Engineering Applications of Artificial Intelligence 177 (2026) 114991

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research paper

A comprehensive review on quantum deep neural networks for prognostics and health management: Fundamentals, challenges and opportunities☆

Mayank Shekhar Jha [a,*], Sameul Yen-Chi Chen [b], Chetan Kulkarni [c], Joongheon Kim [d]

[a] Université de Lorraine, CNRS, CRAN, Nancy, 54000, France
[b] Brookhaven National Laboratory, Upton, NY, USA
[c] KBR Inc, NASA Ames Research Center, Mountain View, USA
[d] Korea University, Seoul, Korea

ARTICLE INFO

Keywords:
Parameterized quantum circuits
Quantum computing
Quantum deep neural network
Quantum deep learning
Prognostics
Quantum prognostics
Remaining useful life

ABSTRACT

The domain of Prognostics and Health Management (PHM) targets accurate prediction of remaining useful life (RUL), efficient estimation of state of health (SoH), and assessment of anomaly indicators from multivariate data emanating from systems under functional as well as structural degradation that are prevalent across various engineering areas. Proactive assessment of SoH and prediction of RUL are core objectives of PHM and remain challenging as degradation dynamics is commonly nonlinear, (plausibly) non-stationary, multi-sensor based, often data-limited, and frequently affected by changing operating conditions. Quantum Deep Neural Networks (QDNNs), typically implemented through parameterized quantum circuits within hybrid quantum–classical workflows, have recently emerged as a potential alternative or complement to classical deep models due to their compact parameterization and expressive quantum feature spaces. This article presents a tutorial-style review of QDNNs for PHM. We first summarize the quantum-computing foundations required to interpret such models, including qubits, measurement, data encoding, parameterized quantum circuits, gradient estimation, and hybrid optimization. We then organize the PHM literature by architecture family: feedforward, recurrent, convolutional, generative, and attention-based quantum models. We discuss, for each family, its mathematical principle, typical encoding choices, datasets, reported metrics, and comparison baselines. The current evidence suggests that QDNNs can be competitive and sometimes more parameter-efficient than selected classical baselines; however, the literature remains heterogeneous, often simulator-dominated, and insufficient to establish systematic quantum advantage. We conclude by identifying the main open challenges as well as emerging opportunities for quantum enhanced prognostics.

1. Introduction

Over the past decade, deep neural networks (DNNs) have emerged as a prominent family of approaches in artificial intelligence leading to fundamental and applicative breakthroughs in computer vision, speech processing, natural language understanding, and predictive analytics (Goodfellow et al., 2016). The success of DNNs stems from hierarchical representation learning, wherein convolutional (CNN), recurrent (RNN/LSTM), and attention-based transformer architectures automatically extract multiscale and temporal correlations from data (Vaswani et al., 2017; He et al., 2016). Several architectural families have proven especially influential. Convolutional Neural Networks (CNNs) introduce convolutional filters and weight sharing to capture local structure and exploit translation equivariance, making them highly effective for grid-structured data such as images and spectrograms (Goodfellow et al., 2016; He et al., 2016). For sequential and time-series data, recurrent neural networks (RNNs) model temporal dependencies by updating a hidden state over time; however, vanilla RNNs can suffer from vanishing/exploding gradients when learning long-range dependencies. Long short-term memory (LSTM) networks address this limitation by introducing gating mechanisms that regulate information flow and

☆ This article is part of a Special issue entitled: 'Quantum Technologies for Practical Applications' published in Engineering Applications of Artificial Intelligence.

* Corresponding author.
E-mail addresses: mayank.jha@univ-lorraine.fr (M.S. Jha), ycchen1989@ieee.org (S.Y.-C. Chen), chetan.s.kulkarni@nasa.gov (C. Kulkarni), joongheon@korea.ac.kr (J. Kim).

https://doi.org/10.1016/j.engappai.2026.114991
Received 30 December 2025; Received in revised form 3 April 2026; Accepted 27 April 2026
Available online 29 April 2026

stabilize learning over long horizons (Hochreiter and Schmidhuber, 1997). More recently, transformers have emerged as a general-purpose sequence modeling framework. They replace recurrence with self-attention, a mechanism that computes context-dependent interactions between all positions in a sequence, enabling efficient parallel training and strong performance in language and other sequence-learning problems (Vaswani et al., 2017). Despite their broad applicability, the growing accuracy of deep models has often been accompanied by rapid increases in parameter count, training compute, and memory/communication demands, raising concerns regarding scalability, accessibility, and energy efficiency (Talaei Khoei et al., 2023). Empirical studies have quantified substantial financial and environmental costs associated with training large deep learning models, motivating the development of more compute- and energy-aware training practices (sometimes summarized under the umbrella of Green AI) (Schwartz et al., 2020; Patterson et al., 2021).

Quantum computing offers a fundamentally different computational paradigm in which information is encoded in quantum states, processed by unitary evolution, and extracted by measurement (Biamonte et al., 2017; Nielsen and Chuang, 2010; Preskill, 2018). For certain problems, quantum algorithms provide provable asymptotic speedups; the best-known examples are Shor's factoring algorithm and Grover's unstructured search (Shor, 1994; Grover, 1996). More generally, complexity-theoretic separations have also been established for certain sampling and adiabatic settings, although evidence for broadly useful practical advantage in generic optimization and machine learning remains problem-dependent and actively debated (Abbas et al., 2024). Importantly, access to exponentially large Hilbert spaces does not by itself imply computational advantage; any useful speedup must arise from algorithmic exploitation of interference, entanglement, and measurement structure.

The convergence of quantum computing and artificial intelligence has led to the emergence of a new paradigm, Quantum Machine Learning (QML), in which quantum resources such as superposition, entanglement, and interference are exploited to enhance learning and inference capabilities (Biamonte et al., 2017). QML promises algorithmic and representational advantages over classical models by encoding data in exponentially large Hilbert spaces and executing computations in superposition (Cerezo et al., 2021; Dunjko and Briegel, 2018). Among the broad landscape of Quantum Machine Learning (QML) approaches, a particularly active and practical family is formed by Quantum Deep Neural Networks (QDNNs). In this review, we use "QDNN" to denote layered quantum (or hybrid quantum–classical) models in which trainable quantum subroutines play the role of learnable layers, in close analogy to classical deep networks. In near-term settings, these layers are most commonly implemented as Parameterized Quantum Circuits (PQCs) or Variational Quantum Circuits (VQCs), i.e., circuits whose gates depend on a set of real-valued trainable parameters and are optimized through a classical outer-loop optimizer using measurement-derived loss values (Benedetti et al., 2019; Cerezo et al., 2021).

This variational (hybrid) workflow is attractive because it can be executed on Noisy Intermediate-Scale Quantum (NISQ) hardware (term to denote pre–fault-tolerant quantum processors with limited qubit counts and imperfect gates (Preskill, 2018)) and because it supports end-to-end training in applications where classical data are encoded into quantum states (via an embedding/feature map) and processed by PQC layers (Schuld et al., 2015; Havlíček et al., 2019). A recurring motivation for QDNNs is parameter efficiency relative to the exponential dimension of the underlying quantum state space. An n-qubit register is described by a state in a 2^n-dimensional complex Hilbert space, but a PQC with p tunable parameters explores only a restricted family of states generated by its circuit ansatz (a fixed gate template and entangling pattern).

Whether and when such models can offer a genuine quantum advantage remains an open and nuanced question. In this review, we use "quantum advantage" in the broad sense of a demonstrable benefit of quantum computation over the best-known classical alternatives under comparable resource assumptions (e.g., runtime, sample complexity, or model size), noting that the appropriateness of this goal in data-driven learning has been explicitly debated in the QML community (H.-Y. Huang et al., 2021). Nevertheless, several well-studied routes by which quantum models may yield advantage have emerged: (i) sampling and generative modeling tasks where certain PQC-generated distributions are conjectured (under complexity-theoretic assumptions) to be hard to reproduce classically (Du et al., 2020); (ii) kernel/feature-map methods in which a quantum embedding yields a similarity kernel that may be expensive to compute classically for carefully chosen feature maps (Havlíček et al., 2019); and (iii) variational learning and optimization settings that leverage PQCs as trainable models, whose practical performance hinges on NISQ trainability, error mitigation, and careful benchmarking against strong classical baselines (Cerezo et al., 2021; Benedetti et al., 2019).

Table 1 presents the list of abbreviations and acronyms used throughout the article.

Quantum mechanics also permits entanglement i.e., non-classical correlations between qubits that cannot be factorized into independent states. Entanglement is the essential resource enabling non-local correlations and the enhanced expressive power of Quantum Neural Networks (QNNs) (Biamonte et al., 2017; Abbas et al., 2021; Bharti et al., 2022). Quantum *entanglement* provides intrinsic multi-feature correlation without explicit weight matrices, plausibly improving learning efficiency. Whether such expressivity translates into improved sample efficiency or generalization (i.e., performance on unseen data) is an active research topic.

Parallel to advances in quantum computing, Prognostics and Health Management (PHM) has emerged, in the past decade, as a core domain that enables predictive maintenance through proactive estimation of a system's Remaining Useful Life (RUL) (Sikorska et al., 2011). This task of predicting RUL is referred to as prognostics and constitutes a core objective of PHM. It shapes a predictive maintenance strategy by integrating the latter for decision support and automation (Sikorska et al., 2011; Jha et al., 2016a; Kanso et al., 2022). Within industrial machinery context, diagnostic functions include fault detection, isolation, and identification (Jha et al., 2016a). While fault detection and isolation (FDI) identifies abnormal operation by comparing measured signals (e.g., pressure, temperature, vibration) against expected behavior (Jha et al., 2016b; Jardine et al., 2006), PHM is distinguished by its focus on forecasting RUL (preferably) under uncertainty (Jha et al., 2016a; Kanso et al., 2022). Prognostic methods are commonly grouped into model-based, data-driven, and hybrid approaches (Sikorska et al., 2011). Hybrid schemes fuse online measurements with *a priori* physics using observers/estimators and often infer state of health (SoH) before propagating it to predict RUL; however, they typically require degradation models (Chelouati et al., 2021; Thuillier et al., 2024; Kanso et al., 2022).

Within PHM, three task families are especially relevant to this review. Remaining Useful Life (RUL) prediction aims to estimate the time, cycles, or usage horizon remaining before a defined failure threshold is reached. State of Health (SoH) tracking instead estimates a continuous health variable that summarizes the current degradation level of the asset relative to healthy and end-of-life conditions. While SoH and RUL are related, they are not identical, in that SoH is a present-state descriptor, whereas RUL is a future-oriented prognostic quantity that depends on both current health and future operating conditions (Sikorska et al., 2011; Jha et al., 2016a; Kanso et al., 2022).

Anomaly detection also requires careful interpretation in PHM. A data-level anomaly refers to an abnormal observation in the measured signal itself, such as an outlier, missing segment, calibration drift, transmission artifact, or sensor corruption. By contrast, an equipment-level anomaly refers to a physically meaningful deviation in system behavior

Table 1
List of abbreviations and acronyms used throughout the article.

Abbreviation	Description
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
QCNN	Quantum Convolutional Neural Network
RNN	Recurrent Neural Network
QRNN	Quantum Recurrent Neural Network
LSTM	Long Short-Term Memory network
QLSTM	Quantum Long Short-Term Memory network
GRU	Gated Recurrent Unit
QWGRU	Quantum-Weighted Gated Recurrent Unit
QBiLSTM	Quantum Bidirectional LSTM
QConvLSTM	Quantum Convolutional Long Short-Term Memory network
QNN	Quantum Neural Network
QDNN	Quantum Deep Neural Network
QGAN	Quantum Generative Adversarial Network
QBM	Quantum Boltzmann Machine
QVAE	Quantum Variational Autoencoder
QREDNN	Quantum Recurrent Encoder–Decoder Neural Network
QAREDNN	Quantum Attention Recurrent Encoder–Decoder Neural Network
QTransformer/QTrans	Quantum Transformer network
SAA-QCNN	Siamese Attention-Augmented Quantum Convolutional Neural Network
HQRNN	Hybrid Quantum Recurrent Neural Network
PQC	Parameterized Quantum Circuit
VQA	Variational Quantum Algorithm
VQE	Variational Quantum Eigensolver
NISQ	Noisy Intermediate-Scale Quantum
PHM	Prognostics and Health Management
RUL	Remaining Useful Life
SoH	State of Health
HI	Health Indicator
RMSE/MAE	Root-Mean-Square Error/Mean Absolute Error
XJTU-SY	Xi'an Jiaotong University – SY bearing degradation dataset
PEMFC	Proton Exchange Membrane Fuel Cell (stack degradation dataset)
GAN	Generative Adversarial Network
VAE	Variational Autoencoder
Adam/SPSA	Adaptive Moment Estimation/Simultaneous Perturbation Stochastic Approximation
POVM	Positive Operator-Valued Measure
CNOT	Controlled-NOT gate
C-MAPSS	Commercial Modular Aero-Propulsion System Simulation

caused by abnormal operation, incipient fault development, or degradation. The two should not be conflated: a model may detect anomalous data without the asset being faulty, and conversely an incipient fault may initially manifest only through subtle distributional changes rather than obvious signal outliers. This distinction is especially important for unsupervised and generative QDNNs, whose outputs may otherwise be interpreted ambiguously by cross-disciplinary readers.

Over the past decade, DNN-based approaches have emerged as powerful tools for modeling nonlinear and non-stationary machinery degradation dynamics (Fink et al., 2020). In prognostics, DNN-based approaches differ fundamentally from physics-based and hybrid schemes in that they directly learn the mapping between multivariate sensor measurements and the desired RUL target, without relying on explicit physical or empirically derived degradation models (de Beaulieu et al., 2022a, 2024), enabling end-to-end mappings from multivariate time-series measurements to RUL without explicit physics or empirically derived degradation models (Schmidhuber, 2015; Fink et al., 2020; de Beaulieu et al., 2022a, 2024). CNN-based regression and recurrent architectures (notably LSTMs) have been extensively studied for RUL prediction on C-MAPSS and battery datasets, with subsequent CNN refinements, CNN–LSTM hybrids, and attention mechanisms improving performance, particularly under varying operating conditions and long sequences (Babu et al., 2016; Yuan et al., 2016; de Beaulieu et al., 2022b; Zheng et al., 2017; Wu et al., 2018; Wang et al., 2018; Liu et al., 2017; Li et al., 2018; An et al., 2020; Bahdanau et al., 2014; da Costa et al., 2019; H. Zhang et al., 2020; Z. Chen et al., 2020; Xiang et al., 2020). This end-to-end learning capability has enabled accurate health-state estimation and long-horizon RUL prediction even under complex degradation patterns, paving the way for fully data-driven PHM frameworks. While effective, these data-driven approaches typically require substantial historical data, careful pre-processing, robust target construction, and strong controls against leakage and operating-condition shift, which motivates exploration of alternative compact and potentially data-efficient learning paradigms such as quantum deep learning.

Fig. 1 presents an illustration of the core tasks of PHM and workflow under DNNs, encompassing raw data acquisition and preprocessing, feature extraction, and subsequent input to DNNs, leading to SoH estimation through single-step and multi-step forecasting and/or RUL prediction.

From an application standpoint, PHM offers a particularly meaningful setting in which to test hybrid quantum models. First, many PHM tasks rely on multivariate sensor signals whose degradation signatures are nonlinear, often non-stationary and strongly coupled across time, sensors, and operating regimes. Second, labeled run-to-failure trajectories are often scarce or expensive to obtain, which makes compact models and data-efficient feature mappings attractive. Third, PHM decisions ideally require not only a point estimate, but also some notion of confidence or uncertainty. These characteristics do not prove that quantum models are preferable, but they do explain why QDNNs have attracted attention in PHM: the combination of quantum feature maps, entangling transformations, and measurement-based outputs offers a potentially compact route for modeling complex degradation patterns within hybrid classical-quantum pipelines (Abbas et al., 2021; Bharti et al., 2022; Caro et al., 2021).

In this context, QDNNs have begun to emerge as compact hybrid alternatives for selected PHM tasks. Early work by Silva and Droguett (2022) presented one of the first PHM-oriented hybrid quantum–classical frameworks, exemplified on a ball-bearing health-state diagnosis problem. Subsequent studies explored battery health prediction with shallow feedforward quantum regressors and hybrid sequential

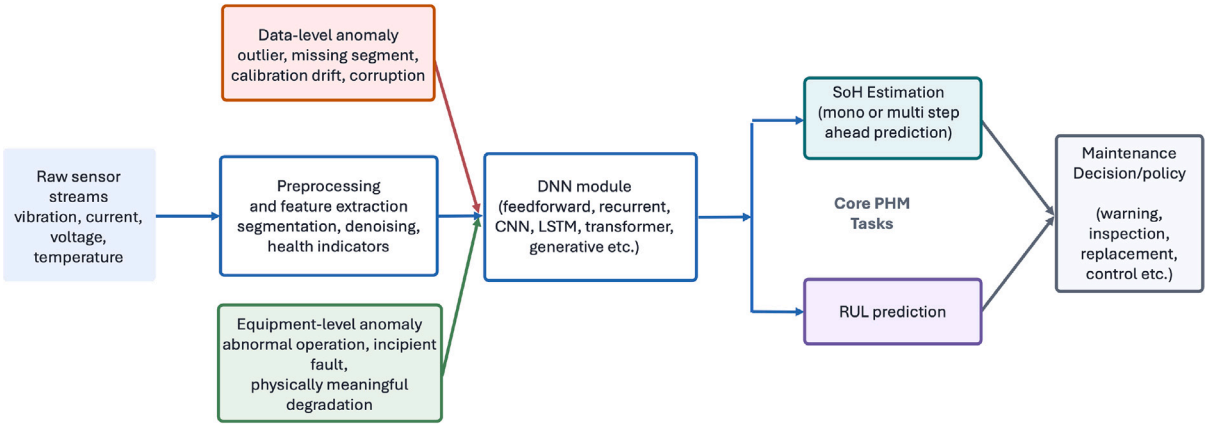

Fig. 1. Illustration of core tasks of PHM and work-flow with DNNs.

models; for example, Ngo et al. (2023) modeled lithium-ion battery capacity degradation with a PQC-based regressor, while Soon and Soon (2025) proposed a sequential QNN-GRU architecture for battery aging. More recent works have introduced quantum recurrent architectures (QLSTM, QGRU), QCNN-based models, and attention-based quantum sequence models for bearings, batteries, PEMFCs, and turbofan engines (Tsurkan et al., 2025; Wang et al., 2024; Xiang et al., 2018; Liang et al., 2025; Zaid et al., 2024; Y. Chen et al., 2020). Collectively, these studies suggest that QDNNs are a promising PHM research direction, particularly in relation to compactness and hybrid composability, although systematic superiority over strong classical baselines remains unestablished.

Before reviewing various QDNNs that are being developed for prognostics, we underline the relevance of the former for the latter using a unifying mathematical view. Let $\mathbf{X}_{1:T} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ denote a multivariate degradation trajectory and let y denote a PHM target such as remaining useful life (RUL), state of health (SoH), anomaly score, or a health-state label. A generic hybrid QDNN can be written as

$$\hat{y} = h_{\boldsymbol{\omega}}\left(\boldsymbol{\mu}_{\boldsymbol{\Theta}}(\mathbf{X}_{1:T})\right), \tag{1}$$

$$\boldsymbol{\mu}_{\boldsymbol{\Theta}}(\mathbf{X}_{1:T}) = \left[\mathrm{Tr}(O_1\rho_{\boldsymbol{\Theta}}(\mathbf{X}_{1:T})), \ldots, \mathrm{Tr}(O_m\rho_{\boldsymbol{\Theta}}(\mathbf{X}_{1:T}))\right], \tag{2}$$

where $\rho_{\boldsymbol{\Theta}}(\mathbf{X}_{1:T})$ is the quantum state prepared by the chosen encoding and trainable circuit, $\{O_j\}_{j=1}^{m}$ are measured observables, and $h_{\boldsymbol{\omega}}(\cdot)$ denotes optional classical post-processing. In other words, the quantum module learns a structured transformation from encoded degradation data to measurement statistics, which are then mapped to the desired PHM output. The architectural families reviewed in this paper differ primarily in the structure imposed on $\rho_{\boldsymbol{\Theta}}(\mathbf{X}_{1:T})$. For example, Feedforward QNNs apply a one-shot transformation to a fixed feature vector or short temporal window, recurrent models reuse a shared quantum cell over time, QCNNs impose locality and hierarchical pooling, generative models learn a density or latent representation of healthy/degrading behavior, and attention-based models perform context-dependent mixing across time-steps or sensor channels. Their differences are therefore not merely taxonomic; they correspond to different inductive biases about how degradation information is organized. This distinction is especially important in PHM because prognostic tasks are heterogeneous. When the input is a compact set of health indicators, a feedforward QNN may be sufficient. When degradation unfolds over long horizons, recurrent or attention-based quantum hybrids are more natural. When local correlations dominate, e.g., across neighboring time windows or time-frequency patches, QCNN-type locality becomes relevant. When the goal is anomaly detection or rare-fault modeling rather than direct RUL regression, generative architectures become more appropriate. The motivation for QDNNs in PHM therefore comes from a potential match between architecture-specific inductive bias and problem structure, rather than from any assumption that a quantum model is automatically superior to a classical one.

The literature at the intersection of QML and PHM is evolving rapidly. General QML surveys continue to emphasize algorithmic, theoretical, or physics-oriented aspects rather than PHM applications (Dunjko and Briegel, 2018; Cerezo et al., 2021; Schuld et al., 2015). More recently, battery-centered reviews of quantum machine learning for energy-storage health prediction and optimization have begun to appear (Khakpour Komarsofla and Kiani, 2026), confirming the growth of the area but also highlighting its current fragmentation around specific asset classes.

However, a tutorial-style comprehensive review explicitly organized by QDNN architecture and PHM task is still lacking in the literature. In particular, what is still needed is a unified comprehensive summary that simultaneously explains the mathematical foundations of QDNNs, differentiates their architectural families, and critically evaluates their usefulness for RUL prediction, SoH estimation, anomaly detection, and degradation forecasting across heterogeneous PHM settings. To address this need, the present work offers a structured, tutorial-style review of quantum deep learning architectures for prognostics, with the following main contributions:

- a concise introduction to basic quantum computation concepts and their correspondence with classical deep learning structures covering qubits, quantum measurement, data encoding, PQCs, and hybrid quantum–classical training, to help new researchers rapidly build intuition in quantum DNNs;
- a systematic survey of all proposed QDNNs based approaches for prognostics, including feedforward, recurrent, convolutional, generative, and attention-based architectures, summarizing datasets, circuit designs, and empirically reported improvements (performance gains or/and parameter efficiency) under specific encodings/ansätze;
- precise identification of standing scientific gaps, challenges and open research opportunities to guide future developments in quantum-enhanced PHM.

This paper identifies in a precise manner existing scientific gaps and proposes future research directions in the domain of quantum enhanced prognostics. Through this work, we aim to equip readers from both the PHM and quantum computing communities with the conceptual tools necessary to understand, design, and evaluate quantum-enhanced prognostic systems. It is noted that this work does not assume a priori that quantum models are superior to classical deep networks for PHM tasks. Rather, it examines the current literature through three complementary questions: (i) which QDNN architectures have actually been explored for PHM-relevant tasks; (ii) under what encoding, dataset, and benchmarking conditions they have reported competitive behavior; and

(iii) what remains missing before stronger claims regarding robustness, scalability, or quantum advantage can be made. This distinction is essential because the current evidence base is heterogeneous and still largely shaped by the constraints of NISQ-era implementations.

The remainder of this paper is organized as follows. Section 2 introduces fundamental quantum computing concepts required for understanding QDNNs, Section 3 reviews the major QDNN model families and training principles, then Section 4 reviews the prognostics works based on the previously presented QDNN models and summarizes representative studies. Further, Section 5 discusses standing challenges in quantum enhanced prognostics, and Section 6 outlines emerging opportunities and future directions. Finally, Section 7 draws conclusions.

2. Fundamentals of quantum computing for deep learning

Quantum computing differs from classical computation in three fundamental aspects: (i) representation via quantum states (qubits) rather than bits; (ii) evolution via reversible unitary operations (and, in practice, noisy quantum channels) rather than irreversible Boolean logic; and (iii) measurement, which yields probabilistic outcomes and expectation values rather than deterministic output. The following subsections present these principles that determine the algorithmic foundations of quantum deep learning, where information is encoded, transformed, and inferred through quantum state manipulation.

2.1. Qubits, superposition, and entanglement

At the core of quantum computing lies the qubit the quantum analogue of a classical bit, which can occupy the basis states $|0\rangle$ and $|1\rangle$, or any linear superposition thereof, thus providing exponentially richer representational capacity than binary logic (Preskill, 2018). A single qubit is described by a normalized state vector in the two-dimensional Hilbert space $\mathcal{H}_2$ spanned by the orthonormal basis $\{|0\rangle, |1\rangle\}$:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \qquad \alpha, \beta \in \mathbb{C},\ |\alpha|^2 + |\beta|^2 = 1, \tag{3}$$

where $|\psi\rangle$ denotes the pure state of the qubit, and α and β are the complex probability amplitudes whose squared magnitudes $|\alpha|^2$ and $|\beta|^2$ correspond to the probabilities of measuring the qubit in states $|0\rangle$ and $|1\rangle$, respectively. Neglecting the global phase (which has no observable effect), any qubit state can equivalently be parameterized by two real angles ϑ and φ as

$$|\psi\rangle = \cos\left(\frac{\vartheta}{2}\right)|0\rangle + e^{i\varphi}\sin\left(\frac{\vartheta}{2}\right)|1\rangle, \tag{4}$$

where $\vartheta \in [0, \pi]$ and $\varphi \in [0, 2\pi)$ define the polar and azimuthal coordinates of the state on the Bloch sphere (Nielsen and Chuang, 2010).

For a system of n qubits, the composite state lies in the 2^n-dimensional Hilbert space $\mathcal{H} = (\mathbb{C}^2)^{\otimes n}$ and can be expressed as

$$|\Psi\rangle = \sum_{i=0}^{2^n-1} c_i|i\rangle, \qquad c_i \in \mathbb{C},\ \sum_i |c_i|^2 = 1,$$

where $\{|i\rangle\}$ denotes the computational basis and the coefficients c_i represent the corresponding probability amplitudes.

In QDNNs, entanglement plays a central representational role. Entanglement refers to non-separable correlations in a composite quantum state, i.e., the joint state of multiple qubits cannot be written as a tensor product of independent single-qubit states. A pure n-qubit state $|\Psi\rangle$ is called (fully) separable if it can be written as a tensor product of single-qubit states,

$$|\Psi\rangle = \bigotimes_{j=1}^{n} |\psi_j\rangle.$$

If no such factorization exists, the state is entangled. In the two-qubit case, this reduces to the condition $|\Psi\rangle \neq |\psi_1\rangle \otimes |\psi_2\rangle$. A canonical example is the two-qubit Bell state

$$|\Phi^+\rangle = \frac{1}{\sqrt{2}}\big(|00\rangle + |11\rangle\big),$$

which exhibits maximal entanglement. When an ansatz includes entangling gates (e.g., CNOT/CZ patterns), measurement outcomes can depend on joint, higher-order interactions among the features encoded across different qubits. From a machine-learning perspective, this provides a mechanism for modeling multivariate dependencies without explicitly constructing large classical weight tensors that enumerate pairwise or higher-order feature couplings. The extent to which a parameterized circuit can explore correlated (entangled) regions of Hilbert space is closely tied to its expressibility and entangling capability, which can be quantified and compared across circuit templates (Biamonte et al., 2017).

2.2. Quantum measurement

Quantum computation differs from classical computation not only in how states are represented but also in how information is extracted. Measurement forms the interface between the quantum and classical worlds: it converts quantum amplitudes into classical probabilities and expectation values that can be processed by classical optimizers, closing the feedback loop in hybrid training schemes (Cerezo et al., 2021; Benedetti et al., 2019). A general quantum measurement is characterized by a set of operators $\{M_m\}$ acting on the system's Hilbert space $\mathcal{H}$, satisfying the completeness relation

$$\sum_m M_m^\dagger M_m = I, \tag{5}$$

where $M_m^\dagger$ is the Hermitian conjugate of M_m, and I is the identity operator on $\mathcal{H}$. For a state $|\psi\rangle$, the probability of obtaining outcome m is

$$p(m) = \langle\psi|M_m^\dagger M_m|\psi\rangle, \tag{6}$$

and the corresponding post-measurement state is

$$|\psi_m\rangle = \frac{M_m|\psi\rangle}{\sqrt{p(m)}}. \tag{7}$$

On current NISQ devices, the hardware does not directly return an exact expectation value. The readers are referred to Section 2.5 for more details on the term NISQ, that is popularized by Preskill to denote pre–fault-tolerant quantum processors with limited qubit counts and imperfect gates (Preskill, 2018). Instead, one performs repeated projective measurements, often of Pauli observables or Pauli strings, and estimates expectation values from the resulting shot statistics (Preskill, 2018; Cerezo et al., 2021). If an observable has spectral decomposition $O = \sum_m m\,\Pi_m$, then for state ρ the Born rule gives $p(m) = \mathrm{Tr}(\Pi_m\rho)$ and the corresponding expectation is

$$\langle O\rangle = \mathrm{Tr}(O\rho) = \sum_m m\,p(m), \tag{8}$$

which is estimated empirically from repeated circuit executions. This scalar quantity serves as the output of a parameterized quantum circuit (PQC) and feeds into a classical optimizer for learning. More generally, measurements are described by a Positive Operator-Valued Measure (POVM) (Heinosaari and Ziman, 2010), defined by positive semi-definite operators $E_m = M_m^\dagger M_m$ satisfying $\sum_m E_m = I$. POVMs extend projective measurements to account for noise and partial detection, providing a realistic framework for measurement modeling in NISQ-era quantum learning. More generally, for a POVM $\{E_m\}$ with assigned outcome values $o_m \in \mathbb{R}$, one has $\mathbb{E}[o] = \sum_m o_m\,\mathrm{Tr}(E_m\rho)$ for a (possibly mixed) state ρ. In most QML studies, the measurement map is fixed in advance, typically through Pauli observables such as Z or tensor products of Pauli operators. However, the choice of observable is part of the effective hypothesis class after measurement, because it determines which statistics of the final quantum state are exposed to the classical optimizer. Recent work has therefore begun to treat the observable itself as a learnable or input-adaptive object, showing that jointly optimizing the circuit and the measurement rule can improve task performance in some settings (Chen et al., 2025a,b).

2.3. Data encoding and quantum feature maps

Once measurement extracts information from quantum states, the next step toward learning is to embed classical data into quantum states. This process, also known as data encoding or state preparation, forms the interface between classical information and quantum computation (Schuld and Killoran, 2019; Mitarai et al., 2018; Abbas et al., 2021; Schuld, 2022). Given a classical feature vector $\mathbf{x} \in \mathbb{R}^d$, data encoding maps it into a quantum state within an n-qubit Hilbert space as

$$|\phi(\mathbf{x})\rangle = U_\phi(\mathbf{x})\,|0\rangle^{\otimes n}, \tag{9}$$

where $|0\rangle^{\otimes n}$ denotes the ground state of all n qubits and $U_\phi(\mathbf{x})$ is a unitary encoding operator that embeds $\mathbf{x}$ into quantum amplitudes, phases, or gate parameters. The mapping $\mathbf{x} \mapsto |\phi(\mathbf{x})\rangle$ defines a quantum feature map whose similarity measure is given by the (fidelity) kernel

$$k(\mathbf{x}, \mathbf{x}') = |\langle\phi(\mathbf{x})|\phi(\mathbf{x}')\rangle|^2, \tag{10}$$

representing the squared overlap between encoded states $|\phi(\mathbf{x})\rangle$ and $|\phi(\mathbf{x}')\rangle$. This kernel acts as the quantum analogue of feature-space similarity in classical kernel methods, capturing nonlinear correlations through the geometry of the Hilbert space. The geometry of this feature map determines the expressive capacity and separability of the model in Hilbert space. Common encoding schemes include: basis encoding, mapping binary inputs to computational basis states ($x_i \in \{0,1\} \to |x_i\rangle$); amplitude encoding, embedding normalized features as state amplitudes, achieving exponential compression but requiring complex preparation; angle (rotation) encoding, where data modulate gate rotations, e.g., $R_Y(x_i) = e^{-ix_iY/2}$, widely used on NISQ devices (Schuld, 2022; Preskill, 2018); phase encoding, using phase shifts to represent data; and hybrid/entangling encodings, which combine local rotations with controlled gates to capture nonlinear correlations (Abbas et al., 2021; Havlíček et al., 2019).

While quantum evolution is linear in the state vector, the induced input–output map $\mathbf{x} \mapsto \langle O\rangle_{\mathbf{x},\theta}$ is generally nonlinear in $\mathbf{x}$ due to the data-dependent unitary $U_\phi(\mathbf{x})$ and the Born rule used to extract observables. Quantum feature maps can induce very high-dimensional feature spaces, and for certain encoding families the effective feature dimension or Fourier spectrum can grow rapidly with qubit number and circuit structure (Mitarai et al., 2018; Schuld, 2022; Havlíček et al., 2019). However, this growth is strongly encoding-dependent and should not be conflated with an automatic practical advantage.

2.3.1. Encoding choices in current quantum prognostics studies

In the prognostics literature reviewed here, angle or *rotation encoding* is by far the most common choice. The reason is primarily practical in that, rotation-based encoding is shallow, hardware-friendly, and naturally compatible with multivariate health indicators, cyclic features, spatio-temporal indicators. In such schemes, each feature (or a subset of features after classical compression) modulates one or more single-qubit rotations such as $R_Y(x_i)$ or $R_Z(x_i)$ (Silva and Droguett, 2022). This makes angle encoding particularly attractive for NISQ implementations, where circuit depth must remain small. Several reviewed prognostic studies therefore adopt a pipeline in which handcrafted or compressed features are first normalized classically and then embedded through rotation gates before variational processing by a PQC (Silva and Droguett, 2022). For example, Silva and Droguett (2022) employs angle encoding of five normalized bearing features into a five-qubit pipeline. A second recurrent pattern is classical compression followed by quantum encoding (Wang et al., 2024). This is especially important in PHM because raw multivariate sensor streams are often too high-dimensional to be embedded directly into the limited number of available qubits. In such cases, classical pre-processing layers, statistical feature extraction, or learned embeddings reduce dimensionality before the reduced representation is encoded into the quantum circuit. This design improves implementability, but it also raises an interpretive issue, if most of the task-relevant structure is already extracted classically, the contribution of the quantum layer must be assessed carefully. More expressive alternatives include entangling feature maps and repeated data re-uploading, which can encode cross-feature interactions more richly than purely local rotations. These strategies may increase representational power by allowing the encoded state to capture nonlinear correlations already at the embedding stage. However, they also increase circuit depth, sensitivity to noise, and optimization difficulty.

By contrast, amplitude encoding is attractive in principle because it can represent 2^n amplitudes using only n qubits, but generic state preparation is typically more demanding than angle encoding and is therefore less common in current PHM-oriented QDNN implementations. Hence, for near-term quantum prognostics, the dominant trade-off is clear: angle-based and hybrid encodings are preferred for implementability and trainability, whereas more compact or more expressive encodings are attractive theoretically but often harder to realize reliably on NISQ hardware.

From a PHM perspective, encoding should not be viewed as a purely technical preprocessing step. It directly determines what degradation structure becomes accessible to the quantum model. Encodings that preserve monotonic health evolution, regime dependence, cross-sensor coupling, or local temporal signatures are likely to be more meaningful than generic embeddings that ignore degradation semantics.

2.4. Parameterized quantum circuits and quantum neural networks

PQCs constitute the trainable core of quantum machine learning, implementing differentiable quantum operations through tunable gates. A PQC is composed of a sequence of L unitary blocks acting on an initial n-qubit state $|0\rangle^{\otimes n}$:

$$U(\boldsymbol{\theta}) = U_L(\theta_L)\cdots U_2(\theta_2)U_1(\theta_1), \qquad U_k(\theta_k) = e^{-i\theta_k H_k/2}, \tag{11}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$ are trainable parameters and H_k are Hermitian generators, commonly chosen from the Pauli basis

$$X = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad Y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \quad Z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix},$$

ensuring unitary, norm-preserving evolution. In particular, for Pauli rotations $U_k(\theta_k) = e^{-i\theta_k P/2}$ with $P \in \{X, Y, Z\}$, the generator has eigenvalues $\pm\frac{1}{2}$ and the standard shifts $\pm\frac{\pi}{2}$ apply. Parameterized rotations such as $R_X(\theta) = e^{-i\theta X/2}$ and $R_Y(\theta) = e^{-i\theta Y/2}$ act as trainable feature transformations in Hilbert space; effective nonlinearity enters through measurement statistics and classical post-processing.

Given classical input $\mathbf{x}$, a fixed encoding map $U_\phi(\mathbf{x})$ embeds the data before the trainable PQC:

$$|\psi(\mathbf{x}, \boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})\,U_\phi(\mathbf{x})\,|0\rangle^{\otimes n}. \tag{12}$$

Task-specific objectives are defined as expectation values of observables O measured on the output state,

$$C(\mathbf{x}, \boldsymbol{\theta}) = \langle\psi(\mathbf{x}, \boldsymbol{\theta})|\, O\, |\psi(\mathbf{x}, \boldsymbol{\theta})\rangle \tag{13}$$

$$= \langle 0^{\otimes n}|U_\phi^\dagger(\mathbf{x})\,U^\dagger(\boldsymbol{\theta})\,O\,U(\boldsymbol{\theta})\,U_\phi(\mathbf{x})|0^{\otimes n}\rangle, \tag{14}$$

optimized via hybrid quantum–classical feedback. For gates whose generators have an appropriate two-eigenvalue spectrum, derivatives of exact expectation values can be expressed by the parameter-shift rule (Mitarai et al., 2018; Cerezo et al., 2021):

$$\frac{\partial C(\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_\ell} = \frac{1}{2}\left[C(\mathbf{x}, \boldsymbol{\theta} + \tfrac{\pi}{2}\mathbf{e}_\ell) - C(\mathbf{x}, \boldsymbol{\theta} - \tfrac{\pi}{2}\mathbf{e}_\ell)\right], \tag{15}$$

where $\mathbf{e}_\ell$ is a unit vector selecting parameter θ_ℓ. On hardware, these shifted expectations are themselves estimated from finite shots. The optimization typically employs algorithms such as Adam, stochastic gradient descent (SGD), or quantum natural gradient (Stokes et al., 2020). When measurement noise or finite sampling (shot noise) increases variance, derivative-free optimizers like SPSA (Spall, 1992) or Bayesian strategies (Ostaszewski et al., 2021), combined with techniques such as adaptive shot allocation or zero-noise extrapolation (Endo et al., 2021; Bharti et al., 2022), help stabilize convergence.

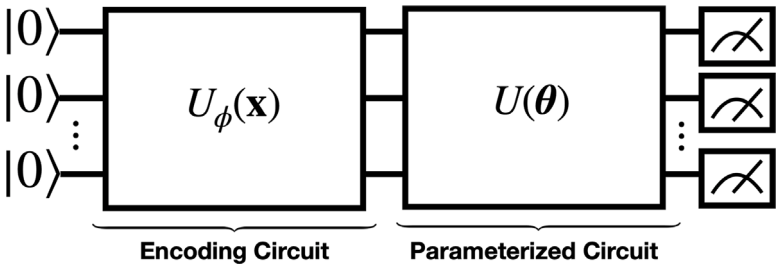

Fig. 2. Quantum Neural Network (QNN).

Remark 2.1. The shifts $\pm\frac{\pi}{2}$ apply to gates of the form $U(\theta) = e^{-i\theta G}$ when G has two distinct eigenvalues $\pm r$ and the gate is parameterized such that $r = \frac{1}{2}$ (e.g., $G = P/2$ for Pauli P). More generally, the shift depends on the generator spectrum and may require generalized shift rules.

A QNN, as illustrated in Fig. 2, is a variational quantum model in which a parameterized quantum circuit (PQC) implements a trainable mapping from an input to one or more classical outputs obtained by measurement. In the most common (NISQ) setting, a classical feature vector $\mathbf{x}$ is first embedded into an n-qubit quantum state $|\phi(\mathbf{x})\rangle$ via a data-encoding circuit $U_\phi(\mathbf{x})$, after which a layered ansatz $U(\theta)$, typically composed of alternating blocks of parameterized single-qubit rotations and entangling two-qubit gates, acts on the register. The model output is then defined through the expectation value(s) of chosen observable(s), e.g., $\langle O\rangle_{\mathbf{x},\theta} = \langle\psi(\mathbf{x},\theta)|\, O\,|\psi(\mathbf{x},\theta)\rangle$, which can be interpreted (possibly after additional classical post-processing) as a regression score, class logit, or learned feature embedding (Benedetti et al., 2019; Tacchino et al., 2019; Beer et al., 2020; Schuld and Killoran, 2019). Although the circuit evolution is unitary (and hence linear at the level of state vectors), the overall input–output map is generally nonlinear in both $\mathbf{x}$ and θ because it is mediated by the Born rule (measurement probabilities) and any subsequent classical nonlinearities. PQC-based QNN modules therefore serve as trainable primitives that constitute several quantum deep-learning architectures, including QCNNs, quantum recurrent models (e.g., QLSTMs/QGRUs), and attention-based quantum transformer variants. Fig. 3 shows the flow of information through a typical variational quantum circuit also referred to as QNN where the classical data are encoded by feature map into a quantum state which, in turn, evolves under a parameterized quantum circuit (PQC). The measurement of the state leads to a classical output inform of expectation value or prediction

Like classical neural networks, QNNs are trained by minimizing a task loss (e.g., Eq. (14)) using a hybrid quantum–classical optimization loop. However, gradient-based training can be hindered by barren plateaus, i.e., regions of the optimization landscape where gradients concentrate around zero and the variance of $\partial C/\partial\theta_\ell$ becomes exponentially small with increasing system size for broad classes of highly expressive or randomly initialized circuits (McClean et al., 2018; Cerezo et al., 2022). Practical mitigation strategies include restricting circuit depth/expressibility via shallow or problem-inspired ansätze, adopting local cost functions that depend only on small subsystems, and using structured initialization or layer-wise/warm-start training to preserve usable gradient signal (Cerezo et al., 2022; Cunningham and Zhuang, 2024; Kashif and Al-kuwari, 2023). In particular, locality-preserving constructions have been reported to improve trainability by stabilizing gradient flow and reducing sensitivity to global concentration effects (Beer et al., 2020).

Remark 2.2. Classical networks propagate gradients via differentiable activations, whereas QNNs estimate them from expectation-value differences because unitaries are non-commuting and measurement-defined. Despite this, both paradigms share an iterative gradient-descent structure, forming a direct analogue of classical backpropagation.

PQCs are not only the core building blocks of QNNs, but also the standard computational primitive underlying variational quantum algorithms (VQAs) (Cerezo et al., 2021). A VQA is a hybrid quantum–classical optimization procedure in which a parameterized quantum circuit prepares a trial state $|\psi(\theta)\rangle$, and a classical optimizer iteratively updates the circuit parameters θ to optimize an objective written as an expectation value of a Hermitian operator (observable) O, i.e., $C(\theta) = \langle\psi(\theta)|O|\psi(\theta)\rangle$. Because the objective is estimated from measurement statistics, VQAs naturally match the capabilities of near-term (NISQ) devices, where one can reliably execute only relatively shallow circuits while delegating optimization and control to classical hardware (Preskill, 2018; Cerezo et al., 2021). Two canonical examples illustrate the VQA paradigm. The variational quantum eigensolver (VQE) (Peruzzo et al., 2014) targets ground-state properties of a problem Hamiltonian H (e.g., in quantum chemistry or materials modeling) by minimizing the energy expectation $E(\theta) = \langle\psi(\theta)|H|\psi(\theta)\rangle$. This is justified by the variational principle: among all physically valid trial states, the minimum achievable expectation value upper-bounds the true ground-state energy, and increasingly expressive ansätze can systematically improve the approximation.

On the other hand, the quantum approximate optimization algorithm (QAOA) (Farhi et al., 2014) is designed for discrete/combinatorial optimization. It encodes an objective function into a cost Hamiltonian H_C and alternates between unitary evolutions generated by H_C and by a mixing Hamiltonian H_B (which promotes exploration of the solution space):

$$U_{\mathrm{QAOA}}(\boldsymbol{\gamma},\boldsymbol{\beta}) = \prod_{k=1}^{p} e^{-i\beta_k H_B}\, e^{-i\gamma_k H_C},$$

where the angles $(\boldsymbol{\gamma},\boldsymbol{\beta})$ are optimized to maximize (or minimize) the expected cost $\langle H_C\rangle$. After optimization, measuring the final state yields candidate bitstrings, with high-probability samples corresponding (ideally) to near-optimal solutions.

From the perspective of quantum deep learning, VQE and QAOA are important because they formalize the same measurement-driven, hybrid training loop used to train QNN/QDNN models: a PQC defines a differentiable hypothesis class, the learning objective is an expectation value (or a function thereof), and a classical optimizer updates circuit parameters based on measurement feedback (Cerezo et al., 2021). Moreover, many QDNN ansätze borrow architectural motifs from VQAs (e.g., alternating structured layers and hardware-efficient entangling patterns), reinforcing the role of VQAs as a conceptual and practical foundation for trainable quantum models.

2.5. Hybrid quantum-classical approach

Hybrid quantum–classical workflows form the computational backbone of modern quantum machine learning and VQAs (McClean et al., 2016; Cerezo et al., 2021). They are particularly well matched to the NISQ era (Preskill, 2018), in which programmable quantum processors provide noisy physical qubits, currently ranging from tens to more than a thousand, depending on platform, but still operate without scalable fault-tolerant error correction. NISQ devices are therefore constrained

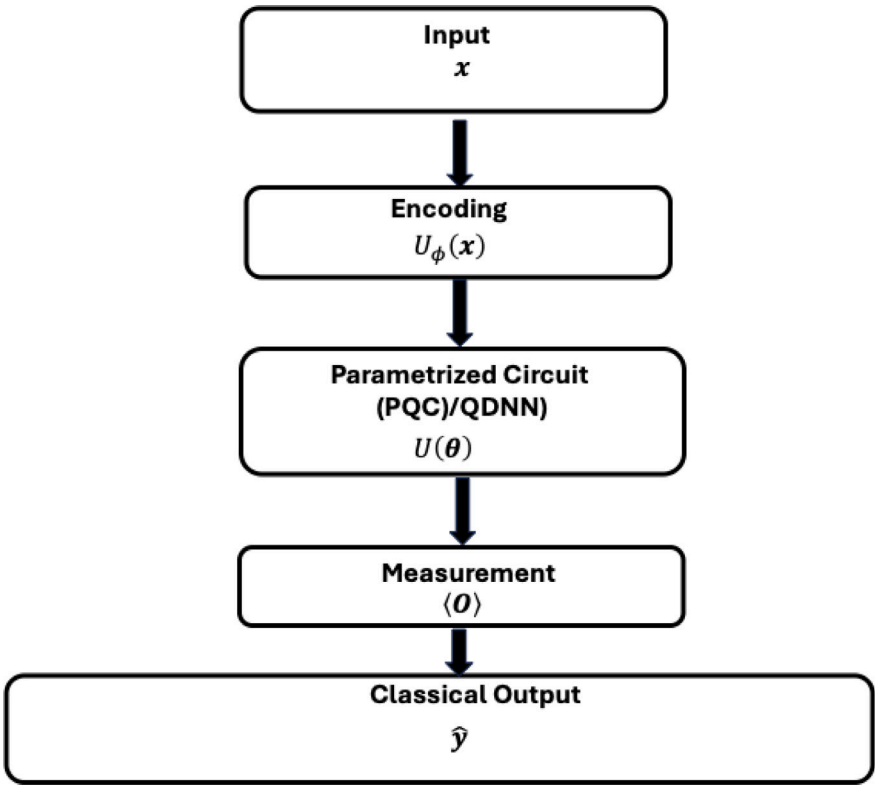

Fig. 3. Information flow through a Quantum Neural Network (QNN): Classical data are encoded by feature map into a quantum state which, in turn, evolves under a parameterized quantum circuit (PQC); measurement leads to a classical output inform of expectation value or prediction.

by decoherence, imperfect gate operations, and crosstalk. Nevertheless, they can support meaningful computation through shallow circuits executed in a hybrid loop, where quantum hardware evaluates PQCs while classical hardware performs optimization, control, and data handling (Bharti et al., 2022; Endo et al., 2021).

A typical hybrid training (or optimization) iteration alternates between classical and quantum stages (illustrated in Fig. 4):

1. **Classical proposal:** a classical optimizer selects circuit parameters θ (and, in learning from classical data, specifies the data-encoding circuit $U_\phi(\mathbf{x})$ for an input $\mathbf{x}$);
2. **Quantum evaluation:** the quantum processor executes the PQC (often $U(\theta)U_\phi(\mathbf{x})$) and performs repeated measurements;
3. **Classical update:** measurement outcomes are aggregated into estimates of expectation values $\langle O \rangle$ (from a finite number of circuit executions, or shots), a scalar objective $C(\theta)$ is computed, and θ is updated to reduce (or maximize) this objective.

This feedback loop is repeated until convergence. In practice, the shot budget controls the variance of $\langle O \rangle$ estimates, introducing a trade-off between statistical accuracy and runtime.

Hybrid execution is supported by software ecosystems that provide circuit programming, simulator/hardware backends, and (often) optimization and differentiation utilities, including PennyLane (Bergholm et al., 2018), Qiskit Machine Learning (Sahin et al., 2025), Cirq (Omole et al., 2020), and CUDA-Q (Li et al., 2025). Because hybrid training may require many circuit evaluations per iteration (especially for gradient estimation), runtime optimizations and near-hardware execution strategies such as batching, reduced host-device round trips, and tighter integration of classical control with quantum execution, are increasingly important for practical scalability (Matsuo et al., 2022; Shaydulin et al., 2023).

In QDNNs, this hybrid loop plays a role analogous to backpropagation-based training in classical deep learning: parameterized unitaries implement trainable feature transformations in Hilbert space, while effective nonlinearity typically arises from the measurement map (state to classical statistics) and any classical post-processing layers (or, where supported, mid-circuit measurement and classical feed-forward control). Circuit parameters are updated by classical optimizers using gradient estimators such as the parameter-shift rule (Mitarai et al., 2018) (or gradient-free methods when shot noise and device noise dominate). Hybrid training thus provides a practical route to implementing QCNNs, quantum recurrent models (e.g., QLSTMs/QGRUs), and attention-based quantum transformer variants on NISQ hardware, while leaving fault-tolerant implementations as a longer-term objective.

Recent studies report robustness to noisy training data in some QNN architectures and structured approaches that improve trainability relative to random deep circuits (Beer et al., 2020; K. Zhang et al., 2020). Separately, theoretical work has derived encoding-dependent generalization bounds for variational quantum models (Caro et al., 2021; Abbas et al., 2021). More recent studies further analyze how robustness and generalization are interrelated in variational quantum learning models (Berberich et al., 2024).

Importantly, while stochasticity can sometimes behave like a regularizer, device noise can also fundamentally hinder optimization: rigorous results show that realistic noise can induce barren plateaus, regions where gradients vanish exponentially with system size, thereby making training increasingly difficult as circuits scale (Cerezo et al., 2021). This trade-off is central in the NISQ setting. Consequently, NISQ-oriented QDNN designs typically emphasize shallow, hardware-efficient ansätze, local cost functions, and noise-aware training and mitigation strategies (Cerezo et al., 2021).

Finally, it is crucial to distinguish representational capacity from computational advantage. Although an n-qubit state lives in a 2^n-dimensional Hilbert space, merely accessing large superpositions does not by itself imply a speedup. Any genuine quantum advantage requires algorithmic structure, typically interference patterns and entanglement to amplify the probability of correct outcomes under measurement, and these effects can be fragile in the presence of noise and finite-sampling uncertainty (Preskill, 2018; Cerezo et al., 2021).

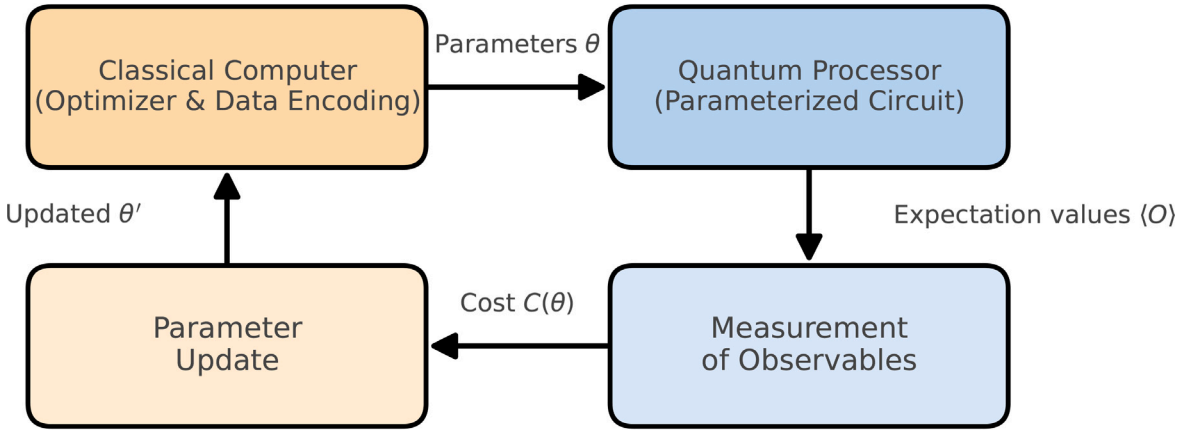

Fig. 4. Hybrid quantum–classical optimization loop. The classical processor encodes input data and proposes parameters θ, the quantum processor executes a parameterized circuit and returns measured observables $\langle O \rangle$, and the classical optimizer updates θ to minimize the cost function $C(\theta)$. This feedback loop underpins all variational algorithms and quantum deep-learning models on NISQ hardware.

Table 2
Various structures for QDNNs in PHM.

PHM setting	Typical data form	Most suitable QDNN family	Why this family is a natural choice	Main caution
Compact health indicators or handcrafted features	Low-dimensional feature vectors, short windows	Feedforward QNN	Direct nonlinear mapping from reduced health representation to SoH, RUL, or health-state label; simple and parameter-efficient	Weak modeling of long temporal dependencies
Long degradation sequences	Windowed multivariate time series	QRNN/QLSTM/QGRU	Shared recurrent quantum cell can model temporal evolution and sequential memory	High circuit-evaluation cost across timesteps; trainability under noise
Locally structured signals	Spectrograms, time–frequency patches, neighboring windows	QCNN	Locality, pooling, and parameter sharing are natural when degradation signatures are spatially or temporally localized	Qubit-layout and pooling design can be restrictive
Rare faults, anomaly detection, unsupervised monitoring	Unlabeled or imbalanced data	Quantum generative models	Density learning, reconstruction, and latent-state modeling are natural for anomaly scoring and data augmentation	PHM evidence still sparse, especially for QGANs and QBMs
Long-range multivariate dependencies	Multi-sensor sequences with operating-condition shifts	Quantum attention/transformer-style models	Context-dependent mixing is attractive when distant timesteps or sensor channels interact strongly	Higher circuit and qubit complexity; limited PHM evidence to date

3. Quantum deep neural network architectures

QDNNs integrate PQCs with classical preprocessing and post-processing components, yielding hybrid architectures that can be trained end-to-end via the variational optimization workflow introduced earlier. This section provides a concise yet comprehensive overview of the principal QDNN architectures, elucidating their operational principles and drawing parallels with their classical counterparts. Furthermore, it reviews existing prognostics-oriented studies employing each respective structure, emphasizing their applicability, relevance, and effectiveness for RUL prediction and broader prognostic objectives. Before we review various QDNN structures, we give below in Table 2 the intuition behind various structures and why they are relevant to PHM tasks.

3.1. Quantum perceptrons and feedforward quantum neural networks

In classical deep learning, a perceptron applies an affine transformation followed by a nonlinear activation function. Quantum systems, however, evolve under unitary dynamics that preserve norm and linearity. Nonlinearity emerges indirectly through the probabilistic nature of quantum measurement. The Quantum Neural Networks (QNNs), extend classical dense-layer operations into the Hilbert space of quantum states (Beer et al., 2020; Tacchino et al., 2019). A quantum perceptron can be interpreted as a parameterized unitary transformation acting on one or several qubits, followed by measurement that maps amplitudes to probabilities (Tacchino et al., 2019). Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$ denote the vector of trainable gate parameters. A general perceptron block is represented by a composite unitary

$$U(\boldsymbol{\theta}) = \prod_{l=1}^{L} U_l(\theta_l), \qquad U_l(\theta_l) = e^{-i\theta_l H_l/2}, \tag{16}$$

where H_l are Hermitian generators, typically chosen from the Pauli set $\{X, Y, Z\}$. When applied to a data-encoded input state $|\phi(\mathbf{x})\rangle$ (see Section 2.3), the circuit produces the variational state $|\psi(\mathbf{x};\boldsymbol{\theta})\rangle = U(\boldsymbol{\theta})|\phi(\mathbf{x})\rangle$. The model output is given by the expectation value of a chosen observable O, $C(\mathbf{x};\boldsymbol{\theta}) = \langle \psi(\mathbf{x};\boldsymbol{\theta})| O |\psi(\mathbf{x};\boldsymbol{\theta})\rangle$, which serves as the perceptron activation or regression score. For binary observables such as the Pauli-Z operator, the measurement outcomes $\{+1, -1\}$ yield class probabilities $p(y|\mathbf{x}) = \frac{1+C(\mathbf{x};\boldsymbol{\theta})}{2}$. A basic QNN layer consists of parameterized single-qubit rotations and entangling gates,

$$U_{\text{layer}} = \prod_i R_Z(\theta_i^{(1)}) R_Y(\theta_i^{(2)}) R_Z(\theta_i^{(3)}) \prod_{(i,j)\in E} \text{CNOT}(i,j), \tag{17}$$

where E defines the set of connected qubit pairs (entanglement topology). The CNOT(i, j) gates act between a control qubit i and target qubit j, introducing correlations among qubit amplitudes analogous to weighted interconnections in classical dense layers. Without these entangling operations, the model reduces to independent single-qubit rotations and loses expressive power. Heuristically, one may view each qubit as playing a neuron-like role: local rotations encode feature-dependent transformations, while entangling gates establish higher-order correlations (Beer et al., 2020). Training follows hybrid quantum–classical optimization, with gradients obtained through the parameter-shift rule (Section 2.4), $\frac{\partial C}{\partial \theta_l} = \frac{1}{2}\left[C(\theta_l + \frac{\pi}{2}) - C(\theta_l - \frac{\pi}{2})\right]$, and parameters

updated using classical optimizers such as Adam or stochastic gradient descent (SGD). Because expectation values are estimated from finite measurement samples, the training process is inherently stochastic, similar to mini-batch learning in classical deep networks (Benedetti et al., 2019; Cerezo et al., 2021).

Data-reuploading circuits admit rigorous approximation results in restricted settings. In particular, repeated data re-uploading enables a single-qubit circuit to realize a universal quantum classifier with classical post-processing (Pérez-Salinas et al., 2020), and single-qubit native QNNs have been shown to approximate broad classes of univariate functions through Fourier representations (Yu et al., 2022). These results are important, but they do not by themselves imply that generic deep QNNs provide universal approximation with favorable scaling for multivariate PHM tasks. More generally, certain carefully designed quantum kernels or model classes may be hard to reproduce classically under specific assumptions, but such hardness claims are highly model- and data-dependent (Havlíček et al., 2019; H.-Y. Huang et al., 2021).

It should be noted that among QDNN families, feedforward QNNs constitute the most direct quantum analogue of classical dense networks and therefore appeared first in early PHM-oriented studies. Their appeal is both conceptual and practical. Conceptually, they implement a compact parametric map from engineered health features or short temporal windows to a scalar diagnostic or prognostic output through a shallow PQC and measurement (Benedetti et al., 2019). Practically, they are compatible with the qubit and depth limitations of NISQ devices, since they do not require persistent temporal memory or repeated circuit unfolding across long sequences, unlike recurrent or attention-based models. As a result, feedforward QNNs are often used either as quantum feature extractors followed by a classical predictor, or as shallow end-to-end quantum regressors/classifiers (Silva and Droguett, 2022). At the same time, this architectural simplicity also defines their main limitation in prognostics. Most feedforward quantum models operate on precomputed statistical descriptors, health indicators, or compressed cycle-level features rather than on raw long-horizon sensor trajectories (Ngo et al., 2023). Consequently, they are well suited to compact nonlinear regression/classification on reduced representations, but less naturally suited to modeling degradation histories with long temporal dependencies. This limitation partly explains the subsequent shift toward hybrid quantum recurrent, convolutional, and attention-based structures in the PHM literature.

3.2. Quantum Recurrent Neural Networks (QRNNs) and quantum LSTMs

In classical deep learning, temporal dependencies are modeled by recurrent architectures such as the RNN, LSTM, and GRU. Their quantum counterparts QRNNs and QLSTMs, exploit superposition and entanglement to encode temporal correlations within a compact, high-dimensional Hilbert space (Chen et al., 2022; Wang et al., 2023; Tsurkan et al., 2025; Wang et al., 2024). While a classical RNN propagates a hidden vector $\mathbf{h}_t$, a quantum recurrent model evolves a joint quantum state that entangles memory and input qubits across time. In an idealized coherent quantum-memory formulation, the recurrent update at each time step t, the input $\mathbf{x}_t$ is first encoded into a quantum state $|\phi(\mathbf{x}_t)\rangle$, which interacts with the previous memory state $|\psi_{t-1}\rangle$ through a parameterized unitary transformation:

$$|\psi_t\rangle = U(\boldsymbol{\theta})\big(|\phi(\mathbf{x}_t)\rangle \otimes |\psi_{t-1}\rangle\big), \tag{18}$$

where $U(\boldsymbol{\theta})$ denotes the variational evolution operator at step t, parameterized by $\boldsymbol{\theta}$. Expectation values $\langle O_j\rangle_t = \langle\psi_t|O_j|\psi_t\rangle$ of selected observables O_j define the recurrent outputs, while the memory register, represented by the unmeasured qubits, retains temporal information through entanglement. This mechanism allows long-term dependencies to persist without explicit recurrent weight matrices. In most practical hybrid QRNN/QLSTM implementations on current hardware, however, the circuit is reinitialized at each timestep and the recurrent memory is maintained partly or entirely in classical form.

A QLSTM (Chen et al., 2022) extends this framework by introducing the integration of QNNs with the classical gating mechanism in LSTM, as illustrated in Fig. 5. The linear operations of an LSTM (originally implemented by classical deep neural networks) are replaced by parameterized quantum circuits acting on the vector $v_t = [h_{t-1} x_t]$, representing the concatenation of input x_t at time-step t and the hidden state h_{t-1} from the previous time-step $t-1$. Mathematically, the QLSTM can be described as follows,

$$f_t = \sigma\left(\mathrm{QNN}_1(v_t)\right) \tag{19a}$$

$$i_t = \sigma\left(\mathrm{QNN}_2(v_t)\right) \tag{19b}$$

$$\tilde{C}_t = \tanh\left(\mathrm{QNN}_3(v_t)\right) \tag{19c}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{C}_t \tag{19d}$$

$$o_t = \sigma\left(\mathrm{QNN}_4(v_t)\right) \tag{19e}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{19f}$$

The original QLSTM architecture can be extended by incorporating additional classical operations to handle input dimensionalities beyond the constraints of current quantum devices and simulation platforms, such as when angle encoding is applied and the input dimension exceeds the number of available qubits. In particular, classical neural networks may be employed as preprocessing modules before the VQC or as postprocessing components after the quantum circuit to improve predictive performance (Cao et al., 2023).

In many hybrid QRNN/QLSTM implementations, the quantum circuit is reinitialized at each timestep and outputs gate values, while recurrent memory is maintained classically; fully coherent quantum-memory variants exist but are more hardware-demanding (Tsurkan et al., 2025; Wang et al., 2024).

Analogously, QGRUs or QWGRUs encode gating coefficients through rotation gates whose expectation values determine the update strength. These architectures are typically parameter-efficient because the same compact PQC structure (and often shared parameters) is reused across time-steps; fully coherent variants may additionally reuse a qubit register as quantum memory, but are more hardware-demanding. Training follows the hybrid variational loop (Section 2.5), where gradients are obtained via the parameter-shift rule and optimized using Adam or SPSA. Backpropagation through time aggregates gradients across timesteps; within each timestep, PQC parameter gradients can be obtained via the parameter-shift rule (or SPSA under high shot noise) and then accumulated by the classical optimizer. To mitigate decoherence and vanishing gradients, implementations employ shallow circuits and truncated sequences. Some empirical studies like Chen et al. (2022) report faster convergence or competitive performance for QLSTM variants relative to classical LSTMs on selected tasks, although the extent to which this is attributable specifically to quantum correlations, circuit inductive bias, or implementation choices remains unresolved.

In practice, quantum analogues of LSTM/GRU gates are realized as variational sub-circuits. Gating coefficients are often derived from measured expectation values and then used (classically) to control subsequent unitaries. In coherent quantum-memory formulations, long-term information is preserved by maintaining a subset of qubits as an unmeasured memory register across timesteps; output qubits may be measured while memory qubits are re-used to sustain temporal coherence (Tsurkan et al., 2025; Wang et al., 2024).

3.3. Quantum Convolutional Neural Networks (QCNNs)

QCNNs extend classical convolutional architectures to quantum information processing by combining local unitary filters with measurement-based pooling operations (Cong et al., 2019; Pesah et al., 2021; Mari et al., 2020; Chen et al., 2023). Originally proposed by Cong et al. (2019) for classifying topological phases of quantum matter,

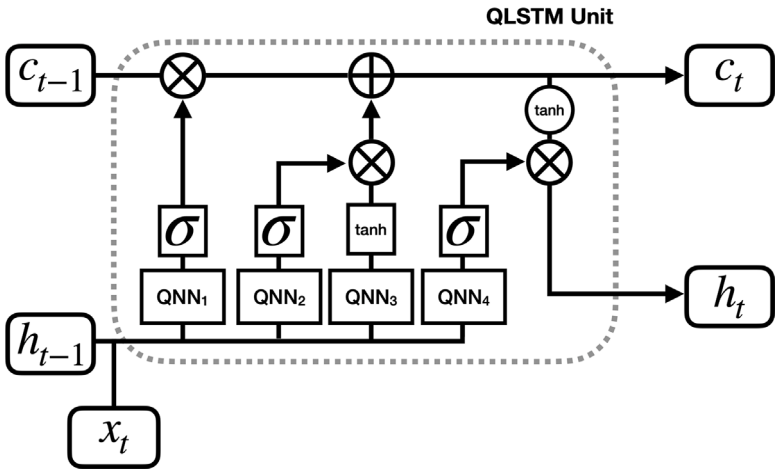

Fig. 5. Quantum Long Short-term Memory (QLSTM).

QCNNs pursue the same goals as classical CNNs: parameter reduction, spatial locality, and hierarchical feature extraction. Unlike fully connected QNNs, where all qubits interact, QCNNs restrict entanglement to neighboring qubits, yielding shallower circuits, reduced gate overhead, and improved noise tolerance on NISQ hardware. Each QCNN layer alternates between: (i) a convolution stage, where parameterized two-qubit unitaries act on adjacent qubits to extract local correlations: $U_{\text{conv}} = \prod_{(i,j)\in\mathcal{N}} U_{ij}(\boldsymbol{\theta}_{ij})$, and (ii) a pooling stage, where selected qubits are measured or traced out to reduce dimensionality: $\rho' = \text{Tr}_{\text{pooled}}\left[U_{\text{pool}}\rho U_{\text{pool}}^{\dagger}\right]$. Stacking several convolution–pooling pairs yields a hierarchical quantum representation analogous to classical feature maps, with final qubit measurements producing the network output. With hierarchical pooling and weight sharing, the number of trainable parameters can scale as $\mathcal{O}(\log n)$ with the input size n. When weights are shared across positions, the number of trainable parameters per layer remains constant, improving trainability and noise robustness on NISQ devices (Pesah et al., 2021; Cong et al., 2019). Their locality also mitigates the barren plateau problem where the gradients vanish less rapidly since cost functions depend only on small subsystems (Pesah et al., 2021). Consequently, QCNNs require fewer circuit evaluations than global variational ansätze. Modern adaptations include the quanvolutional network (Mari et al., 2020), where classical data patches are embedded into quantum circuits whose expectation values serve as classical feature maps, successfully applied to image and signal-processing tasks (Chen et al., 2023).

3.4. Quantum generative models: QBMs, QGANs, and quantum autoencoders

In classical deep learning, generative models approximate a data distribution $p_{\text{data}}(\mathbf{x})$ using a parameterized model $p_{\theta}(\mathbf{x})$. Quantum computing naturally supports generative modeling since a quantum state inherently defines a probabilistic distribution over exponentially many outcomes. Quantum mechanics generalizes this concept by encoding probabilities in the amplitudes of a quantum state,

$$|\psi_{\theta}\rangle = \sum_{\mathbf{b}\in\{0,1\}^n} \psi_{\theta}(\mathbf{b})\,|\mathbf{b}\rangle, \tag{20}$$

where $|\psi_{\theta}(\mathbf{b})|^2$ denotes the probability of observing outcome $\mathbf{b}$. Since an n-qubit system encodes 2^n amplitudes, Quantum Generative Models (QGMs) can represent complex, high-dimensional distributions compactly. The main quantum generative model paradigms of interest in PHM are: Quantum Boltzmann Machines (QBMs), Quantum Generative Adversarial Networks (QGANs), and Quantum Variational Autoencoders (QVAEs) (also referred to simply as quantum autoencoders, QAEs). These models are generative in that they learn the underlying probability distribution or latent representation of data, which can be used to detect anomalies or generate synthetic samples. As such, QGMs could address PHM challenges such as limited failure data (by synthetic data generation) and high-dimensional sensor streams (by exploiting quantum Hilbert spaces for efficient representation).

3.4.1. QBMs

Quantum Boltzmann Machines (QBMs) are quantum generalizations of energy-based Boltzmann models in which the learned distribution is defined by a thermal state of a parameterized Hamiltonian. Unlike classical Boltzmann machines, the non-commuting terms in QBMs allow the model to represent distributions shaped by coherence and entanglement, thereby enlarging the accessible hypothesis class (Amin et al., 2018; Kieferová and Wiebe, 2017). However, the additional expressivity comes at a computational cost. In general, estimating the partition function, the model expectations, and the gradients of the likelihood is harder than in classical restricted Boltzmann machines, precisely because the Hamiltonian terms do not necessarily commute. Practical QBM training therefore relies on approximations such as bound-based likelihood optimization, tomography-inspired methods, or variational Gibbs-state preparation on NISQ hardware (Amin et al., 2018; Kieferová and Wiebe, 2017; Zoufal et al., 2021). Recent theoretical work further suggests that certain visible-unit QBMs admit favorable sample-complexity guarantees under relative-entropy training, although such results currently pertain primarily to fault-tolerant rather than near-term settings (Coopmans and Benedetti, 2024). From a PHM perspective, QBMs are most naturally interpreted as probabilistic models of healthy and degraded operating regimes, where likelihood or free-energy scores could support anomaly detection, rare-regime modeling, or uncertainty-aware health-state representation.

3.4.2. QGANs

QGANs combine the adversarial training paradigm of GANs with quantum circuit models (Lloyd and Weedbrook, 2018). In a classical GAN, a generator neural network tries to produce fake data that mimic the real data distribution, while a discriminator network tries to distinguish fake from real. In a QGAN, one or both of these components are implemented with quantum circuits (or hybrid quantum–classical models). The QGAN (Lloyd and Weedbrook, 2018) adapts adversarial learning to quantum systems: a quantum generator $G(\theta_G)$ produces states approximating a target distribution, while a discriminator $D(\theta_D)$ (quantum or classical) distinguishes real from generated samples via the objective

$$\min_{\theta_G}\max_{\theta_D}\left[\mathbb{E}_{\mathbf{x}\sim p_{\text{data}}}\log D_{\theta_D}(\mathbf{x}) + \mathbb{E}_{\mathbf{z}\sim p_z}\log\left(1 - D_{\theta_D}(G_{\theta_G}(\mathbf{z}))\right)\right]. \tag{21}$$

where p_z denotes the prior distribution of latent variables and $D_{\theta_D}(\cdot)$ outputs the discriminator's estimate that a sample is drawn from the data distribution. In practice, Wasserstein and related adversarial objectives are also frequently used to improve stability. Through iterative measurement and feedback, both networks co-evolve toward equilibrium. The scientific interest lies in the fact that a quantum generator can exploit quantum superposition to generate more diverse or complex data patterns, potentially addressing issues like mode collapse in classical GANs (Lloyd and Weedbrook, 2018), plausibly leading to a form of quantum advantage by representing probability distributions that are hard to sample classically. Moreover, QGANs naturally suit scenarios with limited data: prior studies suggest that QGANs can generalize well in small-sample learning and might offer speedups in generating high-dimensional data (Z. Zhang et al., 2025).

From a modeling perspective, QGANs can operate in several regimes depending on whether the data, the generator, and the discriminator are classical or quantum. In the most PHM-relevant near-term setting, classical sensor windows or health indicators are first encoded into quantum states, after which a PQC-based generator learns to reproduce the target data distribution and the discriminator may be either a classical network or another PQC. When the training data themselves are quantum states, both the generator and discriminator can be defined entirely in the quantum domain, and discrimination may be performed through trainable measurements, state overlaps, or fidelity-based tests (Lloyd and Weedbrook, 2018; Dallaire-Demers and Killoran, 2018; Hu et al., 2019; K. Huang et al., 2021). Training still follows the alternating min–max procedure of adversarial learning, but the quantum setting introduces additional practical issues: loss estimates are shot-limited, gradients may be obtained through parameter-shift or finite-difference rules, and adversarial imbalance can be amplified by device noise, mode collapse, or vanishing gradients. Consequently, practical QGAN implementations often rely on shallow circuits, Wasserstein-type objectives, patch-wise generation, label conditioning, or multiple discriminator updates per generator update to stabilize optimization (Dallaire-Demers and Killoran, 2018; Hu et al., 2019; K. Huang et al., 2021).

Importantly, QGANs have already been demonstrated on quantum and hybrid datasets. Proof-of-principle experiments on superconducting platforms trained a quantum generator to reproduce the statistics of single-qubit mixed states with high average fidelity, while later multi-qubit implementations learned arbitrary mixed states and simple logic distributions using quantum gradients (Hu et al., 2019; K. Huang et al., 2021). QGAN-based image generation has also been demonstrated experimentally on handwritten-digit and grayscale-bar datasets, showing that adversarial quantum models can move beyond purely synthetic toy states (H.-L. Huang et al., 2021). For PHM, these developments are relevant because if degradation data are encoded into quantum states, or if quantum latent states are learned from sensor sequences, QGANs could in principle generate rare-fault quantum states, augment under-represented degradation regimes, or simulate short degradation trajectories for data-scarce monitoring problems. At present, however, such use in PHM remains prospective rather than established, and direct benchmark studies on standard PHM datasets are still missing.

3.4.3. QVAEs

In a classical variational autoencoder (VAE), an encoder $q_\phi(z|x)$ maps an input x to a latent distribution, while a decoder $p_\theta(x|z)$ reconstructs the data from the latent variable z. Training maximizes the evidence lower bound (ELBO),

$$\mathcal{L}_{\mathrm{ELBO}}(\theta,\phi;x) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - D_{\mathrm{KL}}\left(q_\phi(z|x) \parallel p(z)\right), \tag{22}$$

which balances reconstruction quality against regularization of the latent space (Goodfellow et al., 2016). Quantum extensions of this idea split into two related but conceptually distinct families. The original Quantum Variational Autoencoder (QVAE) formulation proposed by Khoshaman et al. (2019) places a QBM in the latent generative process, leading to a quantum analogue of the variational lower bound in which the latent prior is sampled from a quantum Hamiltonian rather than from a simple classical distribution. This is the closest quantum counterpart to a genuine probabilistic VAE, and its motivation is that a quantum latent prior may capture multimodal correlations more compactly than a factorized or RBM-based classical prior. A second line of work, more common in the NISQ setting, uses quantum autoencoders (QAEs) or hybrid classical-quantum autoencoders, where a parameterized quantum circuit is placed in the encoder, the decoder, or the bottleneck to learn a compressed latent representation (Romero et al., 2017; Bondarenko and Feldmann, 2020; Sakhnenko et al., 2022). Unlike a strict VAE, such models do not necessarily enforce an explicit probabilistic latent prior; instead, they are optimized for compression, reconstruction, denoising, or anomaly scoring. This distinction is important for PHM, in that full QVAE formulations remain rare, whereas hybrid QAEs are already relevant to anomaly detection, health-indicator construction, and unsupervised feature learning from high-dimensional sensor data.

3.5. Quantum transformers and attention mechanisms

In classical transformers, attention mechanisms compute similarity between query–key pairs to reweigh value representations,

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{23}$$

where $Q,K,V \in \mathbb{R}^{T\times d_k}$ are the query, key, and value matrices obtained from linear projections of the input, and d_k is the feature dimension. Quantum attention generalizes this operation through unitary evolutions that entangle the quantum states encoding Q, K, and V. Parameterized entangling unitaries can approximate attention-like, context-dependent mixing across query–key–value registers. These circuits do not implement the classical softmax dot-product exactly, but can learn analogous reweighting via interference and entanglement. Cha et al. (2022) pioneered a quantum Transformer for quantum state reconstruction. Their quantum transformer model uses a Transformer's attention mechanism to learn the probability distribution of quantum measurement outcomes, enabling accurate reconstruction of noisy quantum states. Notably, quantum transformer model outperformed prior neural-network approaches to quantum state tomography and the authors attributed its success to the ability to model quantum entanglement across the entire system, much like classical self-attention captures long-range word dependencies.

Li et al. (2024) introduced a Quantum Self-Attention Neural Network (QSANN) for Natural Language Processing (NLP) tasks, marking an early attempt to incorporate self-attention into quantum machine learning. Their QSANN replaces classical attention with a variational quantum circuit, using a Gaussian projected quantum self-attention mechanism. In text classification experiments, QSANN achieved better accuracy than both a classical self-attention network and previous syntax-based Quantum NLP models. Importantly, QSANN was designed for near-term hardware, demonstrating scalability on larger datasets and robustness to quantum noise

Shi et al. (2023) proposed a simplified quantum self-attention that forgoes explicit score matrices: a single parameterized circuit is applied to each query, key, and value register, directly producing weighted output states. This design avoids auxiliary qubits and mirrors the effect of attention entirely within the quantum circuit, yielding effective results on small-scale data. However, it requires encoding all Q–K–V pairs simultaneously, incurring a high qubit count for larger sequences. These NISQ-oriented studies demonstrate viable quantum attention loops and quantum recurrent-attention hybrids.

Quantum Transformer models have also expanded into vision and unsupervised learning. Kerenidis et al. (2024) proposed a Quantum Vision Transformer (Quantum Compound Transformer) that implements global attention via quantum linear algebra.

Table 3
Comparison between classical and QDNNs across key aspects.

Aspect	Classical DNNs	QDNNs
Representation space	Activations in $\mathbb{R}^m$ (or $\mathbb{C}^m$) with explicit parameters (weights) scaling with model size.	States in $\mathbb{C}^{2^n}$ with correlations represented via superposition/entanglement; model parameters scale with circuit ansatz and depth.
Information encoding	Inputs represented as numeric vectors or tensors.	Inputs encoded into quantum states via angle, phase, or amplitude encoding schemes.
Computation model	Layered matrix multiplications followed by nonlinear activations.	Evolution of quantum states under parameterized unitary transformations $U(\theta)$.
Nonlinearity/Activation	Explicit nonlinear functions (ReLU, sigmoid, tanh).	Implicit nonlinearity through quantum measurement and expectation values.
Training mechanism	Gradient backpropagation using the chain rule.	Hybrid optimization using parameter-shift or stochastic gradient estimation; measurements introduce sampling noise.
Convolutional architectures	Spatial feature extraction via kernel convolutions and pooling.	Local entangling unitaries emulate convolution; qubit measurement acts as pooling (QCNN).
Recurrent architectures (LSTM/GRU)	Temporal memory via gated additive updates.	Memory realized either via classical recurrent states with quantum gate subcircuits (hybrid QLSTM/QGRU) or, in coherent variants, via entanglement between input and memory qubits..
Generative models (GAN/VAE)	Learn data distributions via differentiable generator–discriminator or encoder–decoder networks.	Generator or discriminator implemented as PQCs; sampling determined by quantum measurement statistics (QGAN, QVAE).
Attention/Transformer models	Attention weights computed via query–key dot products.	Quantum self-attention via entangling unitaries acting on query–key subspaces (QTransformer).
Regularization/Noise	Explicit techniques (dropout, batch normalization, weight decay).	Noise and finite-shot sampling introduce stochasticity; they generally degrade fidelity but can sometimes act as implicit regularization.

Liu et al. (2024) investigated a quantum-entangled self-attention network by leveraging entangled states for encoding, showing quantum attention can capture complex correlations in spatio-temporal data that classical models might overlook.

Table 3 summarizes the key contrasts between classical and quantum deep networks across representative architectural families presented above. The comparison illustrates how quantum models generalize classical deep learning into exponentially large Hilbert spaces while introducing new challenges in data encoding, training stability, and interpretability.

4. Prognostics using various QDNNs

This section reviews PHM-relevant QDNN studies by architecture family. Depending on the task, the inputs may be low-dimensional health indicators, handcrafted feature windows, or multivariate temporal sequences, while the outputs may be health-state labels, SoH trajectories, RUL estimates, degradation-trend forecasts, or anomaly scores. Consequently, not all studies reviewed here share the same mathematical form: feedforward and recurrent quantum models are typically used in supervised regression or classification settings, whereas QBM-, QGAN-, and QVAE-based approaches are primarily unsupervised or generative. When a quantum model produces one or more expectation values as intermediate outputs, a convenient generic form is

$$\hat{y}(\mathbf{x};\boldsymbol{\theta}) = g\left(\mathrm{Tr}\left[O\,\rho_{\mathbf{x},\boldsymbol{\theta}}\right]\right), \tag{24}$$

where $\rho_{\mathbf{x},\boldsymbol{\theta}}$ is the encoded and processed quantum state, O is a measured observable, and $g(\cdot)$ denotes optional affine scaling or classical post-processing. For direct feedforward regressors, g may simply rescale a scalar expectation value to the target range; in other studies, g may represent a classical classifier, recurrent head, or probabilistic refinement module. This distinction is important because the literature reviewed below spans diagnosis, SoH estimation, RUL prediction, degradation forecasting, and anomaly detection rather than a single unified prognostic task.

Fig. 6 illustrates a comparative schematic of classical and quantum-enhanced PHM pipelines. As shown in Fig. 6(a) classical models process real-valued features through successive parameterized (weighted) layers leading to health states (estimation or prediction) as outputs. On the other hand, as shown in Fig. 6(b), in QDNNs, PHM features are first encoded into quantum states, propagated through trainable unitary transformations implemented by PQCs. These states evolve under the PQCs and are subsequently measured. Measurement of the resulting state yields classical outputs, typically in the form of expectation values or predicted health states.

4.1. Prognostics using quantum perceptrons and feedforward QNNs

A useful way to interpret the current feedforward-QNN literature for PHM is through three recurring design patterns. The first pattern uses a quantum feature extractor followed by a classical predictor. In this setting, the PQC acts as a nonlinear embedding module that transforms low-dimensional health descriptors into a compact latent representation, while the final regression or decision stage remains classical. The second pattern uses a direct feedforward quantum regressor/classifier, where the PQC output itself is mapped to SoH, RUL, or a health-state label. The third pattern consists of enhanced feedforward hybrids, where shallow quantum layers are combined with classical preprocessing, dimensionality reduction, or post-processing blocks to balance expressivity with implementability on near-term hardware. This classification is useful for PHM because it reveals that most early feedforward studies did not attempt full end-to-end sequence learning directly on raw sensor streams. Rather, they targeted reduced representations of degradation, such as statistical descriptors of vibration signals, cycle-level summaries of electrochemical behavior, or health indicators obtained from classical preprocessing. Feedforward QNNs are therefore best interpreted as compact nonlinear estimators operating on compressed health representations. Their main strengths lie in parameter efficiency, architectural simplicity, and ease of deployment in hybrid quantum–classical loops, whereas their main weakness is the limited ability to represent long-range temporal structure without additional recurrent or attention mechanisms.

Silva and Droguett (2022) presented one of the earliest PHM-oriented hybrid quantum–classical learning frameworks. Their study is best interpreted as a health-state diagnosis proof-of-concept rather than a direct RUL regressor: using the MFPT ball-bearing dataset, the authors segmented the vibration signal into windows, extracted five

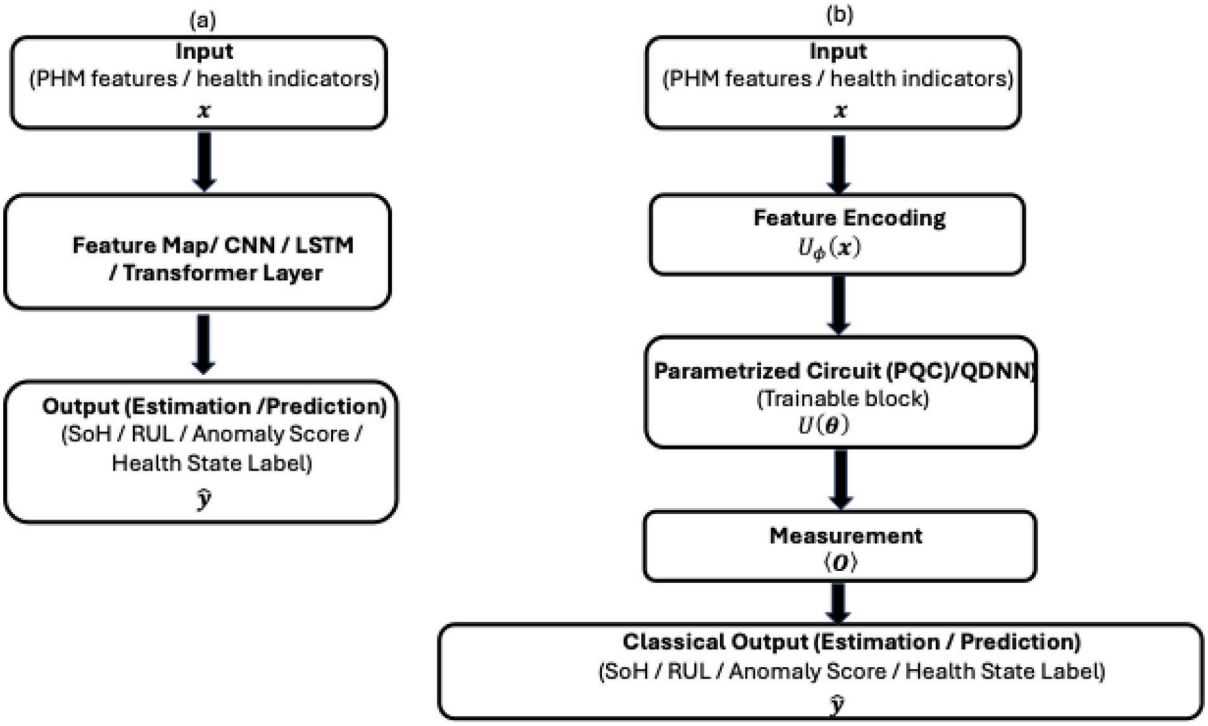

Fig. 6. Comparative schematic of classical and quantum-enhanced PHM pipelines. (a) Classical models transform real-valued features through explicit weighted layers, whereas (b) QDNNs first encode data into quantum states, evolve them through trainable unitaries, and extract outputs through measurement.

handcrafted statistical features (average, variance, maximum amplitude, peak-to-peak, and RMS), angle-encoded these features into a small PQC, and used a classical feedforward head for final classification. The reported mean testing accuracy was 97.4 ± 1.3%, showing that compact quantum feature extraction can be effective on reduced health representations. A second feedforward line is illustrated by Ngo et al. (2023), who formulated lithium-ion battery capacity degradation as a supervised regression problem from operational cycle index to remaining capacity and evaluated a QNN regressor on NASA battery datasets (B05, B06, B18, and B56). Their implementation used angle encoding, a hardware-inspired variational circuit, and a Qulacs simulator, with RMSE and MAPE as the main evaluation metrics. As such, this work should be interpreted primarily as a simulator-based proof-of-concept showing that shallow QNNs can model nonlinear battery degradation trajectories, rather than as a definitive demonstration of superiority over strong classical prognostic baselines. Liu et al. (2025) proposed a hybrid enhanced quantum neural network (EQNN) framework for rolling-bearing RUL prediction on the PRONOSTIA benchmark. The method combines dynamic complexity characteristic entropy, health-index construction through PCA-VAE, and an EQNN predictor, and is presented as a hybrid enhanced-QNN approach for improving robustness and real-time prediction accuracy under complex degradation conditions.

A recent extension of this feedforward-hybrid line is the QNN-GPR framework proposed by Lee et al. (2025) for lithium-ion battery SoH prediction. In that model, three battery health indicators are first extracted, a QNN performs nonlinear feature transformation in quantum feature space to produce intermediate SoH estimates, and a Gaussian Process Regression layer subsequently refines these estimates probabilistically. This design is noteworthy because it couples a compact quantum predictor with an uncertainty-aware classical regressor, illustrating how feedforward QNN modules can be embedded into probabilistic PHM pipelines rather than used only as standalone regressors.

4.2. Quantum recurrent models for prognostics

For most prognostics applications, multivariate time-series signals (e.g. vibration, voltage, temperature) are first normalized and segmented into temporal windows $\mathbf{x}_t$. Each window is then encoded as a quantum state $|\phi(\mathbf{x}_t)\rangle$ via angle or phase encoding on n qubits. For high-dimensional sensor inputs, a classical feature embedding (often Fourier or wavelet-based) is used to compress the input into a lower-dimensional representation before quantum encoding. Temporal dynamics across these encoded sequences are modeled by a recurrent quantum cell. Most current designs use a hybrid quantum–classical approach: a parameterized quantum circuit (PQC) produces the recurrent cell's gating values or state updates, while the memory (hidden state) is maintained (or partially maintained) in classical form between time steps. In more experimental proposals, a coherent quantum memory is used; for example, by reusing the same qubit register across timesteps to entangle past and present states; but noise and decoherence make this approach challenging on today's hardware. Almost all implementations so far are trained in a supervised manner for RUL regression: a classical output layer reads the quantum recurrent state and predicts the RUL at the current time $\widehat{\mathrm{RUL}}_t$. Some architectures also jointly learn a time-varying health index trajectory, using the quantum recurrent branch as a sequence encoder for the degradation process.

Several QRNN variants have been explored in recent studies. Wang et al. (2024) proposed an embedding-layer QLSTM with transfer learning for proton-exchange membrane fuel cell (PEMFC) stacks. A classical embedding compressed voltage sequences before feeding a QLSTM implemented as PQCs, yielding 1.3% higher mid-life RUL accuracy than a classical LSTM and demonstrating strong cross-stack transferability. Tsurkan et al. (2025) developed a Hybrid Quantum Recurrent Neural Network (HQRNN) that integrates quantum-enhanced LSTM gates with classical dense layers for turbofan-engine RUL prediction (NASA C-MAPSS). Despite fewer parameters than a deep LSTM, the HQRNN achieved approximately 5% lower RMSE/MAE and outperformed CNN, Multiple Layered Perceptrons (MLPs), and random-forest baselines.

Xiang et al. (2018) proposed a quantum-weighted gated recurrent unit neural network (QWGRUNN) for performance degradation trend prediction of rotating machinery. The method was validated on full-life vibration data of rolling bearings collected by the University of Cincinnati and was reported to achieve higher prediction accuracy, faster convergence, and lower computation time than a classical GRUNN baseline. Conceptually, the work is important because it illustrates an early recurrent quantum design in which the update mechanism is structured to retain historical degradation information while improving nonlinear approximation.

Recent battery prognostics work also extends the recurrent-hybrid direction. Soon and Soon (2025) proposed a sequential QNN-GRU architecture in which a QNN based on Pauli feature maps, variational quantum circuits, and measurement first learns nonlinear feature transformations, after which a classical GRU models the temporal evolution of battery aging. The study benchmarks the proposed model against LSTM, GRU, and hybrid deep-learning baselines, and complements the predictive analysis with SHAP-based feature attribution. This line of work is important because it illustrates a pragmatic NISQ-era design pattern for PHM, i.e., usage of the quantum module where nonlinear representation learning is most valuable, while retaining classical recurrent dynamics for temporal stability and implementation efficiency.

4.3. QCNNs for prognostics

Convolutional architectures are key to extracting spatial and local features from complex signals such as vibration spectra, voltage curves, or health indicators. While classical CNNs are widely used in RUL prediction, they often require large parameter counts to capture nonlinear degradation correlations. QCNNs present a compact alternative when degradation signatures exhibit local temporal or spatial structure. In idealized QCNN constructions with locality, pooling, and weight sharing, the number of trainable parameters can scale favorably, sometimes logarithmically with input size in theory, although the practical benefit remains architecture- and task-dependent.

Liang et al. (2025) provides one of the clearest PHM examples: the authors proposed a QCNN for stochastic lithium-ion battery SoH estimation and evaluated it on four heterogeneous datasets (CALCE, TJU, XJTU, and MIT), comprising 272 cells under multiple chemistries and operating conditions. Their model used only 39 trainable parameters and reported strong performance in terms of R^2 and RMSE, highlighting the potential of QCNNs for compact feature fusion in battery prognostics. Zaid et al. (2024) introduced a Siamese Attention-Augmented QCNN (SAA-QCNN) for aircraft-engine RUL estimation on NASA C-MAPSS. At the level of this review, the study is best interpreted as an early conference proof-of-concept demonstrating that siamese processing, quantum convolution, and attention can be combined in a prognostics setting; exact metric-level claims relative to all classical baselines should be checked directly against the original conference paper before being stated in detail. As such, QCNNs are attractive for PHM when degradation signatures exhibit local temporal or spatial structure, because locality and pooling can yield compact models with fewer trainable parameters than fully connected variational circuits.

4.4. QBMs for prognostics

To our knowledge, no peer-reviewed paper has yet demonstrated a QBM trained on a mechanical system's sensor data for PHM. However, analogous work in anomaly detection suggests the viability of QBMs for PHM. Stein et al. (2023) presented one of the first unsupervised anomaly detection approaches using a QBM focused on cybersecurity endpoint detection logs rather than a dynamical system. A QBM was implemented on a D-Wave quantum annealer to model normal versus anomalous computer events. Their results showed that the QBM could outperform a classical Restricted Boltzmann Machine in certain cases, achieving better detection quality and requiring fewer training steps. This highlights that quantum annealing-based QBMs can learn complex data distributions and potentially generalize better than classical models.

4.5. QGANs for prognostics

The literature on QGANs for PHM is still very limited. However, a few pioneering works give clue on how QGANs can be applied to anomaly/fault detection as well as prognostics contexts. Herr et al. (2021) introduced a variational quantum–classical Wasserstein GAN for anomaly detection, using a PQC-based generator and a classical discriminator. Although the application domain was credit-card fraud rather than PHM, the study is relevant because it showed that hybrid QGANs can be trained stably for anomaly-detection pipelines and achieved performance on par with classical methods in terms of F1 score.

Hammami et al. (2025) proposed a QGAN architecture for multivariate time-series anomaly detection in a network-intrusion setting, using variational quantum circuits, data re-uploading, and successive data injection. While not a PHM study, its setup is structurally close to sensor-based anomaly detection and therefore provides useful methodological insight for future PHM applications.

4.6. QVAEs for prognostics

There are only a few published PHM-adjacent studies involving quantum autoencoding ideas. Pramanik and Chandra (2021) proposed a possible quantum variational autoencoder circuit at the methodological level, but did not provide a direct PHM case study. Its relevance to PHM is therefore conceptual rather than empirical. A more concrete predictive-maintenance example is provided by Sakhnenko et al. (2022), who developed a Hybrid Classical–Quantum Autoencoder (HAE) for anomaly detection. In their framework, a classical encoder first compresses the input, a 4-qubit parameterized quantum circuit serves as the bottleneck, and a classical decoder reconstructs the signal. The model was tested both on standard benchmarks and on a gas-turbine sensor dataset relevant to predictive maintenance, where the addition of the quantum bottleneck improved precision, recall, and F1 score relative to a purely classical autoencoder. This makes hybrid QAE-type models the most mature PHM-relevant branch of the broader QVAE/QAE family at present.

4.7. Quantum transformers for prognostics

Transformers and attention mechanisms are attractive for prognostics because they model long-range dependencies across time-steps and sensor channels. In the current quantum-prognostics literature, however, the evidence base is still small and is better described as a family of attention-inspired quantum sequence models rather than mature transformer architectures. Y. Chen et al. (2020) proposed a quantum recurrent encoder–decoder neural network (QREDNN) with attention for performance degradation trend prediction of rotating machinery. Their model uses denoised fuzzy entropy extracted from vibration acceleration signals as the degradation feature and was validated on full-life rolling-bearing data from the University of Cincinnati.

4.8. Summary

Beyond the strict QDNN families above, closely related hybrid quantum–classical prognostics models are also beginning to appear. For example, Soon et al. (2025) proposed a quantum-enhanced ensemble for battery SoH forecasting under varying discharge loads, combining quantum support vector regression, a classical GRU, and Gaussian Process Regression. Although such an ensemble is not a pure QDNN in the architectural sense adopted in this review, it is relevant because

Table 4
Representative QDNN studies relevant to prognostics and PHM. Reported qubit count is included when explicitly stated in the source paper; N/R denotes not explicitly reported. In the implementation column, "no hw." indicates that no hardware-level result was reported, even if the study is conceptually hardware-compatible.

Ref.	Architecture	PHM task	Dataset	Qubits	Impl.	Primary metrics	Baseline(s)	Contribution
Silva and Droguett (2022)	Hybrid PQC feature extractor + classical classifier	Bearing health-state diagnosis	MFPT ball-bearing dataset	5	no hw.	Accuracy; cross-entropy	Proof-of-concept setting	Early PHM-oriented hybrid quantum/classical demonstration using five angle-encoded handcrafted features
Ngo et al. (2023)	Feedforward QNN regressor	Battery capacity degradation/SoH-related regression	NASA battery dataset	N/R	Qulacs sim.	RMSE; MAPE	Limited classical comparison	Shallow angle-encoded QNN models nonlinear battery capacity-aging trajectories on a simulator
Wang et al. (2024)	Embedding-layer QLSTM with transfer learning	PEMFC stack RUL prediction	IEEE PHM 2014 PEMFC (FC1, FC2)	N/R	no hw.	RUL accuracy; absolute error; transferability	Classical LSTM-family predictors	Feature compression plus transfer learning improves cross-stack PEMFC prognostics
Tsurkan et al. (2025)	HQRNN	Turbofan RUL prediction	NASA C-MAPSS	4	no hw.	RMSE; MAE; parameter count	RF, LASSO, SVM, KNR, GB, MLP, CNN, stacked LSTM	Hybrid quantum recurrent model improves over simple classical baselines while remaining below some advanced joint classical architectures
Xiang et al. (2018)	QWGRUNN	Bearing degradation-trend prediction	University of Cincinnati full-life rolling-bearing data	N/R	no hw.	Prediction accuracy; convergence speed; computation time	Classical GRUNN	Faster convergence and improved degradation-trend prediction compared with GRUNN
Liang et al. (2025)	QCNN	Battery stochastic SoH estimation	CALCE, TJU, XJTU, MIT	N/R	cloud-comp.; no hw.	R^2; RMSE; parameter count	MLP, CNN, recurrent baselines	Compact 39-parameter QCNN with strong SoH estimation across heterogeneous battery datasets
Zaid et al. (2024)	SAA-QCNN	Aircraft-engine RUL prediction	NASA C-MAPSS	N/R	no hw.	Comparative forecasting performance	Conventional deep RUL baselines	Conference proof-of-concept combining siamese quantum convolution and attention for aircraft-engine prognostics
Liu et al. (2025)	EQNN + PCA-VAE health-index construction	Bearing RUL prediction	IEEE PHM 2012/PRONOSTIA	N/R	no hw.	RUL prediction accuracy; comparative performance	Existing RUL prediction methods	Combines dynamic complexity entropy, latent health-index construction, and EQNN for bearing RUL prediction
Y. Chen et al. (2020)	QREDNN	Rotating-machinery degradation-trend prediction	University of Cincinnati rolling-bearing full-life vibration data	N/R	no hw.	Prediction accuracy; computational cost; generalization	Existing trend-prediction methods	Uses denoised fuzzy entropy and a recurrent-attention quantum architecture for compact degradation forecasting
Lee et al. (2025)	Hybrid QNN–GPR	Battery SoH estimation	NASA battery aging datasets	N/R	no hw.	Prediction error; probabilistic refinement	Classical GPR and neural predictors	Feedforward quantum feature learning followed by probabilistic regression for data-constrained SoH forecasting
Soon et al. (2025)	Quantum-enhanced ensemble (QSVR + GRU + GPR)	Battery SoH forecasting under varying loads	NASA battery datasets	N/R	no hw.	Average prediction error; correlation	Conventional ML and recurrent baselines	Illustrates expansion from pure QDNN blocks toward heterogeneous quantum–classical ensembles

it shows that PHM-oriented quantum learning is expanding beyond standalone PQC layers toward more heterogeneous quantum–classical compositions.

Table 4 summarizes the quantum and hybrid DNN structures that have been presented so far, applied to prognostics. A direct quantitative comparison across the studies in Table 4 remains difficult because the datasets differ substantially in modality, realism, scale, and target definition. For instance, NASA C-MAPSS is a widely used benchmark for multivariate RUL prediction and enables relatively standardized algorithmic comparison, but it is simulated and therefore does not fully capture all forms of sensor non-idealities and real operational variability. PRONOSTIA and XJTU-SY provide real bearing degradation trajectories and are therefore valuable for studying degradation trends and run-to-failure prediction under experimental conditions, but they remain component-specific and comparatively limited in scale. Public battery-aging datasets, such as NASA, CALCE, and Oxford, are highly relevant for SoH/RUL studies, yet they differ in chemistry, protocol, and end-of-life definitions, which complicates strict cross-paper comparison. Consequently, claims regarding superiority of one QDNN family over another should be interpreted with caution unless they are established under a common dataset, target definition, preprocessing pipeline, and baseline-tuning protocol.

It is important to note that contemporary classical models remain highly competitive and, in many cases, more accurate than current QDNNs. For battery prognostics, recent classical hybrid models such as BiGRU–Transformer and hybrid attention-based networks report strong accuracy, robustness, and generalization on public aging datasets (Jia et al., 2023; Zhao et al., 2023). In PEMFC prognostics, a recent transformer-based study reported coefficients of determination of 97% and 99% for two stacks (Gürsoy, 2025). For rolling-bearing RUL prediction, a TCN–Transformer model reduced RMSE and MAE by 14.62% and 9.26%, respectively, while improving SCORE by 13.04% (Jin et al., 2025). Even within the quantum literature, the recent HQRNN study on C-MAPSS improves over stacked LSTM and other conventional baselines, yet explicitly remains below some advanced joint classical architectures (Tsurkan et al., 2025). Therefore, the most justifiable present interpretation is that QDNNs are promising mainly for compactness, parameter efficiency, hybrid composability, and possible small-data benefits, rather than for uniform dominance in absolute predictive accuracy.

5. Standing challenges in quantum prognostics

As of today, Quantum prognostics is best viewed as a NISQ-era instantiation of variational quantum machine learning for multivariate, sequential, and often safety-critical reliability data. It is attractive due to compact parameterizations, expressive feature maps, and inherently probabilistic outputs. The latter must be weighed against bottlenecks that are more severe than in many small QML benchmarks. In PHM, encoding, trainability, noise, runtime, and evaluation protocol are tightly coupled, in that the choice of data interface shapes the effective hypothesis class; noise and finite-shot estimation perturb optimization; and weak benchmarking can obscure whether an observed gain comes from the quantum component, the classical preprocessing, or simply unequal model-selection effort (Preskill, 2018; Cerezo et al., 2021; H.-Y. Huang et al., 2021). The following subsections summarize the main existing challenges to scalable, reproducible, and practically useful quantum-enhanced PHM.

5.1. NISQ hardware constraints and system-level feasibility

Most published quantum prognostics models are designed for Noisy Intermediate-Scale Quantum (NISQ) processors, where limited coherence times, imperfect two-qubit gates, readout errors, crosstalk, and restricted qubit connectivity constrain both circuit depth and qubit layout (Preskill, 2018; Cerezo et al., 2021). These hardware-level limitations are especially acute for PHM because many target problems are sequential, in that, QRNN/QLSTM cells, attention blocks, and repeated data re-uploading can require numerous circuit evaluations per sample and repeated measurements across time. Even when an individual circuit is shallow, the cumulative cost of state preparation, repeated shots, transpilation to hardware-native gates, and classical outer-loop optimization may dominate runtime. As a result, most current studies remain (non-quantum) simulator-based or limited to few-qubit proof-of-concepts, and the gap between algorithmic feasibility and deployable industrial inference remains substantial. For operational PHM,

additional system-level constraints must also be considered, including latency of remote QPU access, calibration drift across acquisition cycles, and the need for stable performance under real-time monitoring conditions (Matsuo et al., 2022; Shaydulin et al., 2023).

5.2. Data interface and encoding bottlenecks for high-dimensional sensor streams

The data interface is arguably the central bottleneck in practical quantum prognostics. PHM datasets such as NASA C-MAPSS, PRONOSTIA, XJTU-SY, and public battery degradation datasets are high-dimensional, multivariate, and sequential; mapping them into n-qubit states therefore requires design choices that determine both expressivity and feasibility. In supervised QML, this is not a peripheral issue in that, the encoding largely defines the effective feature space and, in many settings, the resulting model is best understood through a kernel perspective (Schuld, 2021, 2022). Angle and phase encodings are attractive on NISQ hardware because they are shallow and hardware-friendly, but they scale linearly with feature dimension and therefore often require classical compression, windowing, or handcrafted health indicators before quantum processing. More compact encodings, such as amplitude encoding, are qubit-efficient in principle but can incur substantial state-preparation overhead. Data re-uploading and entangling feature maps can increase expressive power, yet they also deepen circuits and can worsen trainability (Pérez-Salinas et al., 2020; Havlíček et al., 2019). For PHM, this creates a methodological tension in that, if most of the degradation structure is already extracted by classical preprocessing, any observed gain becomes difficult to attribute specifically to the quantum component (H.-Y. Huang et al., 2021).

5.3. Training instability: barren plateaus, noise, and shot-limited gradients

Hybrid training of PQCs is sensitive to optimization pathologies. Barren plateaus (exponentially vanishing gradients) can occur for deep or poorly structured ansatze, making convergence slow or unreliable (McClean et al., 2018; Cerezo et al., 2022). Moreover, realistic noise can further suppress gradients and amplify optimization variance, while finite-shot estimation injects stochasticity into both loss values and gradient estimates. For sequence models, the challenge compounds in that backpropagation-through-time translates into repeated circuit evaluations across timesteps, multiplying shot cost and sensitivity to noise. Thus, trainability constraints often force shallow circuits and small models, limiting the ability to represent long-horizon degradation dynamics.

Although a dedicated PHM-specific optimizer benchmark for QDNNs is not yet available, empirical QML studies provide a useful practical guide. On ideal or mildly noisy simulators, Adam- or AMSGrad-type updates combined with analytic or SPSA-based gradients often converge faster and to lower losses than standard SPSA. Under shot noise and noisy backends, however, SPSA remains attractive because it requires only two objective evaluations per update, irrespective of parameter dimension, and therefore scales more favorably when circuit evaluation is expensive (Wiedmann et al., 2023; Spall, 1992). For PHM, this trade-off becomes concrete: feedforward or QCNN models trained on short windows can often exploit Adam-like optimizers on simulator-based gradients, whereas long-sequence QRNN/QLSTM models and hardware-in-the-loop execution may prefer SPSA-like methods to control the circuit-evaluation budget. Barren plateaus further complicate this choice, especially for long temporal contexts or deeper ansätze, because gradient variance can decay rapidly with system size and noise (McClean et al., 2018; Cerezo et al., 2022).

Table 5 presents optimizer selection rationale for different kinds of PHM oriented QDNNs scenarios that are generally possible.

5.4. Error mitigation overhead and the accuracy-runtime trade-off

Meaningful learning on near-term hardware often requires error-mitigation techniques such as readout-error mitigation, zero-noise extrapolation, probabilistic error cancellation, or symmetry-based post-selection (Endo et al., 2021; Cai et al., 2023). These methods can substantially improve the fidelity of measured expectation values, but they introduce a nontrivial accuracy-runtime trade-off: many mitigation strategies amplify sampling variance, increase the number of circuit executions, or require additional calibration experiments. In a PHM context, this overhead matters on two fronts. During training, mitigation can slow every loss or gradient evaluation and alter optimizer behavior through higher-variance estimates. During deployment, it can compromise the low-latency inference demanded by online monitoring, control, or edge-device settings. Consequently, the relevant question is not only whether mitigation improves predictive accuracy, but whether the improvement persists once shot budget, wall-clock time, and operational reliability are treated as part of the PHM objective (Endo et al., 2021; Cai et al., 2023).

5.5. Benchmarking and reproducibility

A major obstacle to credible claims in quantum prognostics is the absence of standardized benchmarking and reporting. Current studies vary widely in data splits, target definitions, leakage control, preprocessing pipelines, shot budgets, noise models, transpilation settings, optimizer choices, and the strength of classical baselines. This makes cross-paper comparison unreliable, even before accounting for whether a model was evaluated on noiseless simulators, noisy simulators, or real hardware. The need for careful benchmarking is not unique to PHM: recent QML benchmark studies on standard datasets have shown that outcomes can be highly sensitive to feature-map choice, hyperparameterization, and baseline fairness, and that additional quantum cost does not automatically translate into better generalization (H.-Y. Huang et al., 2021; Alvarez-Estevez, 2025). For PHM, reproducibility additionally requires domain-specific reporting items such as the run-to-failure split protocol, operating-condition shift, uncertainty metrics, shot budget, noise model, and wall-clock cost per forecast. Without such standards, it remains difficult to determine whether reported gains reflect genuine methodological progress or experimental confounding.

5.6. Defining and demonstrating "quantum advantage" in PHM tasks

Even when a quantum or hybrid model improves accuracy on a benchmark, a rigorous claim of advantage is subtle and debatable. A central difficulty in quantum prognostics is that quantum advantage is not a single, universally accepted notion. In the PHM setting, advantage may refer to at least five distinct aspects: (i) predictive advantage, i.e., lower RMSE/MAE or higher R^2 on held-out data; (ii) sample-efficiency advantage, i.e., attaining comparable predictive quality from fewer run-to-failure trajectories or fewer labeled degradation histories; (iii) model-efficiency advantage, i.e., achieving comparable performance with fewer trainable parameters or a more compact latent representation; (iv) computational advantage, i.e., lower time-to-solution, runtime, or energy cost under comparable resource accounting; and (v) reliability-oriented advantage, i.e., improved robustness, calibration, or stability under noise and domain shift. Accordingly, a claim of advantage is scientifically meaningful only when the specific notion of advantage is made explicit and the comparison is conducted against carefully tuned classical baselines under matched preprocessing, hyperparameter effort, and evaluation protocol. This distinction is particularly important in QML because improved performance on a learning task does not automatically imply a genuine quantum computational advantage. In data-driven settings, apparent gains may arise from confounding factors such as stronger

Table 5
Optimizer-selection for representative PHM-oriented QDNN scenarios.

Representative PHM scenario	Dominant bottleneck	Optimizer preference	Rationale
Short-window feedforward/QCNN regression on simulators	Smooth optimization, moderate parameter count	Adam/AMSGrad + parameter-shift	Exact or low-variance gradients are available; fast convergence is usually more important than minimizing the number of objective evaluations.
Long-sequence QRNN/QLSTM training (e.g., C-MAPSS or battery-cycle histories)	Repeated circuit evaluations across timesteps	Adam/AMSGrad with truncated sequences and shallow circuits	Stable gradient accumulation is helpful early in training, but the total circuit-evaluation cost grows quickly with sequence length and circuit size.
Noisy hardware or shot-limited execution	Measurement variance and expensive gradient estimation	SPSA or other two-point stochastic methods	Requires only two loss evaluations per update, independent of parameter dimension; often more practical when shot cost dominates.
Simulator-to-hardware transfer	Distribution shift between ideal and noisy execution	Warm-start with Adam on simulator, then SPSA-style fine-tuning on hardware	A practical compromise: low-noise pretraining accelerates convergence, while stochastic fine-tuning adapts more robustly to hardware noise.

classical preprocessing, different model capacities, unequal hyperparameter budgets, or differences in evaluation setup. Moreover, the large Hilbert-space dimension of an n-qubit state should not be conflated with practical learning advantage. In the absence of careful benchmarking, statements such as "more expressive" or "higher-dimensional" remain insufficient to establish superiority over strong classical alternatives. Viewed through this lens, the current PHM literature provides encouraging but still task-specific evidence rather than a conclusive demonstration of quantum advantage. Several reviewed studies report competitive predictive performance with reduced parameter count or compact circuit structure, which may be interpreted as early evidence of model-efficiency advantage under specific encodings and datasets. Other studies suggest improved robustness or transfer behavior, which may be viewed as tentative reliability-oriented gains. However, there is currently no broadly accepted PHM study demonstrating end-to-end computational quantum advantage under matched classical baselines and comparable hardware-level resource accounting. A rigorous future benchmark for quantum prognostics should therefore report, at minimum, the task definition, dataset split protocol, preprocessing pipeline, quantum and classical model capacities, hyperparameter search budget, shot budget, runtime or wall-clock cost, and uncertainty or robustness metrics whenever relevant.

5.7. PHM-specific requirements: uncertainty and interpretability

PHM decisions are safety and cost-critical, so point predictions alone are insufficient. Models must provide calibrated uncertainty (prediction intervals or distributions), behave robustly under operating-condition shifts, and be testable in failure analysis. Quantum models naturally produce measurement distributions, but converting these into calibrated epistemic/aleatoric uncertainty for RUL/SoH and validating reliability under shift remains largely open.

Interpretability is particularly challenging in QDNN-based PHM because the internal representation is a quantum state rather than a directly inspectable feature vector. Intermediate amplitudes are not fully observable without state tomography, measurement itself is stochastic and potentially destructive, and entanglement induces nonlocal correlations that make single-feature attribution less straightforward than in classical models. As a result, explanations based only on final expectation values can obscure how information from different sensors, timesteps, or operating regimes is actually combined inside the circuit. Recent QML work has begun to adapt classical explainability tools to this setting, including local explanation methods such as Q-LIME and feature-based or example-based interpretation strategies for QNNs (Pira and Ferrie, 2024; Tian and Yang, 2024). However, these methods have not yet been systematically translated to PHM-oriented tasks. This limitation matters more in PHM than in many generic benchmarking problems because maintenance decisions are safety- and cost-critical. A useful explanation in PHM should ideally answer at least four questions: which measurements or operating conditions drove the prediction, whether the explanation is consistent with known degradation mechanisms, how reliable the prediction is under uncertainty or distribution shift, and what action should follow operationally. Consequently, interpretability in quantum PHM is not only about feature attribution, but also about traceability to physical mechanisms, trust calibration, and compatibility with engineering failure analysis.

5.8. Resource estimation and long horizon temporal modeling

Resource estimation remains underdeveloped in quantum prognostics. A meaningful estimate should include not only qubit count and circuit depth, but also the number of measurements required per loss or gradient evaluation, the cost of gradient estimation, transpilation overhead, and the classical optimization loop (Cerezo et al., 2021). At present, the literature rarely presents scaling laws that relate these resources to dataset size, prediction horizon, or target accuracy, making near-term feasibility hard to assess beyond proof-of-concept settings. This is particularly problematic for PHM because degradation trajectories in batteries, turbines, and fuel cells often span hundreds to thousands of timesteps. Existing quantum recurrent and attention-based models therefore rely on truncated windows, compressed embeddings, or shallow temporal modules rather than full long-context sequence modeling (Bausch, 2020; Li et al., 2023). Developing temporally efficient architectures, whether through structured recurrence, local attention, tensor-network compression, or problem-informed memory mechanisms, remains a prerequisite for applying quantum models to realistic long-horizon prognostics.

6. Emerging opportunities and future directions

Despite the challenges discussed in Section 5, quantum prognostics remains a scientifically fertile area because the field is still at an early stage of co-evolution between hardware capability, model design, and PHM evaluation practice. In the near term, the most credible advances are unlikely to come from wholesale replacement of classical PHM pipelines by fully quantum ones. Rather, progress is more likely to arise from co-designed hybrid workflows in which the quantum component, the PHM task, and the benchmarking protocol are matched deliberately to one another. In the longer term, continued progress in logical qubits and fault-tolerant architectures may open qualitatively different algorithmic regimes, but the relevance of those regimes to PHM will still depend on whether they improve practically meaningful quantities such as forecasting accuracy, uncertainty quality, robustness under shift, and inference feasibility. The following directions appear especially consequential for moving quantum prognostics from proof-of-concept studies toward deployable methodology (Preskill, 2018; Cerezo et al., 2021; Bharti et al., 2022; Salinas-Camus et al., 2025).

6.1. Hardware progress and the path to fault tolerance

From a PHM standpoint, the most relevant hardware quantity is not merely the number of physical qubits, but the amount of reliable logical computation that can be executed per prediction. Recent advances in quantum error correction, including the demonstration that logical performance can improve with increasing surface-code distance and the more recent observation of below-threshold logical memories, indicate that the field is beginning to transition from isolated NISQ demonstrations toward early fault-tolerant operation (Google Quantum AI, 2023; Google Quantum AI and Collaborators, 2025). If this trajectory continues, quantum prognostics could eventually exploit circuit depths and memory mechanisms that are currently inaccessible, including deeper temporal models, repeated mid-circuit measurements with feed-forward control, and coherent latent memory across multiple sensing windows.

For PHM, such progress would be especially relevant to tasks that are currently depth-limited, such as long-context degradation forecasting, multi-sensor temporal fusion, or uncertainty-aware sequence modeling. However, the existence of better logical qubits does not automatically imply a practical advantage for PHM. What will matter is whether the additional reliable depth can be converted into better predictive quality, better-calibrated uncertainty, or more informative decision support under realistic latency and cost constraints. Thus, a key future research question is not only how hardware improves, but which PHM subroutines would genuinely benefit from the transition from shallow NISQ circuits to logically protected computation.

6.2. Hybrid design patterns

The most promising near-term direction remains principled hybridization. In practice, the relevant design question is not whether to use a quantum model somewhere in the PHM pipeline, but where to place a quantum subroutine so that it contributes the most value per qubit, per shot, and per unit runtime. Several insertion points appear particularly plausible: (i) quantum feature-map or kernel layers for nonlinear sensor fusion; (ii) compact recurrent or attention-style quantum blocks for temporally compressed sequence encoding; (iii) quantum autoencoding bottlenecks for anomaly detection and latent health-indicator construction; and (iv) probabilistic or generative quantum modules for rare-fault augmentation and density modeling. The common principle is selective use of quantum resources at the stages where classical models are either parameter-hungry, weakly calibrated, or sensitive to small-data regimes (Cerezo et al., 2021; Bharti et al., 2022; H.-Y. Huang et al., 2021).

This modular perspective is also more scientifically testable than end-to-end claims of superiority. It naturally supports ablation studies in which the quantum block is inserted, removed, or repositioned while keeping the rest of the PHM pipeline fixed. More broadly, recent task-specific demonstrations outside PHM suggest that targeted quantum subroutines are more plausible than wholesale replacement of mature classical architectures, especially when the quantum component is aligned with the structure of the underlying learning problem (Yin et al., 2025). A major research opportunity is therefore to derive theory- and hardware-guided rules for hybrid placement: for example, determining when a quantum feature map is more useful than a quantum temporal block, or when a probabilistic classical head should be preferred after a quantum latent representation.

6.3. Beyond current PQC-centric QDNNs

Although the present review focuses on PQC-based QDNNs and closely related hybrid models currently explored in PHM, several adjacent directions deserve explicit attention because they may shape the next stage of quantum-enhanced prognostics.

A first direction, evidently, is the development of fully quantum models, in which the data transformation and learning pipeline are implemented as entirely quantum operations, with minimal reliance on classical hidden layers or classical post-processing. Recent work has begun to explore fully quantum neural architectures trained through fidelity-driven objectives and quantum-native gradient surrogates (Ewald, 2025). While such approaches remain early-stage and are currently demonstrated on small benchmark tasks, they are conceptually important because they probe the limit of how much of the learning stack can migrate from hybrid to genuinely quantum execution.

A second direction in this area concerns tensor-network-based quantum learning. Tensor networks such as matrix product states, tree tensor networks, and MERA provide structured low-entanglement representations that can be mapped naturally to quantum circuits (Rieser et al., 2023). Their relevance to PHM is potentially high: long degradation trajectories and multiscale sensor dependencies often demand compression, locality, and controllable correlation structure, all of which are natural strengths of tensor-network models.

A third potential direction is quantum reinforcement learning (QRL). While QRL is not primarily a direct RUL predictor, it is highly relevant to the decision layer of PHM, where the objective is not only to estimate health but also to optimize maintenance scheduling, inspection timing, sensing policy, or control actions under uncertainty. Since PHM is increasingly viewed as a coupled design–development–decision process, QRL may eventually become relevant at the interface between prognostics and maintenance decision-making (Hu et al., 2022; Wu et al., 2025).

Finally, quantum diffusion models constitute a promising generative direction beyond QGANs. Diffusion-based generative models are attractive because they often provide more stable training than adversarial methods, and early quantum versions have already been explored for few-shot learning and image generation (Wang et al., 2025; J. Zhang et al., 2025). In the PHM context, such models could become relevant for synthetic degradation generation, rare-fault data augmentation, or learning latent health distributions under limited labeled data.

6.4. Quantum generative modeling for rare-fault data and unsupervised monitoring

Generative modeling aligns naturally with PHM needs because labeled failure data are scarce and failure modes are diverse. Although direct applications of quantum generative models to PHM remain limited, their relevance becomes clearer when viewed through the roles already played by classical generative models in PHM. In the classical literature, generative and reconstruction-based models are mainly used in three ways: (i) for anomaly detection, by learning the distribution of healthy operation and flagging low-likelihood or poorly reconstructed observations; (ii) for latent health representation learning, by compressing high-dimensional sensor measurements into lower-dimensional indicators that can later be used for degradation monitoring or prognostics; and (iii) for data augmentation, especially when failure trajectories are scarce, imbalanced, or costly to acquire. Quantum generative models inherit these same potential roles. Consequently, even when direct prognostics demonstrations are still sparse, their PHM relevance should be evaluated according to whether they can improve density modeling, reconstruction fidelity, latent-state compactness, or data efficiency under realistic sensor and hardware constraints.

Direct applications of QBMs to PHM tasks are still nascent. In the PHM context, the most plausible near-term role of a QBM is therefore not yet long-horizon RUL regression, but rather probabilistic modeling of normal and abnormal operating regimes. Such a model could, in principle, assign low probability to emerging degradation patterns before failure labels become available, thereby supporting early-warning anomaly detection or health-state discrimination. Hence, QBMs should currently be viewed as an early-stage probabilistic modeling direction for PHM rather than as a mature prognostic tool.

At the time of writing, no published study has specifically applied a QGAN to a PHM dataset. For PHM, QGANs could be used to augment scarce fault data (by generating synthetic sensor signals for rare failure modes) or to perform anomaly detection by learning the distribution of normal system behavior and identifying out-of-distribution samples. The major role of QGANs is likely to arise in rare-fault and imbalanced-data regimes. In such cases, the objective is not merely to generate plausible signals, but to improve downstream diagnosis or prognostics by enriching under-represented degradation patterns. This is directly analogous to the role played by classical adversarial models for fault-data augmentation and anomaly detection. However, until QGANs are evaluated on standard PHM datasets with explicit downstream gains in SoH/RUL estimation or fault detection, their value for prognostics should be regarded as promising but still preliminary. The potential use-cases here are numerous. One likely application is data augmentation for predictive maintenance: QGANs can generate realistic synthetic sensor signals to supplement limited real failure examples. (In classical PHM, GAN-based augmentation has been used to create additional vibration signals for rare gearbox faults, for example. A QGAN could fulfill a similar role with possibly better fidelity or requiring fewer training samples.) Indeed, recent research on QGANs for data augmentation in other fields shows promising results. Herr et al. (2021) and Hammami et al. (2025) provide early evidence that quantum generative-adversarial approaches can detect anomalies comparably (or even superiorly) to classical methods. The key contributions of these works are showing how to integrate quantum generators into anomaly detection pipelines and illustrating the parameter efficiency and stability benefits. As quantum hardware scales, one can anticipate seeing QGANs applied to PHM use-cases such as tool wear monitoring, where they could generate tool-failure signals to train predictors, or to semi-supervised fault detection where the QGAN learns the nominal equipment behavior and flags deviations. The field is poised for growth, with QGANs offering a novel way to model the probabilistic patterns of machine health data in a quantum-enhanced manner.

6.5. QBiLSTM for prognostics

Beyond unidirectional QLSTMs, Wang et al. (2023) proposed a QBiLSTM with attention to capture dependencies in both temporal directions, for sequence classification in a different domain (adverse drug reaction detection), an approach yet to be explored for prognostics tasks. By leveraging information from both directions in the sequence, this bidirectional quantum model improved detection accuracy over a classical BiLSTM. While QBiLSTM architectures have not yet been applied to prognostics, they present an immediate opportunity to model degradation processes that have temporal dependencies in both directions (for instance, when future operating conditions might influence the interpretation of past degradation). Implementing such quantum bidirectional networks for SoH forecasting especially in a multiple steps ahead manner, is yet to be explored.

On the other hand, while QBiLSTM-style models may be useful for offline degradation assessment, retrospective sequence labeling, or fixed-window health-state inference, their role in online causal RUL prediction is less direct because future context is not available at inference time. Exploring how bidirectional quantum sequence models can be adapted to sliding-window or retrospective PHM tasks therefore remains an open research direction.

6.6. QCNNs and hybrid ensembles for prognostics

Xu et al. (2024) proposed the QConvLSTM, embedding quantum convolutional filters within LSTM cells; although validated on Moving-MNIST data, the concept readily extends to RUL prediction through joint spatial–temporal encoding. When combined with quantum recurrent or classical modules, as in QCNN–QLSTM or hybrid ensembles, they unify local feature extraction, temporal modeling, and classical numerical stability, providing a compact and interpretable framework for industrial prognostics. A recent work in the classical domain (Jha et al., 2025), has shown the effectiveness of such an approach for long term prognostics.

6.7. QVAEs for prognostics

Among the quantum generative families reviewed here, autoencoder type models presently appear to be the most mature from a PHM standpoint. The reason is that their objective aligns naturally with a central PHM need: extracting compact latent structure from high-dimensional sensor data while preserving degradation-relevant information. This makes QAEs and QVAEs particularly attractive for anomaly detection, health-indicator construction, and unsupervised representation learning. By contrast, QBMs and QGANs remain at an earlier stage, with fewer direct PHM demonstrations and less standardized evaluation. However, there are only two existing works that involve QVAE for PHM. Of the two, Sakhnenko et al. (2022) demonstrates the success of a QAE on gas turbine data and is a strong proof-of-concept that quantum autoencoders are viable for real-world predictive maintenance. There are multiple potential possibilities here including extension of quantum autoencoder techniques to other PHM settings: for instance, a QVAE could be trained on normal vibration signals of a rotating machine and then used to detect anomalous vibration patterns indicating an impending bearing failure (much like classical VAEs have been used). Another possibility is using QVAE-derived latent features as health indicators for RUL prediction – i.e. compressing each time-step of a degradation process into a quantum latent vector and then modeling the trajectory of that latent vector over time to predict when it crosses a failure threshold (like in the classical unsupervised/supervised case de Beaulieu et al., 2022a, 2024). No published work has accomplished this yet, but it is an intriguing direction given that classical autoencoder-based health indices are common in RUL literature. QAEs offer a route to combine classical sensor data with quantum computing in a hybrid model that can be deployed on NISQ machines. As larger quantum computers become available, one could envision full QVAEs where both encoder and decoder are quantum, potentially encoding richer probability distributions over the latent space.

6.8. QBMs for probabilistic generative modeling

As noted earlier, no peer-reviewed study has yet established QBMs as a mature tool for machinery prognostics. Nevertheless, they remain an intellectually important direction because they address a different objective from most current PHM QDNNs: “probabilistic modeling of operating regimes” rather than direct point prediction. Since a QBM represents a Gibbs state of a trainable Hamiltonian, it can in principle capture correlated healthy and degraded behaviors through non-commuting interactions and quantum correlations that are absent from simpler classical energy-based priors (Amin et al., 2018; Kieferová and Wiebe, 2017; Zoufal et al., 2021). In PHM, this makes QBMs particularly attractive for density estimation of nominal behavior, anomaly scoring via low-likelihood regions, and latent probabilistic modeling of rare degradation states.

At the same time, the main obstacle is computational rather than conceptual. QBM training requires repeated estimation of model expectations and partition-function-related quantities, which remains expensive even in small systems and becomes substantially more demanding on near-term hardware (Kieferová and Wiebe, 2017; Zoufal et al., 2021; Coopmans and Benedetti, 2024). Consequently, the most realistic near-term PHM role of QBMs is likely as compact probabilistic models of normal versus abnormal regimes, rather than as full long-horizon predictors of RUL. A critical next step for the field would be the first dedicated PHM case studies in which QBMs are evaluated not only by detection accuracy, but also by likelihood-based criteria, calibration, and comparison against strong classical generative baselines.

6.9. Encoding and physics informed feature-map design

Encoding is not a mere implementation detail; it largely determines the effective hypothesis class of the quantum model. A major future opportunity is therefore to move beyond generic embeddings toward problem-structured feature maps tailored to PHM. For prognostics, relevant structure includes monotonic health evolution, periodicity and harmonics in rotating machinery, regime-dependent dynamics, cross-sensor coupling, and multiscale temporal signatures. This direction is consistent with broader results in QML showing that generalization and practical model behavior depend strongly on how classical data are embedded into the circuit (Caro et al., 2021; Schuld, 2022; Havlíček et al., 2019; Abbas et al., 2021). In other words, for PHM, encoding design should be viewed as part of the model itself rather than as a preliminary technical step.

A second, closely related direction is the development of physics-informed quantum models. In condition monitoring and reliability more broadly, physics-informed machine learning has already shown that incorporating governing structure, expert constraints, or mechanistic priors can improve robustness and interpretability (Wu et al., 2024). By contrast, physics-informed QML remains embryonic and has so far been explored mainly in PDE-oriented settings, where hybrid quantum–classical PINN-type models have recently begun to appear (Berger et al., 2025). Translating this philosophy to PHM suggests several concrete possibilities: embedding degradation monotonicity or operating-envelope constraints in the loss, assigning qubit groups to physically meaningful sensor subsets, or coupling observer-based latent variables with PQC layers. A general framework for physics-informed quantum prognostics is still absent, but its development could substantially improve the trustworthiness and domain relevance of future QDNN models.

6.10. Uncertainty-aware quantum prognostics

Uncertainty quantification is one of the most important open directions for making QDNNs relevant to PHM practice. In real deployment, point accuracy alone is insufficient. A useful prognostic models must provide calibrated predictive intervals or confidence bounds around the prediction (for instance, RUL prediction), distinguishing epistemic from aleatoric uncertainty, whilst remain trustworthy under distribution shift and changing operating conditions. Recent PHM reviews and benchmarks increasingly emphasize that uncertainty, together with robustness, interpretability, and feasibility, should be treated as a first-class evaluation dimension rather than an optional add-on (Salinas-Camus et al., 2025; Basora et al., 2025). For quantum-enhanced PHM, this requirement is even stronger because safety-critical maintenance decisions depend not only on the expected forecast, but also on the confidence attached to that forecast.

Quantum models create both an opportunity and a complication. On the one hand, measurement outputs are intrinsically probabilistic and may offer natural access to distribution-valued predictions. On the other hand, measurement variance, shot noise, and hardware noise are not themselves equivalent to predictive uncertainty about future degradation. A major research problem is therefore to separate useful uncertainty about asset health from variability induced by finite sampling or imperfect hardware. Emerging QML work has begun to adapt Bayesian, dropout, and ensemble-style uncertainty methods to hybrid quantum–classical models (Wendlinger et al., 2025). The next PHM-specific step is to translate these ideas into actionable outputs such as calibrated RUL intervals, anomaly-risk scores, and decision thresholds, and to evaluate them with metrics that matter operationally, including calibration error, interval coverage, sharpness, and cost-sensitive maintenance utility.

6.11. Interpretable and trustworthy quantum prognostics

A major future direction is the development of interpretable and trustworthy quantum prognostics pipelines. One promising route is to move beyond purely post-hoc explanation and instead design circuits whose structure already reflects PHM priors, for example by assigning qubit groups to sensor subsets, degradation stages, or operating regimes. Another route is to develop explanation methods directly at the level of observables, qubit groups, and entangling blocks, so that prediction-relevant interactions can be related back to physical degradation mechanisms. Such methods could be combined with counterfactual analyses (e.g., how the prediction changes if a sensor trend is removed), concept-bottleneck health indicators, or physics-constrained ansätze.

For PHM deployment, interpretability should also be coupled with uncertainty and decision relevance. A transparent QDNN should not only indicate which measurements mattered, but also communicate whether the prediction is stable under realistic perturbations, missing data, or hardware noise, and whether the explanation aligns with known failure signatures. Developing such trustworthy quantum PHM models will likely require joint advances in circuit design, explainability, calibration, and human-centered maintenance analytics.

6.12. Toward a prognostics oriented metrics for quantum advantage evaluation

Unlike the prognostics metrics as established in Saxena et al. (2008), specifically designed to assess the efficiency of prognostics process and prediction accuracy, there are currently no existing metrics to assess the so called quantum advantage in the area of quantum prognostics. As such, this presents an important possibility.

Here, we discuss, in a very general manner, some standing possibilities in this direction toward establishing such metrics that can assess the quantum advantage. In this context, we believe that a useful future benchmark for quantum prognostics should move beyond broad qualitative claims and adopt a PHM-oriented metric for quantum advantage. In our view, the most informative formulation is vector-valued rather than scalar:

$$\mathcal{A}_{\mathrm{QPHM}} = \left(A_{\mathrm{pred}}, A_{\mathrm{unc}}, A_{\mathrm{model}}, A_{\mathrm{lat}}, A_{\mathrm{rob}}, A_{\mathrm{cost}}\right), \tag{25}$$

where each component compares a quantum or hybrid model against a strong, carefully tuned classical baseline under matched preprocessing and evaluation conditions. For example, for a regression task one may define

$$A_{\mathrm{pred}} = 1 - \frac{\mathrm{RMSE}_Q}{\mathrm{RMSE}_C}, \qquad A_{\mathrm{lat}} = 1 - \frac{T_Q^{\mathrm{inf}}}{T_C^{\mathrm{inf}}}, \qquad A_{\mathrm{rob}} = 1 - \frac{\Delta_Q}{\Delta_C}, \tag{26}$$

where RMSE_Q and RMSE_C denote the quantum and classical prediction errors, T^{inf} denotes per-window inference latency (including queue overhead if cloud QPUs are used), and Δ denotes the performance degradation under a specified perturbation such as hardware noise, operating-condition shift, or missing data. Analogous definitions can be used for R^2, AUROC, calibration error, or coverage-based uncertainty metrics, depending on the task.

The remaining components should capture engineering constraints that are particularly relevant in PHM: A_{model} may reflect trainable parameters, qubit count, circuit depth, and shot budget; A_{unc} may reflect calibration quality, interval coverage, or decision-oriented risk measures; and A_{cost} may reflect QPU time, monetary cost, or energy cost per fixed number of inferences. We emphasize that Eq. (25) is best interpreted as a reporting framework or benchmark dashboard, not as a single universal number. Different PHM applications may assign different priorities to accuracy, latency, robustness, or interpretability, and the benchmark should reflect those priorities explicitly.

Table 6 provides in a concise manner a qualitative mapping between the main standing challenges in Section 5 and the corresponding future directions in Section 6.

Table 6
Mapping between the main standing challenges in Section 5 and the corresponding future directions in Section 6.

Challenge in Section 5	Matching future direction in Section 6	Example technical lever
NISQ hardware constraints and feasibility	Hardware progress and the path to fault tolerance	Better qubit quality, deeper executable circuits, lower calibration drift
Encoding bottlenecks for high-dimensional sensor streams	Encoding and physics-informed feature-map design	Problem-structured feature maps, compression with degradation priors
Training instability, barren plateaus, and shot-limited gradients	Hybrid design patterns	Shallow circuits, local costs, module placement, optimizer choice
Error-mitigation overhead	Hardware progress and hybrid design patterns	Runtime-aware mitigation, selective mitigation, simulator-to-hardware transfer
Benchmarking and reproducibility	PHM-oriented scorecard for quantum advantage	Standardized reporting of qubits, depth, shots, runtime, and robustness
Uncertainty, domain shift, and trustworthiness	Uncertainty-aware and interpretable quantum PHM	Calibration metrics, shift tests, explainability, decision relevance
Long-horizon temporal modeling and resource estimation	QCNN/QLSTM hybrids, tensor networks, and quantum transformers	Structured sequence encoders, low-entanglement compression, sparse attention

7. Conclusions

This work presents a tutorial-style review of Quantum Deep Neural Networks (QDNNs) for Prognostics and Health Management (PHM), structured around the mathematical primitives of hybrid quantum learning and the main architecture families that currently appear in PHM: feedforward QNNs, quantum recurrent models, QCNNs, quantum generative models, and attention-based quantum architectures. The original contribution of the review lies in its tutorial style, cross-asset, architecture-centered perspective. Unlike general QML surveys, the present work jointly connects quantum learning fundamentals, QDNN architectural design, and PHM tasks across bearings, batteries, PEMFCs, turbofan engines, and anomaly-detection settings.

Clearly, QDNNs are scientifically relevant to PHM because their structured quantum feature maps, entangling transformations, and measurement-based outputs provide compact ways of modeling nonlinear couplings, temporal dynamics, and probabilistic outputs. However, the current evidence is promising but not yet definitive, in that, several studies report competitive accuracy and reduced parameter count, yet the literature remains heterogeneous, dominated by simulator-based experiments, and highly sensitive to encoding choice, classical preprocessing, and benchmarking protocol. Further, systematic quantum advantage in PHM has not yet been established. In many benchmark settings, strong classical recurrent, convolutional, and transformer-based models still achieve equal or better predictive accuracy.

The present review also has clear scope limitations. It focuses primarily on PQC-based QDNNs and closely related hybrid architectures; adjacent QML paradigms such as pure quantum kernel methods, annealing-based approaches, and quantum-inspired optimization were discussed only when directly relevant to PHM. Moreover, the available evidence base is still too small and heterogeneous to support a formal meta-analysis. Further, this work identifies the standing scientific challenges and proposes future research directions in a precise manner. Future work should involve physics-aware encoding design, rigorous comparison against strong classical baselines, reproducible hardware-in-the-loop studies, uncertainty-aware evaluation, and resource accounting in terms of qubits, circuit depth, shot budget, and runtime. Under these conditions, QDNNs may evolve from promising exploratory models into practically useful components of next-generation PHM systems.

As quantum hardware continues to advance, QDNNs are expected to evolve into practical tools for predictive maintenance, bridging the gap between quantum computation and classical PHM community.

CRediT authorship contribution statement

Mayank Shekhar Jha: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Sameul Yen-Chi Chen:** Writing – review & editing, Writing – original draft. **Chetan Kulkarni:** Validation. **Joongheon Kim:** Writing – review & editing.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the first author used ChatGPT-5 in order to improve linguistic quality as well as for making the content concise at certain places. After using this tool/service, the first author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

Abbas, A., Ambainis, A., Augustino, B., Bärtschi, A., Buhrman, H., Coffrin, C., Cortiana, G., Dunjko, V., Egger, D.J., Elmegreen, B.G., et al., 2024. Challenges and opportunities in quantum optimization. Nat. Rev. Phys. 718–735.

Abbas, A., et al., 2021. The power of quantum neural networks. Nat. Comput. Sci. 1, 403–409.

Alvarez-Estevez, D., 2025. Benchmarking quantum machine learning kernel training for classification tasks. IEEE Trans. Quantum Eng. 6, 1–15. http://dx.doi.org/10.1109/TQE.2025.3541882.

Amin, M.H., Andriyash, E., Rolfe, J., Kulchytskyy, B., Melko, R., 2018. Quantum Boltzmann machine. Phys. Rev. X 8 (2), 021050.

An, Q., Tao, Z., Xu, X., El Mansori, M., Chen, M., 2020. A data-driven model for milling tool remaining useful life prediction with convolutional and stacked LSTM network. Measurement 154, 107461.

Babu, G.S., Zhao, P., Li, X.-L., 2016. Deep convolutional neural network based regression approach for estimation of remaining useful life. In: International Conference on Database Systems for Advanced Applications. pp. 214–228.

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Basora, L., Viens, A., Arias Chao, M., Olive, X., 2025. A benchmark on uncertainty quantification for deep learning prognostics. Reliab. Eng. Syst. Saf. 253, 110513. http://dx.doi.org/10.1016/j.ress.2024.110513.

Bausch, J., 2020. Recurrent quantum neural networks. In: Advances in Neural Information Processing Systems. Vol. 33.

Beer, K., Bondarenko, D., Farrelly, T., Osborne, T.J., Salzmann, R., Scheiermann, D., Wolf, R., 2020. Training deep quantum neural networks. Nat. Commun. 11 (1), 808.

Benedetti, M., et al., 2019. Parameterized quantum circuits as machine learning models. Quantum Sci. Technol. 4 (4), 043001.

Berberich, J., Fink, D., Pranjić, D., Tutschku, C., Holm, C., 2024. Training robust and generalizable quantum models. Phys. Rev. Res. 6 (4), 043326.

Berger, S., Hosters, N., Möller, M., 2025. Trainable embedding quantum physics informed neural networks for solving nonlinear PDEs. Sci. Rep. 15, 18823. http://dx.doi.org/10.1038/s41598-025-02959-z.

Bergholm, V., Izaac, J., Schuld, M., Gogolin, C., Ahmed, S., Ajith, V., Alam, M.S., Alonso-Linaje, G., AkashNarayanan, B., Asadi, A., et al., 2018. Pennylane: Automatic differentiation of hybrid quantum-classical computations. arXiv preprint arXiv:1811.04968.

Bharti, K., et al., 2022. Noisy intermediate-scale quantum (NISQ) algorithms. Rev. Modern Phys. 94 (1), 015004.

Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., Lloyd, S., 2017. Quantum machine learning. Nature 549 (7671), 195–202.

Bondarenko, D., Feldmann, P., 2020. Quantum autoencoders to denoise quantum data. Phys. Rev. Lett. 124 (13), 130502. http://dx.doi.org/10.1103/PhysRevLett.124.130502.

Cai, Z., Babbush, R., Benjamin, S.C., Endo, S., Huggins, W.J., Li, Y., McClean, J.R., O'Brien, T.E., 2023. Quantum error mitigation. Rev. Modern Phys. 95 (4), 045005. http://dx.doi.org/10.1103/RevModPhys.95.045005.

Cao, Y., Zhou, X., Fei, X., Zhao, H., Liu, W., Zhao, J., 2023. Linear-layer-enhanced quantum long short-term memory for carbon price forecasting. Quantum Mach. Intell. 5 (2), 26.

Caro, M.C., Gil-Fuster, E., Meyer, J.J., Eisert, J., Sweke, R., 2021. Encoding-dependent generalization bounds for parametrized quantum circuits. Quantum 5, 582.

Cerezo, M., et al., 2021. Variational quantum algorithms. Nat. Rev. Phys. 3 (9), 625–644.

Cerezo, M., et al., 2022. Cost function dependent barren plateaus in shallow parametrized quantum circuits. Nat. Commun. 13 (1), 1791.

Cha, P., Ginsparg, P., Wu, F., Carrasquilla, J., McMahon, P.L., Kim, E.-A., 2022. Attention-based quantum tomography. Mach. Learn.: Sci. Technol. 3 (1), 01LT01.

Chelouati, M., Jha, M.S., Galeotta, M., Theilliol, D., 2021. Remaining useful life prediction for liquid propulsion rocket engine combustion chamber. In: 2021 5th International Conference on Control and Fault-Tolerant Systems. SysTol, pp. 225–230.

Chen, G., Chen, Q., Long, S., Zhu, W., Yuan, Z., Wu, Y., 2023. Quantum convolutional neural network for image classification. Pattern Anal. Appl. 26 (2), 655–667.

Chen, S.Y.-C., Tseng, H.-H., Lin, H.-Y., Yoo, S., 2025a. Learning to measure quantum neural networks. In: 2025 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops. ICASSPW, IEEE, pp. 1–5.

Chen, S.Y.-C., Tseng, H.-H., Lin, H.-Y., Yoo, S., 2025b. Learning to program quantum measurements for machine learning. In: 2025 IEEE International Conference on Quantum Computing and Engineering. QCE, Vol. 01, pp. 1826–1836. http://dx.doi.org/10.1109/QCE65121.2025.00200.

Chen, Z., Wu, M., Zhao, R., Guretno, F., Yan, R., Li, X., 2020. Machine remaining useful life prediction via an attention-based deep learning approach. IEEE Trans. Ind. Electron. 68 (3), 2521–2531.

Chen, S.Y.-C., Yoo, S., Fang, Y.-L.L., 2022. Quantum long short-term memory. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8622–8626.

Chen, Y., Zhao, W., Zhou, W., Guo, Y., 2020. Quantum recurrent encoder–decoder neural network for performance degradation trend prediction of rotating machinery. Knowl.-Based Syst. 204, 106232. http://dx.doi.org/10.1016/j.knosys.2020.106232.

Cong, I., Choi, S., Lukin, M.D., 2019. Quantum convolutional neural networks. Nat. Phys. 15, 1273–1278.

Coopmans, L., Benedetti, M., 2024. On the sample complexity of quantum Boltzmann machine learning. Commun. Phys. 7, 274. http://dx.doi.org/10.1038/s42005-024-01763-x.

Cunningham, J., Zhuang, J., 2024. Investigating and mitigating barren plateaus in variational quantum circuits: A survey. arXiv preprint arXiv:2407.17706.

da Costa, P.R.d.O., Akcay, A., Zhang, Y., Kaymak, U., 2019. Attention and long short-term memory network for remaining useful lifetime predictions of turbofan engine degradation. Int. J. Progn. Health Manag. 10, 034.

Dallaire-Demers, P.-L., Killoran, N., 2018. Quantum generative adversarial networks. Phys. Rev. A 98 (1), 012324. http://dx.doi.org/10.1103/PhysRevA.98.012324.

de Beaulieu, M.H., Jha, M.S., Garnier, H., Cerbah, F., 2022a. Unsupervised prognostics based on deep virtual health index prediction. In: PHM Society European Conference. Vol. 7, pp. 193–199.

de Beaulieu, M.H., Jha, M.S., Garnier, H., Cerbah, F., 2022b. Unsupervised remaining useful life prediction through long range health index estimation based on encoders-decoders. IFAC Pap. 55 (6), 718–723.

de Beaulieu, M.H., Jha, M.S., Garnier, H., Cerbah, F., 2024. Remaining useful life prediction based on physics-informed data augmentation. Reliab. Eng. Syst. Saf. 252, 110451.

Du, Y., Hsieh, M.-H., Liu, T., Tao, D., 2020. Expressive power of parameterized quantum circuits. Phys. Rev. Res. 2 (3), 033125. http://dx.doi.org/10.1103/PhysRevResearch.2.033125.

Dunjko, V., Briegel, H.J., 2018. Machine learning and artificial intelligence in the quantum domain: a review of recent progress. Rep. Progr. Phys. 81 (7), 074001. http://dx.doi.org/10.1088/1361-6633/aab406.

Endo, S., Cai, Z., Benjamin, S.C., Yuan, X., 2021. Hybrid quantum–classical algorithms and quantum error mitigation. J. Phys. Soc. Japan 90 (3), 032001. http://dx.doi.org/10.7566/JPSJ.90.032001.

Ewald, D., 2025. The proposal of a fully quantum neural network and fidelity-driven training using directional gradients for multi-class classification. Electronics 14 (11), 2189. http://dx.doi.org/10.3390/electronics14112189.

Farhi, E., Goldstone, J., Gutmann, S., 2014. A quantum approximate optimization algorithm. arXiv preprint arXiv:1411.4028.

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., Ducoffe, M., 2020. Potential, challenges and future directions for deep learning in prognostics and health management applications. Eng. Appl. Artif. Intell. 92, 103678.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning, vol. 1, MIT press Cambridge.

Google Quantum AI, 2023. Suppressing quantum errors by scaling a surface code logical qubit. Nature 614 (7949), 676–681. http://dx.doi.org/10.1038/s41586-022-05434-1.

Google Quantum AI and Collaborators, 2025. Quantum error correction below the surface code threshold. Nature 638 (8052), 920–926. http://dx.doi.org/10.1038/s41586-024-08449-y.

Grover, L.K., 1996. A fast quantum mechanical algorithm for database search. In: Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing. pp. 212–219.

Gürsoy, M.İ., 2025. Short and long-term prognostics of the remaining useful life of a proton exchange membrane fuel cell using deep learning and transformer model. Int. J. Hydrog. Energy 143, 1120–1132. http://dx.doi.org/10.1016/j.ijhydene.2024.12.512.

Hammami, W., Cherkaoui, S., Wang, S., 2025. Enhancing network anomaly detection with quantum GANs and successive data injection for multivariate time series. In: 2025 International Wireless Communications and Mobile Computing. IWCMC, IEEE, pp. 1667–1672.

Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., Gambetta, J.M., 2019. Supervised learning with quantum-enhanced feature spaces. Nature 567, 209–212. http://dx.doi.org/10.1038/s41586-019-0980-2.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Heinosaari, T., Ziman, M., 2010. An Invitation to Quantum Tomography. Springer, http://dx.doi.org/10.1007/978-3-642-14001-9.

Herr, D., Obert, B., Rosenkranz, M., 2021. Anomaly detection with variational quantum generative adversarial networks. Quantum Sci. Technol. 6 (4), 045004.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Hu, Y., Miao, X., Si, Y., Pan, E., Zio, E., 2022. Prognostics and health management: A review from the perspectives of design, development and decision. Reliab. Eng. Syst. Saf. 217, 108063. http://dx.doi.org/10.1016/j.ress.2021.108063.

Hu, L., Wu, S.-H., Cai, W., Ma, Y., Mu, X., Xu, Y., Wang, H., Song, Y., Deng, D.-L., Zou, C.-L., Sun, L., 2019. Quantum generative adversarial learning in a superconducting quantum circuit. Sci. Adv. 5 (1), eaav2761. http://dx.doi.org/10.1126/sciadv.aav2761.

Huang, H.-Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H., McClean, J.R., 2021. Power of data in quantum machine learning. Nat. Commun. 12 (1), 2631.

Huang, H.-L., Du, Y., Gong, M., Zhao, Y., Wu, Y., Wang, C., Li, S., Liang, F., Lin, J., Xu, Y., Yang, R., Liu, T., Hsieh, M.-H., Deng, H., Rong, H., Peng, C.-Z., Lu, C.-Y., Chen, Y.-A., Tao, D., Zhu, X., Pan, J.-W., 2021. Experimental quantum generative adversarial networks for image generation. Phys. Rev. Appl. 16 (2), 024051. http://dx.doi.org/10.1103/PhysRevApplied.16.024051.

Huang, K., Wang, Z.-A., Song, C., et al., 2021. Quantum generative adversarial networks with multiple superconducting qubits. Npj Quantum Inf. 7, 165. http://dx.doi.org/10.1038/s41534-021-00503-1.

Jardine, A.K., Lin, D., Banjevic, D., 2006. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mech. Syst. Signal Process. 20 (7), 1483–1510.

Jha, M.S., Bressel, M., Ould-Bouamama, B., Dauphin-Tanguy, G., 2016a. Particle filter based hybrid prognostics of proton exchange membrane fuel cell in bond graph framework. Comput. Chem. Eng. 95, 216–230.

Jha, M.S., Dauphin-Tanguy, G., Ould-Bouamama, B., 2016b. Particle filter based hybrid prognostics for health monitoring of uncertain systems in bond graph framework. Mech. Syst. Signal Process. 75, 301–329.

Jha, M., Theilliol, D., Belleoud, P., Oriol, S., 2025. Deep learning based prognostics of nonlinear systems under degradation in closed-loop. In: 6th International Conference on Control and Fault-Tolerant Systems. SysTol'25.

Jia, C., Tian, Y., Shi, Y., Jia, J., Wen, J., Zeng, J., 2023. State of health prediction of lithium-ion batteries based on bidirectional gated recurrent unit and transformer. Energy 285, 129401. http://dx.doi.org/10.1016/j.energy.2023.129401.

Jin, X., Ji, Y., Li, S., Lv, K., Xu, J., Jiang, H., Fu, S., 2025. Remaining useful life prediction for rolling bearings based on TCN–Transformer networks using vibration signals. Sensors 25 (11), 3571. http://dx.doi.org/10.3390/s25113571.

Kanso, S., Jha, M.S., Galeotta, M., Theilliol, D., 2022. Remaining useful life prediction with uncertainty quantification of liquid propulsion rocket engine combustion chamber. IFAC Pap. 55 (6), 96–101.

Kashif, M., Al-kuwari, S., 2023. ResQNets: A residual approach for mitigating barren plateaus in quantum neural networks. EPJ Quantum Technol. 10 (1), 16. http://dx.doi.org/10.1140/epjqt/s40507-023-00216-8.

Kerenidis, I., Mathur, N., Landman, J., Strahm, M., Li, Y.Y., et al., 2024. Quantum vision transformers. Quantum 8, 1265.

Khakpour Komarsofla, M., Kiani, A., 2026. Quantum machine learning approaches to state-of-health prediction and optimization in energy storage devices. J. Energy Storage 153, 120939. http://dx.doi.org/10.1016/j.est.2026.120939.

Khoshaman, A., Vinci, W., Denis, B., Andriyash, E., Sadeghi, H., Amin, M.H., 2019. Quantum variational autoencoder. Quantum Sci. Technol. 4 (1), 014001.

Kieferová, M., Wiebe, N., 2017. Tomography and generative training with quantum Boltzmann machines. Phys. Rev. A 96 (6), 062327. http://dx.doi.org/10.1103/PhysRevA.96.062327.

Lee, P.W.L., Soon, K.L., Soon, L.T., 2025. Quantum neural network and Gaussian process framework for lithium battery state of health prediction. Energy Storage 7 (6), e70262. http://dx.doi.org/10.1002/est2.70262.

Li, X., Ding, Q., Sun, J.-Q., 2018. Remaining useful life estimation in prognostics using deep convolution neural networks. Reliab. Eng. Syst. Saf. 172, 1–11.

Li, J., Rubinshtein, E., Martonosi, M., 2025. Statistical assertions for debugging quantum circuits and states in CUDA-Q. arXiv preprint arXiv:2507.16255.

Li, Y., Wang, Z., Han, R., Shi, S., Li, J., Shang, R., Zheng, H., Zhong, G., Gu, Y., 2023. Quantum recurrent neural networks for sequential learning. Neural Netw. 166, 148–161. http://dx.doi.org/10.1016/j.neunet.2023.07.003.

Li, G., Zhao, X., Wang, X., 2024. Quantum self-attention neural networks for text classification. Sci. China Inf. Sci. 67 (4), 142501.

Liang, C., Tao, S., Huang, X., Wang, Y., Xia, B., Zhang, X., 2025. Stochastic state of health estimation for lithium-ion batteries with automated feature fusion using quantum convolutional neural network. J. Energy Chem. 106, 205–219.

Liu, R., Meng, G., Yang, B., Sun, C., Chen, X., 2017. Dislocated time series convolutional neural architecture: An intelligent fault diagnosis approach for electric machine. IEEE Trans. Ind. Inform. 13 (3), 1310–1320.

Liu, H., Yuan, T., Zhang, X., Xu, H., 2024. Quantum entanglement and self-attention neural networks: An investigation into passengers and stops characteristics for optimal bus stop localization. Inf. Fusion 112, 102527.

Liu, Y., Zhang, J., Li, M., Zhao, Y., 2025. Method for predicting remaining useful life of rolling bearings based on dynamic complexity entropy and quantum neural networks. Eng. Fail. Anal. 170, 109315. http://dx.doi.org/10.1016/j.engfailanal.2025.109315.

Lloyd, S., Weedbrook, C., 2018. Quantum generative adversarial learning. Phys. Rev. Lett. 121 (4), 040502.

Mari, A., Bromley, T.R., Killoran, N., 2020. Transfer learning in hybrid quantum-classical neural networks. Quantum Sci. Technol. 5 (4), 044003.

Matsuo, S., et al., 2022. Low-latency classical–quantum interface for real-time feedback control of superconducting qubits. Phys. Rev. Appl. 17 (6), 064045. http://dx.doi.org/10.1103/PhysRevApplied.17.064045.

McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H., 2018. Barren plateaus in quantum neural network training landscapes. Nat. Commun. 9 (1), 4812.

McClean, J.R., Romero, J., Babbush, R., Aspuru-Guzik, A., 2016. The theory of variational hybrid quantum-classical algorithms. New J. Phys. 18 (2), 023023. http://dx.doi.org/10.1088/1367-2630/18/2/023023.

Mitarai, K., Negoro, M., Kitagawa, M., Fujii, K., 2018. Quantum circuit learning. Phys. Rev. A 98 (3), 032309. http://dx.doi.org/10.1103/PhysRevA.98.032309.

Ngo, A.P., Le, N., Nguyen, H.T., Eroglu, A., Nguyen, D.T., 2023. A quantum neural network regression for modeling lithium-ion battery capacity degradation. In: 2023 IEEE Green Technologies Conference. GreenTech, IEEE, pp. 164–168.

Nielsen, M., Chuang, I., 2010. Quantum Computation and Quantum Information. Cambridge University Press.

Omole, V., Tyagi, A., Carey, C., Hanus, A., Hancock, A., Garcia, A., Shedenhelm, J., 2020. Cirq: A python framework for creating, editing, and invoking quantum circuits. URL https://github.com/quantumlib/Cirq.

Ostaszewski, M., Grant, E., Benedetti, M., 2021. Structure optimization for parameterized quantum circuits. Quantum 5, 391. http://dx.doi.org/10.22331/q-2021-02-08-391.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., Dean, J., 2021. Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., Latorre, J.I., 2020. Data re-uploading for a universal quantum classifier. Quantum 4, 226. http://dx.doi.org/10.22331/q-2020-02-06-226.

Peruzzo, A., McClean, J., Shadbolt, P., Yung, M.-H., Zhou, X.-Q., Love, P.J., Aspuru-Guzik, A., O'Brien, J.L., 2014. A variational eigenvalue solver on a photonic quantum processor. Nat. Commun. 5, 4213. http://dx.doi.org/10.1038/ncomms5213.

Pesah, A., Cerezo, M., Wang, S., Volkoff, T., Sornborger, A.T., Coles, P.J., 2021. Absence of barren plateaus in quantum convolutional neural networks. Phys. Rev. X 11 (4), 041011.

Pira, L., Ferrie, C., 2024. On the interpretability of quantum neural networks. Quantum Mach. Intell. 6, 52. http://dx.doi.org/10.1007/s42484-024-00191-y.

Pramanik, S., Chandra, M.G., 2021. On a possible quantum variational autoencoder circuit. In: 2021 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1–6.

Preskill, J., 2018. Quantum computing in the NISQ era and beyond. Quantum 2, 79.

Rieser, H.-M., Köster, F., Raulf, A.P., 2023. Tensor networks for quantum machine learning. Proc. R. Soc. A 479 (2275), 20230218. http://dx.doi.org/10.1098/rspa.2023.0218.

Romero, J., Olson, J.P., Aspuru-Guzik, A., 2017. Quantum autoencoders for efficient compression of quantum data. Quantum Sci. Technol. 2 (4), 045001. http://dx.doi.org/10.1088/2058-9565/aa8072.

Sahin, M.E., Altamura, E., Wallis, O., Wood, S.P., Dekusar, A., Millar, D.A., Imamichi, T., Matsuo, A., Mensa, S., 2025. Qiskit machine learning: an open-source library for quantum machine learning tasks at scale on quantum hardware and classical simulators. arXiv preprint arXiv:2505.17756.

Sakhnenko, A., O'Meara, C., Ghosh, K.J., Mendl, C.B., Cortiana, G., Bernabé-Moreno, J., 2022. Hybrid classical-quantum autoencoder for anomaly detection. Quantum Mach. Intell. 4 (2), 27.

Salinas-Camus, M., Goebel, K., Eleftheroglou, N., 2025. A comprehensive review and evaluation framework for data-driven prognostics: Uncertainty, robustness, interpretability, and feasibility. Mech. Syst. Signal Process. 237, 113015. http://dx.doi.org/10.1016/j.ymssp.2025.113015.

Saxena, A., Celaya, J., Balaban, E., Goebel, K., Saha, B., Saha, S., Schwabacher, M., 2008. Metrics for evaluating performance of prognostic techniques. In: 2008 International Conference on Prognostics and Health Management. IEEE, pp. 1–17.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Netw. 61, 85–117.

Schuld, M., 2021. Supervised quantum machine learning models are kernel methods. http://dx.doi.org/10.48550/arXiv.2101.11020, arXiv:2101.11020.

Schuld, M., 2022. Quantum machine learning models are kernel methods. Nat. Mach. Intell. 4 (6), 421–432.

Schuld, M., Killoran, N., 2019. Quantum machine learning in feature Hilbert spaces. Phys. Rev. Lett. 122 (4), 040504.

Schuld, M., Sinayskiy, I., Petruccione, F., 2015. An introduction to quantum machine learning. Contemp. Phys. 56 (2), 172–185.

Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O., 2020. Green AI. Commun. ACM 63 (12), 54–63.

Shaydulin, R., Safro, I., Kozlowski, W., et al., 2023. Evaluating latency in hybrid quantum–classical algorithms on real hardware. IEEE Trans. Quantum Eng. 4, 1–10. http://dx.doi.org/10.1109/TQE.2023.3265551.

Shi, S., Wang, Z., Li, J., Li, Y., Shang, R., Zheng, H., Zhong, G., Gu, Y., 2023. A natural NISQ model of quantum self-attention mechanism. arXiv preprint arXiv:2305.15680.

Shor, P.W., 1994. Algorithms for quantum computation: discrete logarithms and factoring. In: Proceedings 35th Annual Symposium on Foundations of Computer Science. IEEE, pp. 124–134.

Sikorska, J.Z., Hodkiewicz, M., Ma, L., 2011. Prognostic modelling options for remaining useful life estimation by industry. Mech. Syst. Signal Process. 25 (5), 1803–1836.

Silva, G.S.M., Droguett, E.L., 2022. Quantum machine learning for health state diagnosis and prognostics. In: 2022 Annual Reliability and Maintainability Symposium. RAMS, IEEE, pp. 1–7.

Soon, K.L., Soon, L.T., 2025. A hybrid quantum neural network and classical gated recurrent unit for battery state of health forecasting incorporating SHAP analysis. J. Energy Storage 136, 118596.

Soon, K.L., Tan, W., Lim, C.H., Lee, H., 2025. A quantum-enhanced ensemble model to forecast battery state of health under varying discharging load. Measurement 256, 118318. http://dx.doi.org/10.1016/j.measurement.2025.118318.

Spall, J.C., 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans. Autom. Control 37 (3), 332–341. http://dx.doi.org/10.1109/9.119632.

Stein, J., Schuman, D., Benkard, M., Holger, T., Sajko, W., Kölle, M., Nüßlein, J., Sünkel, L., Salomon, O., Linnhoff-Popien, C., 2023. Exploring unsupervised anomaly detection with quantum boltzmann machines in fraud detection. arXiv preprint arXiv:2306.04998.

Stokes, J., Izaac, J., Killoran, N., Carleo, G., 2020. Quantum natural gradient. Quantum 4, 269. http://dx.doi.org/10.22331/q-2020-05-25-269.

Tacchino, F., et al., 2019. Quantum feedforward neural networks. Phys. Rev. Lett. 122, 060501.

Talaei Khoei, T., Ould Slimane, H., Kaabouch, N., 2023. Deep learning: systematic review, models, challenges, and research directions. Neural Comput. Appl. 35 (31), 23103–23124.

Thuillier, J., Jha, M.S., Le Martelot, S., Theilliol, D., 2024. Prognostics aware control design for extended remaining useful life: Application to liquid propellant reusable rocket engine. Int. J. Progn. Health Manag. 15 (1).

Tian, J., Yang, W., 2024. Explainable quantum neural networks: Example-based and feature-based methods. Electronics 13 (20), 4136. http://dx.doi.org/10.3390/electronics13204136.

Tsurkan, O., Konstantinova, A., Sedykh, A., Zhiganov, D., Senokosov, A., Tarpanov, D., Anoshin, M., Fedichkin, L., 2025. Hybrid quantum recurrent neural network for remaining useful life prediction. arXiv preprint arXiv:2504.20823.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wang, F.-K., Kebede, G.A., Lo, S.-C., Woldegiorgis, B.H., 2024. An embedding layer-based quantum long short-term memory model with transfer learning for proton exchange membrane fuel stack remaining useful life prediction. Energy 308, 133054.

Wang, B., Lei, Y., Li, N., Li, N., 2018. A hybrid prognostics approach for estimating remaining useful life of rolling element bearings. IEEE Trans. Reliab. 69 (1), 401–412.

Wang, R., Wang, Y., Liu, J., Koike-Akino, T., 2025. Quantum diffusion models for few-shot learning.

Wang, X., et al., 2023. Adverse drug reaction detection from social media based on quantum Bi-LSTM with attention. IEEE Access 11, 16194–16202.

Wendlinger, M., Tscharke, K., Debus, P., 2025. Old rules in a new game: Mapping uncertainty quantification to quantum machine learning. In: 2025 IEEE International Conference on Quantum Computing and Engineering. QCE, http://dx.doi.org/10.1109/QCE65121.2025.00198.

Wiedmann, M., Hölle, M., Periyasamy, M., Meyer, N., Ufrecht, C., Scherer, D.D., Plinge, A., Mutschler, C., 2023. An empirical comparison of optimizers for quantum machine learning with SPSA-based gradients. In: 2023 IEEE International Conference on Quantum Computing and Engineering. QCE, pp. 450–456. http://dx.doi.org/10.1109/QCE57702.2023.00058.

Wu, S., Jin, S., Wen, D., Han, D., Wang, X., 2025. Quantum reinforcement learning in continuous action space. Quantum 9, 1660. http://dx.doi.org/10.22331/q-2025-03-12-1660.

Wu, Y., Sicard, B., Gadsden, S.A., 2024. Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring. Expert Syst. Appl. 255, 124678. http://dx.doi.org/10.1016/j.eswa.2024.124678.

Wu, Y., Yuan, M., Dong, S., Lin, L., Liu, Y., 2018. Remaining useful life estimation of engineered systems using vanilla lstm neural networks. Neurocomputing 275, 167–179.

Xiang, W., Li, F., Wang, J., Tang, B., 2018. Quantum weighted gated recurrent unit neural network and its application in performance degradation trend prediction of rotating machinery. Neurocomputing 313, 85–95.

Xiang, S., Qin, Y., Zhu, C., Wang, Y., Chen, H., 2020. LSTM networks based on attention ordered neurons for gear remaining life prediction. ISA Trans. 106, 343–354.

Xu, Z., Yu, W., Zhang, C., Chen, Y., 2024. Quantum convolutional long short-term memory based on variational quantum algorithms in the era of NISQ. Information 15 (4), 175.

Yin, Z., Agresti, I., de Felice, G., Brown, D., Toumi, A., Pentangelo, C., Piacentini, S., Crespi, A., Ceccarelli, F., Osellame, R., Coecke, B., Walther, P., 2025. Experimental quantum-enhanced kernel-based machine learning on a photonic processor. Nat. Photonics 19 (9), 1020–1027. http://dx.doi.org/10.1038/s41566-025-01682-5.

Yu, Z., Yao, H., Li, M., Wang, X., 2022. Power and limitations of single-qubit native quantum neural networks. Adv. Neural Inf. Process. Syst. 35, 27810–27823.

Yuan, M., Wu, Y., Lin, L., 2016. Fault diagnosis and remaining useful life estimation of aero engine using 1stm neural network. In: IEEE International Conference on Aircraft Utility Systems. AUS, IEEE, pp. 135–140.

Zaid, A.A.-M., Al-Jonaid, A.M.A., Al-Qubati, A.A., Wang, C., 2024. Remaining useful life estimation of aircraft engines using siamese attention-augmented quantum convolutional neural networks. In: Proc. 5th International Conference on Computer Engineering and Applications. ICCEA 2024, IEEE, pp. 1366–1371.

Zhang, J., Che, X., Fan, Y., Peng, S., Chen, G., Ma, Q., Hu, J., 2025. Denoising diffusion models with optimized quantum implicit neural networks for image generation. Future Gener. Comput. Syst. http://dx.doi.org/10.1016/j.future.2025.107875.

Zhang, K., Hsieh, M.-H., Liu, L., Tao, D., 2020. Toward trainability of quantum neural networks. arXiv preprint arXiv:2011.06258.

Zhang, Z., Wu, Y., Ma, X., 2025. Quantum machine learning based wind turbine condition monitoring: State of the art and future prospects. Energy Convers. Manage. 332, 119694.

Zhang, H., Zhang, Q., Shao, S., Niu, T., Yang, X., 2020. Attention-based LSTM network for rotatory machine remaining useful life prediction. IEEE Access 8, 132188–132199.

Zhao, H., Chen, Z., Shu, X., Shen, J., Lei, Z., Zhang, Y., 2023. State of health estimation for lithium-ion batteries based on hybrid attention and deep learning. Reliab. Eng. Syst. Saf. 232, 109066. http://dx.doi.org/10.1016/j.ress.2022.109066.

Zheng, S., Ristovski, K., Farahat, A., Gupta, C., 2017. Long short-term memory network for remaining useful life estimation. In: 2017 IEEE International Conference on Prognostics and Health Management. ICPHM, IEEE, pp. 88–95.

Zoufal, C., Lucchi, A., Woerner, S., 2021. Variational quantum Boltzmann machines. Quantum Mach. Intell. 3, 7. http://dx.doi.org/10.1007/s42484-020-00033-7.