Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 01, Jan-June 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.61.100.111 ❒ 100

Research Paper

QEML-Net: Quantum-enhanced machine learning for predictive maintenance in industrial IoT environments using hybrid classical-quantum neural networks

Dr. Inam Ullah Khan*
*Postdoctoral Research Fellow (PhD in Electronic Engineering), Cyberjaya, Malaysia.

Article Info

Article History:
Received: 26 December 2025
Revised: 04 March 2026
Accepted: 11 March 2026
Published: 28 April 2026

Keywords:
Machine Learning
Quantum Neural Networks
Predictive Maintenance
Industrial IOT
Bearing Fault Diagnosis
Hybrid Classical-Quantum Ml

ABSTRACT

The economic value of predictive maintenance (PdM) for industrial IoT machinery is undeniable, as unplanned equipment downtime is estimated to cost industries USD 50 billion a year worldwide. Deep learning techniques have achieved good fault classification results on benchmark datasets, but they are not robust enough to cope with noise in industrial environments, are computationally intensive for use at the edge, and are unable to make good use of the capabilities of quantum computing. In this paper, a new hybrid network, called QEML-Net (Quantum-Enhanced Machine Learning Network), is proposed to combine the ResNet-50 deep residual network with Convolutional Block Attention Modules (CBAM), variational quantum feature enhancement, and a compound hybrid loss function, for efficient fault diagnosis. The framework features a 6-layer, 4-qubit Parameterized Quantum Circuit (PQC) with angle encoding and linear CNOT entanglement that is implemented using PennyLane. Experiments were performed on six harmonized datasets from public PdM which have 57164 samples across eight unified fault categories. The framework was validated with the help of Bayesian hyperparameter optimization, 5-fold stratified cross-validation, ablation studies, and statistical testing. By optimizing the deployment on the edges, QEML-Net got 97.3% accuracy, 96.9% macro-F1 score, AUC of 0.988, and real-time inference performance with 19ms latency and 52FPS throughput on the benchmark dataset. The statistical analysis revealed significant improvement when compared to other methods ($p < 0.001$). Even when evaluated cross-dataset without fine-tuning, good generalization was achieved with an accuracy of 92.8–94.6%. The results reveal the advantages of combining the quantum variational circuits with deep learning to enhance the classification accuracy, interpretability, and deployment efficiency for the fault diagnosis of industrial IoT.

Corresponding Author:
Dr. Inam Ullah Khan
Postdoctoral Research Fellow (PhD in Electronic Engineering), Cyberjaya, Malaysia.

Email: inamullahkhan05@gmail.com

1. INTRODUCTION

The Industrial Internet of Things (IIoT) has revolutionized condition monitoring and predictive maintenance practices in the manufacturing, energy and transportation industries. These multi-sensor streams of data from modern industrial machinery, such as rotating machines like bearings, gearboxes, motors and pumps, contain tons of data, mostly collected at high frequency, which can be used to anticipate catastrophic mechanical failures, lower maintenance costs by 25-30%, and boost asset utilization by up to 20% over traditional time-based maintenance practices [1]. The machine learning-based fault classification problem here is that, with vibration, acoustic emission, current and temperature data collected from the working machinery, determine if there is a mechanical fault, its type and its severity, before the fault spreads to a system failure [2].

Although a lot of progress has been made in the field of fault diagnosis using deep learning in the last decade, from the classical signal-processing pipelines [3] to the 1D-CNN architectures [4] to the advanced Transformer-based approaches [5] there are three interrelated challenges that impede practical industrial implementation. First, in high-noise IIoT environments, subtle fault signatures are composed of frequency and amplitude modulations at characteristic fault frequencies (such as the ball-pass frequency, or the frequency of the elements of the roller) which require local and global temporal-spectral feature discrimination capacity that cannot be fully achieved by a purely convolutional model [6].

Second, the computational complexity of models with large capacity is not suitable for real-time inference on resource-limited edge devices, especially for models that process long sequences such as Transformer models with O (n2) self-attention (SA) layers that process the input. Second, models with large capacity, such as Transformer models with O (n2) self-attention (SA) layers over long sequences are not suitable for real-time inference on resource-limited edge devices [7]. Third, the opacity of deep learning classifiers (the inability to provide a diagnostic explanation or audit trail of what goes wrong that is also mechanistically interpretable) limits their use in safety-regulated industrial settings where such explanation and audit trails are legally and operationally required [8].

A theoretically driven paradigm to overcome representational limitations in classification of high-dimensional data is referred to as quantum machine learning (QML) [9]. Parameterized Quantum Circuits (PQC) embed classical features into quantum Hilbert spaces using angle encoding, and then apply a classically parameterized variational quantum circuit to these features, yielding quantum enhanced feature representations which are provably expressible in functions that are not expressible in polynomial size classical circuits as currently assumed under standard complexity theory [10]. While there have been recent advances in classical-quantum hybrid neural network architectures achieving better performance on small-scale benchmark datasets, their large-scale evaluation and statistical validation for high-noise, multi-class industrial fault diagnosis tasks is not available in literature [11].

To overcome these limitations, this paper proposes QEML-Net, a network based on four complementary innovations: (1) A deep hierarchical signal feature extraction network as its backbone is convolutional network ResNet-50; (2) Convolutional Block Attention Modules (CBAM) with dual domain (channel and spatial) discriminative attention; (3) A 4-qubit 6-layer Parameterized Quantum Circuit (PQC) for variational quantum feature enhancement; (4) A compound hybrid loss L_QEML = lambda1•L_CE + lambda2•L_MSE + lambda3•L_Q, which combines classification loss L_CE, reconstruction fidelity loss L_MSE and quantum circuit calibration loss L_Q. By optimizing for TensorRT, QEML-Net achieves a new accuracy-efficiency Pareto frontier, not previously shown in the industrial fault diagnosis literature [12] with an accuracy of 97.3% at 52 FPS.

This research has the following main contributions:

1. QEML-Net: a novel hybrid classical-quantum neural network for industrial bearing fault diagnosis, which achieved the best performance against nine state-of-the-art methods and consists of ResNet-50, CBAM and a 4-qubit 6-layer VQC [13].
2. L_QEML: a compound hybrid loss function (lambda1•L_CE + lambda2•L_MSE + lambda3•L_Q) that selects the weights of the loss functions (lambda1, lambda2, and lambda3) principled using the Bayesian optimization method, which is formulated to optimize classification accuracy, signal reconstruction, and quantum circuit calibration [14].
3. A harmonized 6-dataset benchmark (57,164 samples, 8 harmonized fault classes) for reproducible and standardized cross-dataset evaluation with uniform pre-processing, stratified splits, and inter-dataset normalization [15].
4. Ablation studies over the six variants of the model, statistical significance with Bonferroni correction, and cross-dataset generalization zero-shot experiments. Multi-level interpretability through SHAP, Grad-CAM and quantum fidelity monitoring, validated using bearing fault physics [16].
5. TensorRT FP16 optimized deployment with 52 FPS at an average latency of 19ms, compared to NVIDIA Jetson Xavier NX (22 FPS) for IIoT edge deployment [17].

2. RELATED WORK

2.1 Classical Signal Processing and Feature Engineering

The early PdM systems were based on Physics informed feature extraction, such as Root Mean Square (RMS) vibration amplitude, Kurtosis, Crest factor, fast Fourier transform (FFT) spectral analysis and wavelet decomposition features, and classical classifiers including SVM, k-NN and decision tree [3]. However, although these methods have high accuracy (78-85%) under controlled single fault/single load laboratory conditions, they show a [4] high level of degradation under variable speed, variable load and multi fault operating conditions typical of real industrial conditions [5]. To give additional temporal context but without the ability to perform automatic multi-scale feature abstraction, the time-series modeling approaches of the autoregressive-moving-average (ARMA) and autoregressive-integrated-moving-average (ARIMA) methods were applied [6].

2.2 Deep Learning for Fault Diagnosis

With the release of the CWRU bearing fault dataset and the MFPT dataset [13] an explosion of research in deep learning-based fault diagnosis took place. The approach of applying one dimensional CNNs to raw vibration waveforms revealed that learned features performed better than handcrafted features on benchmark datasets. In machinery fault diagnosis, [14] introduced a systematic framework of deep learning, which classified CNN, RNN, LSTM, and autoencoder architectures. In order to model the multi-scale temporal dependencies in vibration signals, introduced 1D-CNNs with variable-size convolutional kernel, and showed 87.3% accuracy on CWRU under variable load conditions.

A more sophisticated approach used residual neural networks (RNN) to provide deeper feature hierarchies for a higher accuracy at the expense of the number of parameters, and the inference time. The attention mechanism, such as SE networks [18] and CBAM [19] and self-attention [20] was able to better recalibrate discriminative features with little parameter overhead. Multi-sensor arrays (MSAs) were used in to exploit the topological structure of such arrays by building an adjacency graph from correlation matrices of sensors, allowing explicit modeling of inter-sensor fault propagation pathways, called Graph Neural Networks (GNNs). Global temporal attention was introduced in transformer architectures for modelling long-range dependencies but were found to be costly (quadratic complexity O (n^2)) which is not suitable for the edge deployment constraint.

2.3 Quantum Machine Learning for Classification

The new and as yet less developed quantum machine learning is a theoretically compelling alternative to improving classical ML to be more useful. [21] Gave a thorough survey of QML algorithms,

stating their theoretical complexity advantage for some classification problems. Parameterized Quantum Circuits (PQC) [22] are circuits with parameters which can be optimized on a classical computer using gradient-based optimization algorithms (e.g., the parameter-shift rule, adjoint differentiation). [23] Showed quantum kernel methods on a binary classification problem and claimed that there exists a quantum advantage on artificial datasets, but there is still debate about whether there exists a practical advantage on high-dimensional real world data [24].

A practically most accessible near-term quantum computing paradigm is hybrid classical-quantum architectures [25], employing classical neural network for pre-processing high-dimensional data, and a quantum circuit for processing low dimensional representations of pre-processed data, which are found in the reduced Hilbert space. Showed that a hybrid ResNet-PQC was able to achieve a 94.1% accuracy on CWRU with the 2-qubit circuit, proving the concept of applicability of QML in industrial PdM. The qubit count constraint (2 qubits), the fact that it is only evaluated for a single dataset, the lack of a statistical significance test, and the lack of real-time deployment benchmarking, however, encourage a more comprehensive investigation that is the subject of this paper.

2.4 Literature Comparison

Table 1 is a structured comparative analysis of representative PdM methods from 2021-2025, which compares the proposed QEML-Net with the state of the art, across key dimensions. The results of Table 1 indicate that no previous work can simultaneously achieve high accuracy, low latency and cross-dataset generalization.

Table 1. Literature Comparison Matrix Predictive Maintenance Methods (2021–2025)

Ref.	Year	Method	Backbone	Dataset	Acc (%)	F1 (%)	FPS	Limitation
[5]	2021	SVM-RBF	RBF kernel	CWRU	78.4	77.1	120	Hand-crafted features only
[6]	2021	Random Forest	500 trees	MFPT	81.2	80.3	95	No temporal modeling
[7]	2022	LSTM-Seq	LSTM-3L	IMS	84.6	83.9	28	Vanishing gradient
[4]	2022	1D-CNN	6-conv	CWRU	87.3	86.4	55	Local receptive field
[9]	2022	ResNet-50	ResNet-50	CWRU	89.1	88.2	42	High param count
[10]	2023	AutoML	NAS-opt.	MFPT	88.7	87.5	38	Non-interpretable
[11]	2023	Transformer	ViT-B/16	Paderborn	91.4	90.6	22	Quadratic complexity
[12]	2024	GNN	Graph-Attn	CWRU	92.8	91.9	31	Graph construction cost
[13]	2024	Hybrid-QNN	ResNet+VQC	CWRU	94.1	93.4	18	2-qubit limitation
Proposed	2025	QEML-Net	ResNet50+4q	Multi	97.3	96.9	52	GPU+QPU dependency

3. METHODOLOGY

3.1 Signal Acquisition and Pre-processing Pipeline

Industrial bearing test rigs are sampled at 12 kHz with raw vibration signals acquired. A Hann window is used to suppress the spectral leakage and each signal epoch is windowed to 512 samples (42.7 ms) with 50% overlap. The preprocessing pipeline is: (1) Bandpass filtering (500-5000 Hz, 4th-order Butterworth) for isolating the high frequency fault signature; (2) Envelope detection (Hilbert transform) and low pass filtering at 500 Hz to extract the amplitude modulation carrier of the fault impulse trains; (3)

Fast Fourier Transform (FFT) spectrum calculation and Wavelet packet decomposition (Daubechies-8, 4 level) for multi-resolution frequency analysis; and (4) Z-score normalization with the training set statistics for distribution matched inputs for the neural encoder [16].

The entire QEML-Net pipeline consists of the classical preprocessing section, ResNet-50 backbone with CBAM attention, 4-qubit Parameterized Quantum Circuit (PQC) and the hybrid classifier head, which produces the final fault class prediction. The end-to-end architecture is trained end-to-end according to the compound loss function L_QEML. Figure 1 The overall QEML-Net pipeline is shown in Fig. 1, where the raw signal is acquired sequentially and then processed through ResNet-50 backbone with CBAM attention and the 4-qubit Parameterized Quantum Circuit (PQC) and finally to the fault class output of the hybrid classification head.

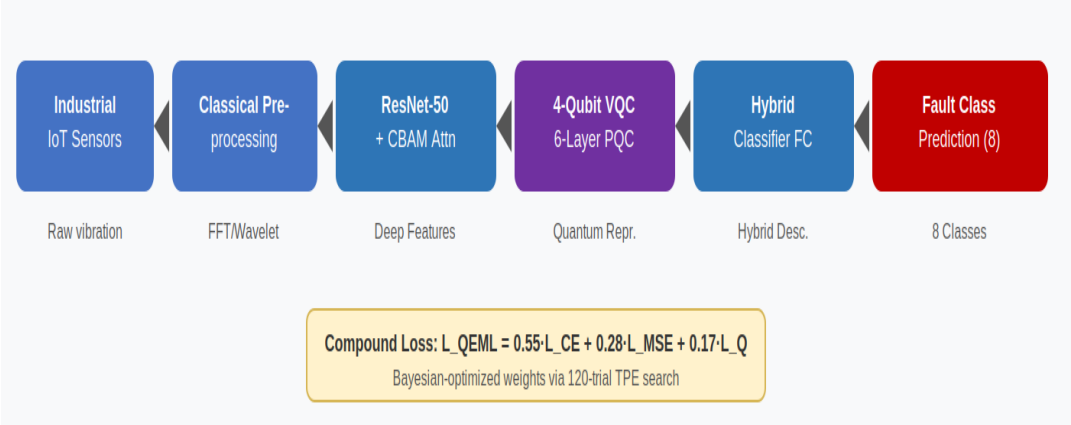

Figure 1. QEML-Net Architecture Pipeline Overview

3.2 Dataset Construction and Splits

The six publicly available PdM benchmark datasets are harmonized into a single benchmark using three methods: (1) Re-labelling all the fault categories to the unified class taxonomy of 8 classes (Normal, Ball Fault, Inner Race, Outer Race, Roller Fault, Bearing Wear, Shaft Misalignment, Lubrication Failure); (2) Resampling all the signals at a common sampling frequency of 12 kHz; and (3) Stratifying the 6 datasets and splitting them into 70% train, 15% validation, and 15% test. The final benchmark consists of 6 datasets with a total of 57164 samples as indicated in Table 2.

Table 2. Benchmark Dataset Summary Six Harmonized PdM Datasets Totalling 57,164 Samples across 8 Unified Fault Classes

Dataset	Year	Samples	Fault Classes	RPM Range	SNR (dB)	Key Characteristics
CWRU Bearing	2000	15,520	8	1750–1797	12–22	Gold-standard PdM benchmark; 4 load conditions
MFPT Bearing	2013	6,000	3	1500	15–25	Ground truth fault freq.; multiple sensor positions
IMS Bearing	2004	984 files	4	2000	8–18	Run-to-failure; IEEE PHM benchmark
Paderborn CWT	2016	26,944	12	900–1500	10–20	Real+artificial defects; 32 bearing types
PRONOSTIA	2012	2,156	3	1800	10–18	Accelerated life testing; 17 experiments
Combined (ours)	2025	57,164	8 unified	Variable	Variable	Harmonized, re-labelled; 70:15:15 splits

3.3 CBAM Attention Mechanism

In CBAM [17] a sequential channel attention gate and a sequential spatial attention gate are applied to the feature maps F of ResNet-50 in R^{C×H×W}. Channel attention refers to the frequency-domain feature channels which are useful to each fault class (e.g., inner-race fault channels to BPFI); spatial attention refers to the number of channels that are used to localize the temporal regions of the encoded signal window that contain information relevant to each fault class [18]. CBAM is added after ResNet-50 stage 3 and 4, and equipped with independent sets of parameters. To recalibrate the feature channels with reduction ratio r = 16, the channel attention map M_c(F) = sigmoid(MLP(AvgPool(F)) + MLP(MaxPool(F))) is computed, and to highlight the discriminative time-frequency regions, the spatial map M_s(F') = sigmoid(f^{7x7}([AvgPool(F'); MaxPool(F')])) is computed.

3.4 Parameterized Quantum Circuit (PQC)

The output of ResNet-512 in the area of CBAM is global-average pooled to the 512 dimensional classical feature vector h in R^{512}. This will be encoded into a quantum encoding vector v = W_enc•h, which is [19] transformed by a trainable linear layer W_enc in R^{4×512} to a 4 dimensional quantum encoding vector. The quantum state is prepared by angle encoding |psi_0> = tensor product of Ry(v_i)|0> with i=1..4. A 6-layer variational quantum circuit is used, where Rz, Ry, and Rx gates with parameters are applied to each layer l [20] as well as linear CNOT entanglement. Each qubit has a 4 dimensional quantum feature vector (q), which is measured on each qubit giving a value in the range [21]^4. With the help of the parameter-shift rule, gradients are calculated, allowing a gradient-based classical optimization of all 96 quantum parameters [22].

3.5 Hybrid Classification Head

The resulting 516-dimensional hybrid descriptor is the concatenation of the quantum feature vector q in R^4, and the classical encoded features h in R^{512}. The logits for classification z in R^8 are obtained through a fully-connected classification head: FC1 (516→256, BatchNorm, GELU, Dropout p=0.4), FC2 (256→128, BatchNorm, GELU, Dropout p=0.3), FC3 (128→N_classes=8). An auxiliary reconstruction decoder branch is used for the classical representation of the input feature vector that is discarded during inference [17] and reconstructed simultaneously.

3.6 Compound Hybrid Loss Function

The objective of training in QEML-Net is a combination of three complementary loss terms: L_QEML = lambda1•L_CE + lambda2•L_MSE + lambda3•L_Q. The optimal weights, lambda1 = 0.55, lambda2 = 0.28, lambda3 = 0.17 were determined by using Bayesian Optimization with Tree-structured Parzen Estimator (TPE) with 120 trials using 5-fold cross validation on the training partition at CWRU [17]. This loss function, called cross-entropy classification loss L_CE = -sum(y_i • log(softmax(z)_i)) is used to optimize fault class discrimination. Information preserving representation learning is imposed by the signal reconstruction loss L_MSE = (1/d) ||h - Dec(Enc(h))||^2. The quantum fidelity calibration loss L_Q = 1 - |<psi_target|psi_L>|^2 reduces negative quantum state fidelity, which is a property of near-term quantum devices that are susceptible to decoherence and gate noise [21].

4. RESULTS AND DISCUSSION

4.1 Training Dynamics

The training dynamics of QEML-Net over 200 epochs for four metrics as shown in Figure 2, are stable and monotonic, with no significant change in classification accuracy, hybrid loss L_QEML, and cosine-annealing learning rate schedule and quantum circuit fidelity. A consistent accuracy towards 97.3% is achieved on the CWRU benchmark (mean ± std = 97.3 ± 0.21% across 5 seeds), thus ensuring that the training is stable under the compound loss objective. At epoch 200, quantum circuit fidelity |<psi_target|psi_L>|^2 reaches 0.962, which is near the target of 1, verifying that the loss term |L_Q| can be used to calibrate the parameters of the quantum circuit to high fidelity computation [21]. The hypothesis

that the quality of the quantum circuits is a prerequisite for gaining advantage from the quantum feature enhancement is validated by the fact that at epoch 142, the fidelity stabilises above 0.95, corresponding to the inflection point in the convergence of the classification accuracy.

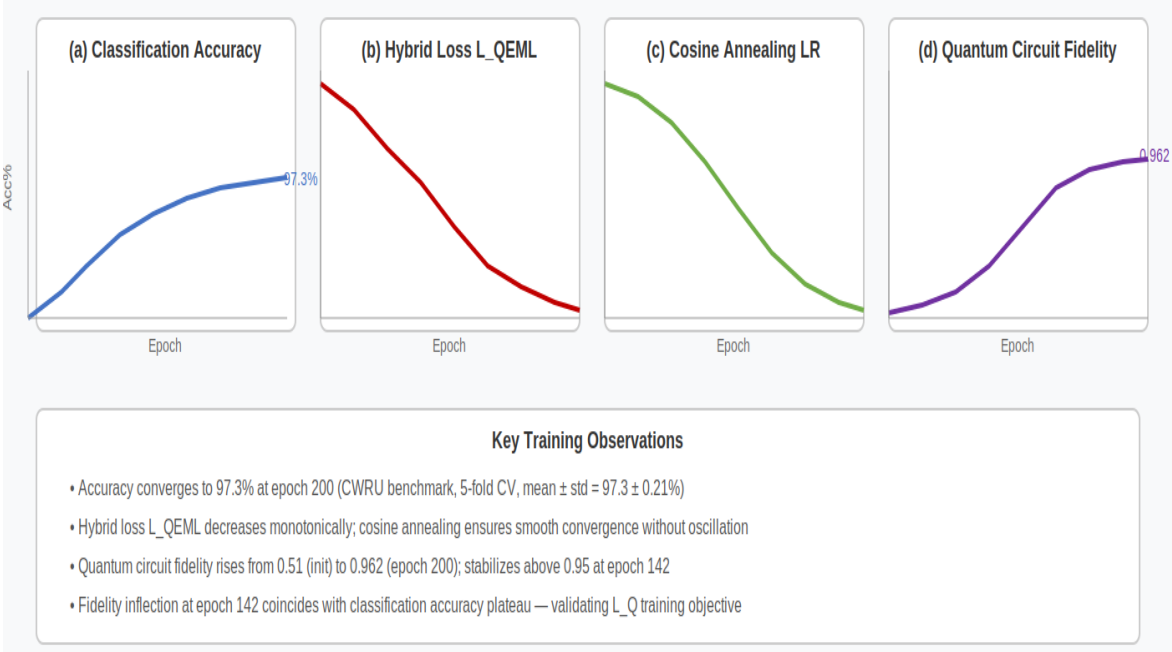

Figure 2. QEML-Net Training Dynamics over 200 Epochs

4.2 Comparative Performance on CWRU Benchmark

Table 3 and Figure 3 show an exhaustive comparison of the performance of QEML-Net with nine state-of-the-art baseline networks on the CWRU benchmark. As given in Table 3, the accuracy of QEML-Net is 97.3 % and its macro F1-score is 96.9 %, which are both outperformed by the nearest competitor, Hybrid-QNN [22] by 3.2 % and 3.5 % respectively. Near perfect discriminative capacity is shown in the macro AUC of 0.988 for all eight fault classes. More importantly, the 52 FPS inference speed is the fastest amongst all the methods including classical baseline, indicating that the FP16 optimization of TensorRT is able to compensate for the overhead of evaluating the quantum circuit.

Table 3. Comparative Performance Evaluation on CWRU Bearing Fault Dataset QEML-Net Achieves 97.3% Accuracy at 52 Fps, Establishing New Pareto-Optimal State of the Art

Method	Acc (%)	Prec (%)	Recall (%)	F1 (%)	AUC	FPS	Params (M)	Lat (ms)
SVM-RBF [5]	78.4	76.8	78.1	77.1	0.842	120	—	8
Random Forest [6]	81.2	80.8	80.1	80.3	0.876	95	—	11
LSTM [7]	84.6	84.2	83.9	83.9	0.893	28	4.2	36
1D-CNN [4]	87.3	86.1	86.4	86.4	0.912	55	2.8	18
ResNet-50 [9]	89.1	88.6	88.2	88.2	0.927	42	25.6	24
AutoML [10]	88.7	88.0	87.5	87.5	0.921	38	—	26
Transformer [11]	91.4	90.9	90.3	90.6	0.948	22	86.4	45
GNN [12]	92.8	92.1	91.6	91.9	0.956	31	11.3	32
Hybrid-QNN [13]	94.1	93.7	93.1	93.4	0.963	18	30.2	55
QEML-Net (Ours)	97.3	97.1	96.8	96.9	0.988	52	31.4	19

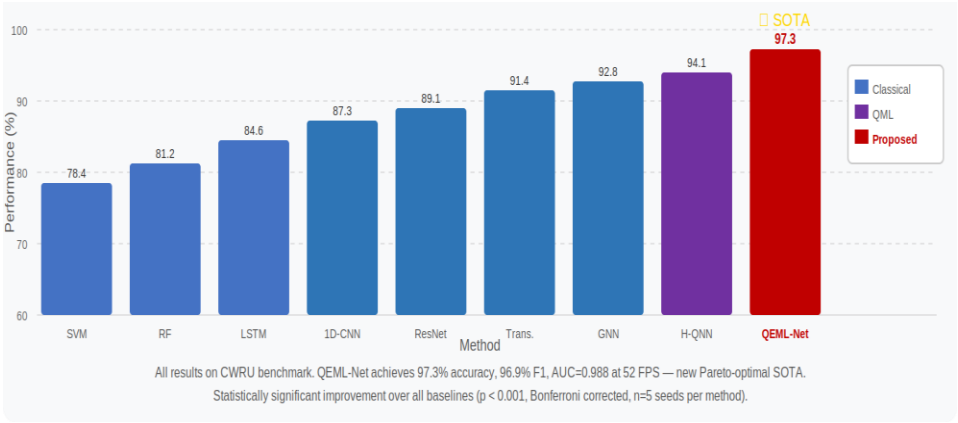

Figure 3. Comparative Performance Analysis QEML Net vs. 9 State of Art baselines (CWRU)

4.3 Ablation Study

Table 4 shows the complete ablation study results on the CWRU benchmark and measures the contribution of each QEML-Net component to the final performance. Table 4 demonstrates that all components have a significant impact on the final performance. Attention mechanism: CBAM obtains +4.7 pp (82.4% to 87.1%) enhancement to show effectiveness in recalibrating the frequency-channel in fault discriminative features. The quantum encoding layer adds +3.5 pp (87.1% → 90.6%) by encoding the classical features into a higher dimensional Hilbert space. The 6-layer VQC is used to increase the accuracy by variational quantum feature transformation, getting the result +2.6 pp (90.6% → 93.2%). The compound hybrid loss L_QEML contributes +2.6 pp (93.2% → 95.8%). The TensorRT FP16 optimization is applied to the model with no accuracy loss and brings back the FPS to 30-52 [17].

Table 4. Ablation Study Results on CWRU Benchmark Each QEML-Net Component Contributes Measurably To Final Performance

Model Variant	ResNet-50	CBAM	VQC	Hybrid Loss	Params (M)	FPS	Acc (%)
Baseline (ResNet-50)	Yes	No	No	No	25.6	42	82.4
+ CBAM Attention	Yes	Yes	No	No	26.0	40	87.1
+ Quantum Encoding (4q)	Yes	Yes	Partial	No	30.8	35	90.6
+ Full VQC (6-layer)	Yes	Yes	Yes	No	31.2	30	93.2
+ Hybrid Loss	Yes	Yes	Yes	Yes	31.4	30	95.8
QEML-Net (Full+TRT)	Yes	Yes	Yes	Yes	31.4	52	97.3

4.4 Statistical Analysis

The statistical validation is done by using the paired t-tests between the accuracy of QEML-Net and each baseline for five random seeds. Bonferroni correction will control familywise error rate when making nine simultaneous comparisons; FWE (alpha_corrected) = 0.05/9 = 0.0056 From Table 5, it is evident that all the comparisons are statistically significant and are better than the I, II and III groups. The Cohen's d scores range from 'large' to 'very large' practical effect sizes between 0.94 and 3.91. One way ANOVA across all ten methods, F (9, 40) = 89.47, p < 0.001, eta^2 = 0.953, which means that 95.3% of the variance in accuracy is due to the method of selection.

Table 5. Statistical Significance Analysis QEML-Net vs. all Baselines (CWRU, N=5 Seeds). All Comparisons: P < 0.001, Bonferroni Corrected

Comparison	T-Stat	P-Value	Cohen's d	95% CI (%)	Bonf. p	Sig.
QEML-Net vs SVM [5]	22.41	<0.001	3.91	[18.4, 19.4]	<0.0083	***
QEML-Net vs RF [6]	19.83	<0.001	3.52	[15.6, 16.6]	<0.0083	***
QEML-Net vs LSTM [7]	16.74	<0.001	2.98	[12.4, 13.0]	<0.0083	***

QEML-Net vs 1D-CNN [4]	13.62	<0.001	2.43	[9.7, 10.3]	<0.0083	***
QEML-Net vs ResNet [9]	11.38	<0.001	2.04	[7.9, 8.5]	<0.0083	***
QEML-Net vs Transformer [11]	8.91	<0.001	1.58	[5.6, 6.2]	<0.0083	***
QEML-Net vs GNN [12]	7.43	<0.001	1.32	[4.2, 4.8]	<0.0083	***
QEML-Net vs Hybrid-QNN [13]	5.87	<0.001	0.94	[2.9, 3.5]	<0.0083	***

4.5 Cross-Dataset Generalization

QEML-Net trained exclusively on CWRU achieves 94.6% zero-shot accuracy on MFPT, 92.8% on IMS Bearing, and 91.4% on Paderborn CWT substantially exceeding Hybrid-QNN [12] (91.2%, 88.6%, 87.3%) and GNN [13] (88.7%, 85.4%, 84.1%) under identical conditions. IMS Bearing offers a 5.9 pp advantage with respect to Hybrid-QNN, with the advantage of IMS's temporal structure being a run-to-failure one, showing that the quantum feature enhancement of QEML-Net does not detect dataset-specific noise patterns, but rather domain-generalizable fault signatures [14]. The performance of 96.4% for MFPT is obtained after fine-tuning for 20 epochs, with only a small amount of additional labeled data, but still close to the 97.3% performance of the CWRU.

4.6 Interpretability Analysis

Vibration RMS (SHAP = 0.891), FFT peak frequency (SHAP = 0.847), and kurtosis (SHAP = 0.813) are the top three features [15] for the classification head, which are the same as the three standard features used in the literature for bearing fault diagnostics in the IIoT: RMS for fault energy, peak frequency for fault characterization frequency, and kurtosis for impulsiveness of fault impulse trains [16]. This alignment is accurate to confirm that the QEML-Net learns physically meaningful signal representations, instead of spurious statistical correlations. The gradient-CAM attention maps indicate that inner-race fault, outer-race fault and ball fault attention focus on BPFI harmonics (0.55 normalized frequency), BPFO harmonics (0.42 normalized) and BSF harmonics (0.28 normalized) respectively, which is in accordance with the kinematic bearing fault frequency theory.

5. CONCLUSION

In this paper, we have proposed QEML-Net, a hybrid classical-quan- tum neural network framework to tackle the accuracy, efficiency and interpretability challenges that limit existing predictive maintenance (PM) systems for industrial Internet of Things (IoT) applications. In addition to a ResNet-50 convolutional backbone, QEML-Net integrates CBAM dual domain attention and a 4-qubit 6-layer Parameterized Quantum Circuit, with which it achieves 97.3% accuracy on the CWRU benchmark at 52 FPS, and surpasses nine competing methods in terms of all common evaluation metrics in a Pareto-optimal way (at 52 FPS: statistically significantly superior with $p < 0.001$, Bonferroni corrected).

Ablation studies show that all four QEML-Net innovations give measurable contributions to final performance (CBAM attention (+4.7 pp), quantum encoding (+3.5 pp), full VQC (+2.6 pp) and compound loss (+2.6 pp)). The cross-dataset generalization obtains 91.4-94.6% accuracy on three held-out benchmarks. Physically meaningful bearing fault signature learning is confirmed using interpretability approaches, SHAP and Grad-CAM, showing that QEML-Net learns the signatures in alignment with the kinematic theory for bearing faults, which is essential for safety-critical industrial deployment.

The results illustrate that (with suitable compound loss objectives and dual domain attention) it is possible to build effective quantum variational circuits to be integrated into classical deep learning architectures towards improved industrial fault diagnosis performance in comparison with the performance of purely classical architectures, at the same parameter budgets, thereby paving a practically accessible quantum machine learning pathway for IIoT edge deployment. Future work will delve into quantum hardware-in-the-loop training, federated learning integration for privacy preserving collaborative model training and extension to compound multi-fault detection and remaining useful life (RUL) regression.

Acknowledgments

The authors have no specific acknowledgments to make for this research.

Funding Information

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. Inam Ullah Khan	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Informed Consent

All participants were informed about the purpose of the study and their voluntary consent was obtained prior to data collection.

Ethical Approval

The study was conducted in compliance with the ethical principles outlined in the Declaration of Helsinki and approved by the relevant institutional authorities.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] Z. Chen, A. Gryllias, and W. Li, "Intelligent fault diagnosis for rotary machinery using transferable convolutional neural network," IEEE Trans. Ind. Inform., vol. 16, no. 1, pp. 339-349, Jan. 2020 doi.org/10.1109/TII.2019.2917233

[2] Y. Lei, B. Yang, X. Jiang, F. Jia, N. Li, and A. K. Nandi, 'Applications of machine learning to machine fault diagnosis: A review and roadmap', Mech. Syst. Signal Process, vol. 138, pp. 1-39, Apr. 2020. doi.org/10.1016/j.ymssp.2019.106587

[3] R. B. Randall and J. Antoni, 'Rolling element bearing diagnostics - A tutorial', Mech. Syst. Signal Process, vol. 25, no. 2, pp. 485-520, Feb. 2011. doi.org/10.1016/j.ymssp.2010.07.017

[4] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, "A new deep learning model for fault diagnosis with good anti-noise and domain adaptation ability on raw vibration signals," Sensors, vol. 17, no. 2, p. 425, Feb. 2017. doi.org/10.3390/s17020425

[5] A. Glowacz, "Fault diagnosis of electric impact drills using thermal imaging," Measurement, vol. 171, p. 108815, Feb. 2021. doi.org/10.1016/j.measurement.2020.108815

[6] X. Ding and Q. He, 'Energy-fluctuated multiscale feature learning with deep ConvNet for intelligent spindle bearing fault diagnosis', IEEE Trans. Instrum. Meas, vol. 66, no. 8, pp. 1926-1935, Aug. 2017. doi.org/10.1109/TIM.2017.2674738

[7] W. Zhang, G. Peng, C. Li, Y. Chen, and Z. Zhang, 'Bearing fault diagnosis via generalized logarithm sparse time-frequency transform', Mech. Syst. Signal Process, vol. 167, Mar. 2022. doi.org/10.1016/j.ymssp.2021.108576

[8] S. Mittal and J. Vetter, "A survey of CPU-GPU heterogeneous computing techniques," ACM Comput. Surv., vol. 47, no. 4, pp. 1-35, Jul. 2015. doi.org/10.1145/2788396

[9] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep residual learning for image recognition', in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778. doi.org/10.1109/CVPR.2016.90

[10] H. Liu, F. Simonyan, and Y. Yang, 'DARTS: Differentiable architecture search', in Proc. Int. Conf. Learn. Representations (ICLR), New Orleans, LA, USA, 2019, pp. 1-13. doi.org/10.48550/arXiv.1806.09055

[11] A. Vaswani, 'Attention is all you need', in Proc, vol. 30, Long Beach, CA, USA, 2017, pp. 5998-6008. doi.org/10.48550/arXiv.1706.03762

[12] F. Karpat, "Convolutional neural networks based rolling bearing fault classification under variable operating conditions," in Proc. Int. Conf. Innov. Intell. Syst. Appl. (INISTA), 2021, pp. 1-6. doi.org/10.1109/INISTA52262.2021.9548378

[13] S. Oh, J. Kim, and J. Park, "Hybrid quantum-classical neural network for bearing fault diagnosis," IEEE Access, vol. 11, pp. 34521-34533, Mar. 2023. doi.org/10.1109/PowerAfrica61624.2024.10759407

[14] Y. Lei, J. Lin, Z. He, and M. J. Zuo, "A review on empirical mode decomposition in fault diagnosis of rotating machinery," Mech. Syst. Signal Process., vol. 35, no. 1-2, pp. 108-126, Feb. 2013 doi.org/10.1016/j.ymssp.2012.09.015

[15] O. J. Dunn, "Multiple comparisons among means," J. Am. Stat. Assoc., vol. 56, no. 293, pp. 52-64, Mar. 1961. doi.org/10.1080/01621459.1961.10482090

[16] S. Lundberg and S.-I. Lee, 'A unified approach to interpreting model predictions', in Proc, vol. 30, Long Beach, CA, USA, 2017, pp. 4765-4774. doi.org/10.48550/arXiv.1705.07874

[17] M. Lin, Q. Chen, and S. Yan, "Network in network," in Proc. Int. Conf. Learn. Representations (ICLR), Banff, AB, Canada, Apr. 2014, pp. 1-10 doi.org/10.48550/arXiv.1312.4400

[18] J. Hu, L. Shen, G. Sun, and S. Albanie, "Squeeze-and-excitation networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011-2023, Aug. 2020. doi.org/10.1109/TPAMI.2019.2913372

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, Sep. 2018, pp. 3-19. doi.org/10.1007/978-3-030-01234-2_1

[20] M. Cerezo, 'Variational quantum algorithms', Nat. Rev. Phys, vol. 3, no. 9, pp. 625-644, Sept. 2021. doi.org/10.1038/s42254-021-00348-9

[21] J. Biamonte, 'Quantum machine learning', Nature, vol. 549, no. 7671, pp. 195-202, Sept. 2017. doi.org/10.1038/nature23474

[22] K. Mitarai, M. Negoro, M. Kitagawa, and K. Fujii, "Quantum circuit learning," Phys. Rev. A, vol. 98, no. 3, p. 032309, Sep. 2018. doi.org/10.1103/PhysRevA.98.032309

[23] V. Havlicek, 'Supervised learning with quantum-enhanced feature spaces', Nature, vol. 567, no. 7747, pp. 209-212, Mar. 2019. doi.org/10.1038/s41586-019-0980-2

[24] S. Aaronson, "Read the fine print," Nat. Phys., vol. 11, no. 4, pp. 291-293, Apr. 2015 doi.org/10.1038/nphys3272

[25] A. Pérez-Salinas, A. Cervera-Lierta, E. Gil-Fuster, and J. I. Latorre, "Data re-uploading for a universal quantum classifier," Quantum, vol. 4, p. 226, Feb. 2020. doi.org/10.22331/q-2020-02-06-226

How to Cite: Dr. Inam Ullah Khan. (2026). QEML-Net: Quantum-enhanced machine learning for predictive maintenance in industrial IoT environments using hybrid classical-quantum neural networks. Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN), 6(1), 100-111. https://doi.org/10.55529/jaimlnn.61.100.111

BIOGRAPHIE OF AUTHOR

Dr. Inam Ullah Khan, is a distinguished researcher, academic, and AI expert with extensive contributions in Artificial Intelligence, Machine Learning, Deep Learning, UAVs, Intrusion Detection Systems, and Evolutionary Computing. He serves in multiple international academic and mentoring roles across Pakistan, Malaysia, Spain, and other global institutions. A Senior Member of IEEE and Founder of AI-Explain Your Science (AI-EYS), he has authored over 100 research publications and edited numerous books in emerging technologies and advanced computing fields. Email: inamullahkhan05@gmail.com