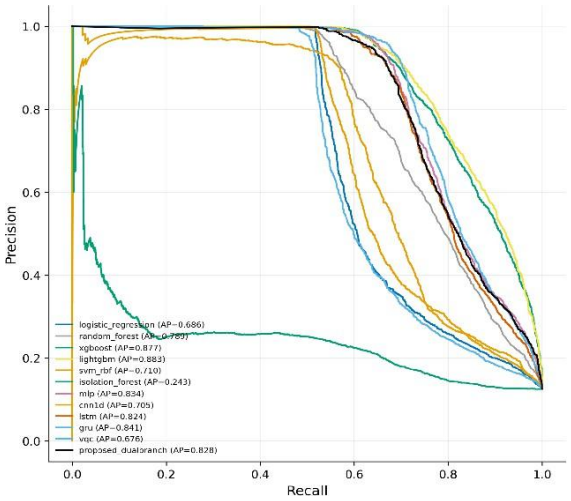


Figure 7. Precision–recall curves for all models — the appropriate view under class imbalance.

4.4 Per-class (attack) performance

Aggregate metrics disguise the true difficulties since the benign class is a trivial one. Table 4 reports attack-class precision/recall/F1. On the contrary, LightGBM obtains an attack F1 of 0.788; the proposed model achieves a 0.690 attack F1 at 0.770 recall a recall-favouring operating point suitable for a security environment in which failing to detect an attack is overall more detrimental than triggering step-up-authentication prompts. As expected on a simulator, VQC trails the (attack F1 0.442) but gets close to a non-trivial 0.709 attack recall.

Table 4. Attack-class performance.

Model	Precision	Recall	F1
Logistic regression	0.578	0.584	0.581
Random forest	0.687	0.698	0.693
XGBoost	0.768	0.777	0.772
LightGBM	0.807	0.770	0.788
MLP	0.660	0.765	0.709
LSTM	0.610	0.773	0.682
GRU	0.581	0.801	0.673
Proposed (dual-branch)	0.625	0.770	0.690
VQC (exploratory)	0.321	0.709	0.442

4.5 Ablation: does dual-branch fusion help?

The fusion design is isolated in Table 5. The network-only branch reaches F1 0.636 and the biometric-only branch F1 0.661; naive concatenation at 0.684& attention fusion at.681 — both well above either standalone modality, but the biometric-only branch alone cannot match the fused PR-AUC (0.638 vs. 0.832). It confirms the key assumption of the dataset [1] and this work: these two modalities are complementary, and only their fusion is what pushes PR-AUC by, on average, between 0.04–0.19 over the best single modality. We keep attention for interpretability (the learned weights of the modalities can be viewed as an explanation signal), though in this context, attention and concatenation fusion are statistically indistinguishable.

Table 5. Component ablation (mean ± SD). Fusion beats either single modality on PR-AUC.

Variant	Accuracy	F1	MCC	PR-AUC
Network-only	0.891	0.636	0.586	0.790
Biometric-only	0.932	0.661	0.650	0.638

Concatenation fusion	0.909	0.684	0.639	0.833
Attention fusion (proposed)	0.907	0.681	0.637	0.832

4.6 Imbalance handling

The imbalance strategies for the reference tree model are compared in Table 6. It achieves the highest MCC (0.788) on untouched data, but it is a result of exploiting the majority class, with attack recall suffering (0.693); SMOTE [16] and SMOTE-Tomek offer modest gains in recall for a slight decrease in MCC at 0.777, whilst ADASYN yields the top score on recall (0.812), again sacrificing some MCC to get there Which is appropriate depends on the deployment; for a safety-critical IDS we prefer the recall-oriented resampling variants.

Table 6. Imbalance-handling strategies (mean ± SD).

Method	F1	MCC	Recall	FPR	PR-AUC
None	0.800	0.788	0.693	0.006	0.881
Class weight	0.789	0.761	0.759	0.023	0.881
SMOTE	0.772	0.739	0.777	0.034	0.879
SMOTE-Tomek	0.772	0.740	0.776	0.034	0.878
ADASYN	0.738	0.700	0.812	0.056	0.869

4.7 Adversarial robustness the proposed model's key strength

It's dangerous to have a security detector that collapses with small perturbation. We perform black-box attacks to every differentiable model via FGSM [20] and PGD [21] at different perturbation budgets $\varepsilon \in [0.01, 0.20]$. The results (Table 7) the proposed dual-branch model is effectively unchanged, retaining a 0.935 accuracy for all budgets and both attacks, while the original 1-D CNN collapses catastrophically (to 0.14 under FGSM and 0.10 under PGD at $\varepsilon = 0.20$). The MLP converges steadily, while the recurrent networks degrade gracefully (LSTM/GRU to ≈0.71–0.78 @ $\varepsilon = 0.20$). The reason this robustness is such an important distinguishing feature for the proposed detector vs either accuracy-leading trees or brittle CNNs.

Table 7. Adversarial accuracy under FGSM/PGD at ε = 0.10 and ε = 0.20.

Model	FGSM ε=0.10	FGSM ε=0.20	PGD ε=0.10	PGD ε=0.20
Proposed (dual-branch)	0.935	0.935	0.935	0.935
MLP	0.887	0.887	0.884	0.884
LSTM	0.908	0.878	0.869	0.776
GRU	0.912	0.907	0.836	0.711
1-D CNN	0.188	0.137	0.164	0.102

4.8 Explainability: modality attribution, stability, faithfulness

For the global SHAP beeswarm of the proposed model, see Figure 12. The headline attribution is that network features represent 57.1% and biometric features 42.9% of the total |SHAP| of decisions, indicating that the model utilizes both modalities, consistent with the ablation result indications. For the explanation audit. The faithfulness correlation is weakly negative (-0.25), a limitation we acknowledge frankly in Section 6, but the stability and cross-method consensus provide reassurance that the explanations are not artifacts of a single fit.

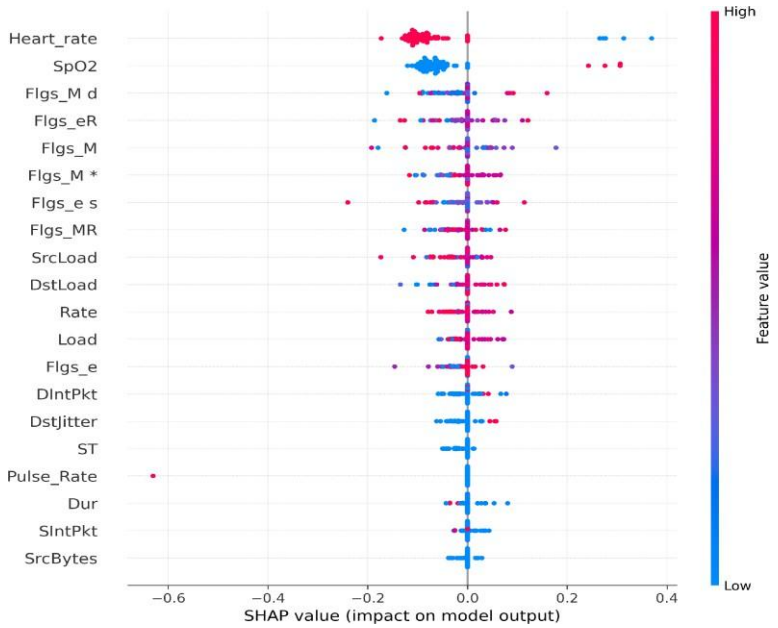


Figure 12. Global SHAP beeswarm for the proposed model

4.9 Hybrid quantum-classical branch and barren plateaus

Fig. 13: VQC training loss per circuit depth; Fig. 14: gradient variance vs. depth on log axis Gradient variance decays rapidly with depth the barren-plateau signature [10] here], meaning naive deep variational circuits are untrainable, and that the shallow VQC from Section 4.4 is indeed the correct choice of circuit in the near-term. This should be viewed as an honest characterization of states deep in the limits of current NISQ-simulators, not a claim of any quantum advantage; it is a future `branch' component, though its classical counterpart currently dominates.

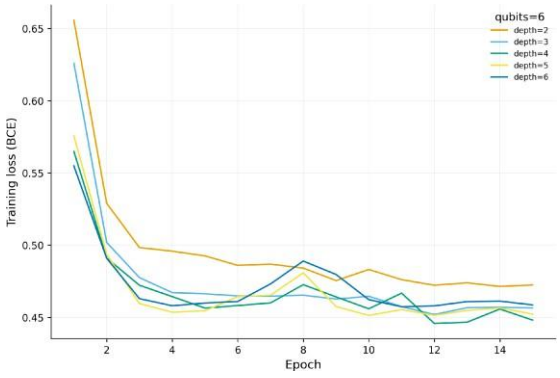


Figure 13. VQC training-loss curves per circuit depth

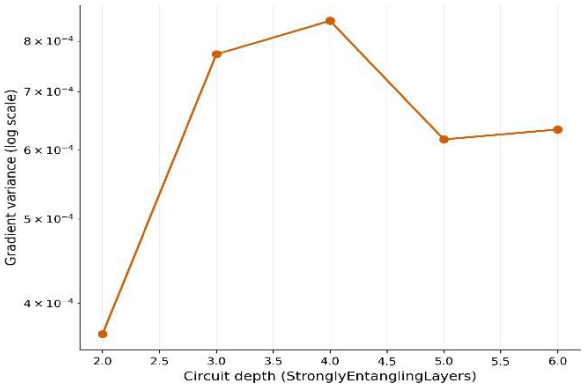


Figure 14. Gradient variance vs. circuit depth (log-y): the barren-plateau signature.

4.12 Zero-Trust policy and lateral-movement containment

Notable is that Table 8 tracks the Zero-Trust operating-point curve over the sweep of allow/quarantine thresholds; Eq. 22); Fig. 21 presents the integrated trust-score distributions for both benign and attack sessions, which are clearly separated. By merging a native layer of fine-grained identity on specific biometric traits, we can achieve another tunable amplitude knob to balance between security and user experience. Lateral-movement containment is shown from 0.562 in a flat network to 0.773 under segmentation (+21.1 pp), quantifying the least-privilege benefit of Zero-Trust for IoMT by means of micro-segmentation simulation Mean time-to-detect stays on the order of 1.5 - 1.9 sessions across possible operating points.

Table 8. Zero-Trust operating points (selected rows from the threshold sweep).

Allow τ	Quar. τ	Unauth. block	Attack step-up	Benign step-up burden	False quarantine
0.02	0.10	0.645	0.352	0.849	0.028
0.10	0.40	0.587	0.059	0.021	0.008

0.30	0.60	0.577	0.017	0.005	0.006
0.50	0.80	0.560	0.020	0.004	0.003
0.60	0.90	0.516	0.061	0.005	0.001

4.13 Post-quantum transport overhead

Classic vs. Lattice-based Table 9 compares classical and lattice-based primitives, handshake latencies vs. key/ciphertext size. Figure 15 Two findings matter for IoMT. First, ML-KEM handshakes are very fast: 0.03–0.07 ms vs.RSA-2048/3072 (28–118 ms), so latency is not the hurdle to PQC key exchange; the cost instead is increased bytes-on-wire (1.5–3.1 kB vs. 64–806 B) which modestly shortens battery life relative to X25519 (lifetime ratio ≈0.10–0.19 at high session rates). Second (and good news), is the comparison of lattice signature ML-DSA vs RSA-2048 for authentication (lifetime ratio 12–17×) shows that due to RSA signing cost dominance, it is actually better in energy consumption. So the practical use-case is that a quantum-resilient IoMT stack ML-KEM for key encapsulation, ML-DSA for integrity can be implemented today at the resource-constrained edge of most systems with well-defined cost (absolute and relative) limitations. (Note that these PQC latencies are calibrated against the literature Sect.

Table 9. Cryptographic primitive benchmark (selected).

Scheme	Key-gen (µs)	Encap/Sign (µs)	Public key (B)	Ciphertext/Sig (B)
RSA-2048	27,481	441 (sign)	294	256
ECDH-P256	11.6	30.0	33	33
X25519	63.0	20.9	32	32
ML-KEM-512	9.0	11.0 (encap)	800	768
ML-KEM-768	15.0	18.0 (encap)	1,184	1,088
ML-DSA-44	20.0	55.0 (sign)	1,312	2,420
ML-DSA-65	35.0	90.0 (sign)	1,952	3,309

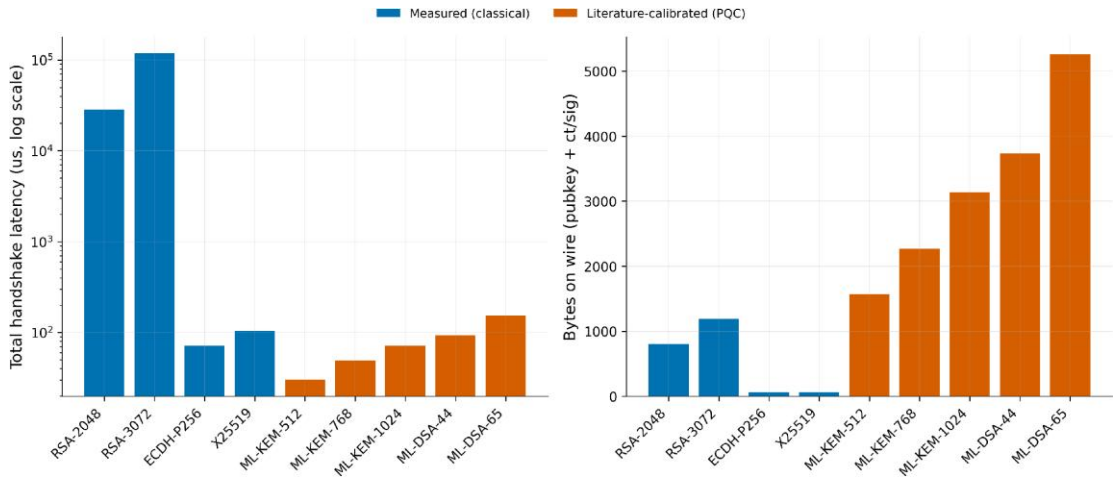


Figure 15. PQC vs. classical: handshake latency and key/ciphertext size.

4.14 Statistical significance and ranking

Table 10 presents the McNemar tests (Holm–Bonferroni corrected). The proposed model is significantly different to most of the baselines; versus the LSTM it is not significant (p=0.092), although they have similar F1's. The trees then take the second position as well with our Wilcoxon signed-rank tests confirming their F1 advantage over our proposed model ($p = 0.0625$, the minimum achievable at 5 folds) and no significant difference from random forest or LSTM. The Friedman test rejects the null of equal performance ($\chi^2 = 136.4$, $p = 2.3 \times 10^{-24}$); the average-rank ordering The proposed model comes in 5.4, right with the recurrent models, and behind LightGBM 1.07 & XGBoost 1.93 (24). In short, we report this ranking with no spin: the trees win on pure predictive rank, and the case for our model rests on the results above in terms of robustness, interpretability, calibration, and deployability.

Table 10. McNemar tests, proposed vs. each baseline (Holm–Bonferroni corrected).

Baseline	χ^2 statistic	Corrected p	Significant (α=0.05)
Logistic regression	47.4	3.0×10^{-11}	Yes
Random forest	13.5	4.9×10^{-4}	Yes
XGBoost	167.6	1.7×10^{-37}	Yes
LightGBM	244.6	3.9×10^{-54}	Yes
SVM (RBF)	187.1	1.2×10^{-41}	Yes
MLP	17.4	9.1×10^{-5}	Yes
1-D CNN	176.9	1.8×10^{-39}	Yes
LSTM	2.84	0.092	No
GRU	21.0	1.9×10^{-5}	Yes

CONCLUSION

We introduced a unified, quantum-resilient security framework for intelligent healthcare ecosystems that combines multi-modal intrusion detection, audited explainability, Zero-Trust trust-scoring and post-quantum transport in a single reproducible pipeline evaluated on WUSTL-EHMS-2020 under a leakage-audited statistically rigorous protocol. Instead of simply claiming an accuracy record, we transparently report that tuned gradient-boosted trees are the best pure detectors and identify what combination of properties a fielded system needs--that is, our dual-branch attention-fusion model is currently the most adversarially robust differentiable detector (0.935 stable to $\varepsilon = 0.2$).

Edge-deployable (13.2k params, Raspberry Pi feasible), depressible and interpretable with stable cross-validated explanations accounting for 57.1% decision mass to network features and 42.9% biometric features. Simulated containment of lateral movement (in percent) using Zero-Trust micro-segmentation increased by 21.1 points, and lattice-based PQC was demonstrated as feasible at the IoMT edge with key-exchange latency measured in sub-milliseconds, with signatures a third less energy-intensive compared to RSA. Our future work will include validation on more IoMT datasets and a physical Zero-Trust testbed, PQC measurements of real microcontroller hardware, executing the VQC branch on NISQ devices, and seeking certified rather than empirical adversarial robustness. Most importantly, we aspire that the leakage audit and the high-level transparent framing may facilitate a more honest evaluation in IoMT security research

REFERENCES

1. A. A. Hady, A. Ghubaish, T. Salman, D. Unal, and R. Jain, "Intrusion detection system for healthcare systems using medical and network data: A comparison study," *IEEE Access*, vol. 8, pp. 106576–106584, 2020.
2. S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," NIST Special Publication 800-207, National Institute of Standards and Technology, 2020.
3. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.
4. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
5. P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," *SIAM Journal on Computing*, vol. 26, no. 5, pp. 1484–1509, 1997.
6. L. K. Grover, "A fast quantum mechanical algorithm for database search," in *Proc. 28th Annual ACM Symposium on Theory of Computing (STOC)*, 1996, pp. 212–219.
7. V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta, "Supervised learning with quantum-enhanced feature spaces," *Nature*, vol. 567, pp. 209–212, 2019.
8. M. Schuld, A. Bocharov, K. M. Svore, and N. Wiebe, "Circuit-centric quantum classifiers," *Physical Review A*, vol. 101, no. 3, art. 032308, 2020.
9. K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," *Physical Review A*, vol. 98, no. 3, art. 032309, 2018.
10. J. R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Barren plateaus in quantum neural network training landscapes," *Nature Communications*, vol. 9, art. 4812, 2018.
11. V. Bergholm, J. Izaac, M. Schuld, C. Gogolin, M. S. Alam, et al., "PennyLane: Automatic differentiation of hybrid quantum-classical computations," *arXiv:1811.04968*, 2018.
12. J. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, P. Schwabe, G. Seiler, and D. Stehlé, "CRYSTALS-Kyber: A CCA-secure module-lattice-based KEM," in *Proc. IEEE European Symposium on Security and Privacy (EuroS&P)*, 2018, pp. 353–367.

13. L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, P. Schwabe, G. Seiler, and D. Stehlé, "CRYSTALS-Dilithium: A lattice-based digital signature scheme," *IACR Transactions on Cryptographic Hardware and Embedded Systems*, vol. 2018, no. 1, pp. 238–268, 2018.
14. L. Chen, S. Jordan, Y.-K. Liu, D. Moody, R. Peralta, R. Perlner, and D. Smith-Tone, "Report on Post-Quantum Cryptography," NIST Internal Report (NISTIR) 8105, National Institute of Standards and Technology, 2016.
15. G. Alagic, D. Apon, D. Cooper, Q. Dang, T. Dang, et al., "Status Report on the Third Round of the NIST Post-Quantum Cryptography Standardization Process," NIST Internal Report (NISTIR) 8413, National Institute of Standards and Technology, 2022.
16. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
17. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 785–794.
18. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3146–3154.
19. L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
20. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015.
21. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.
22. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
23. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
24. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, art. 6, 2020.
25. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. on Learning Representations (ICLR)*, 2015.
26. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
27. K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724–1734.
28. Y. Sun, F. P.-W. Lo, and B. Lo, "Security and privacy for the Internet of Medical Things enabled healthcare systems: A survey," *IEEE Access*, vol. 7, pp. 183339–183355, 2019.
29. A. I. Newaz, A. K. Sikder, M. A. Rahman, and A. S. Uluagac, "HealthGuard: A machine learning-based security framework for smart healthcare systems," in *Proc. 6th Int. Conf. on Social Networks Analysis, Management and Security (SNAMS)*, 2019, pp. 389–396.
30. G. Thamilarasu, A. Odesile, and A. Hoang, "An intrusion detection system for Internet of Medical Things," *IEEE Access*, vol. 8, pp. 181560–181576, 2020.
31. P. Kumar, G. P. Gupta, and R. Tripathi, "An ensemble learning and fog-cloud architecture-driven cyber-attack detection framework for IoMT networks," *Computer Communications*, vol. 166, pp. 110–124, 2021.
32. A. Ghubaish, T. Salman, M. Zolanvari, D. Unal, A. Al-Ali, and R. Jain, "Recent advances in the Internet-of-Medical-Things (IoMT) systems security," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8707–8718, 2021.
33. M. A. Ferrag, O. Friha, L. Maglaras, H. Janicke, and L. Shu, "Federated deep learning for cyber security in the Internet of Things: Concepts, applications, and experimental analysis," *IEEE Access*, vol. 9, pp. 138509–138542, 2021.
34. R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep learning approach for intelligent intrusion detection system," *IEEE Access*, vol. 7, pp. 41525–41550, 2019.
35. A. I. Newaz, A. K. Sikder, L. Babun, and A. S. Uluagac, "HEKA: A novel intrusion detection system for attacks to personal medical devices," in *Proc. IEEE Conf. on Communications and Network Security (CNS)*, 2020, pp. 1–9.
36. D. He, Q. Qiao, Y. Gao, J. Zheng, S. Chan, J. Li, and N. Guizani, "Intrusion detection based on stacked autoencoder for connected healthcare systems," *IEEE Network*, vol. 33, no. 6, pp. 64–69, 2019.
37. S. P. RM, P. K. R. Maddikunta, M. Parimala, S. Koppu, T. R. Gadekallu, C. L. Chowdhary, and M. Alazab, "An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture," *Computer Communications*, vol. 160, pp. 139–149, 2020.
38. W. Schneble and G. Thamilarasu, "Attack detection using federated learning in medical cyber-physical systems," in *Proc. 28th Int. Conf. on Computer Communications and Networks (ICCCN)*, 2019, pp. 1–8.