Journal of Transactions in Systems Engineering
https://journals.tultech.eu/index.php/jtse.
ISSN: 2806-2973
Volume 4, Issue 3
DOI: https://doi.org/10.15157/JTSE.2026.4.3.745-771
Received: 07.09.2026; Revised: 07.10.2026; Accepted: 15.11.2026

Research Article

Q-Trace: A Noise-Aware Quantum Feature Attribution and Explanation Stability Framework for Hybrid Learning Models

Surapaneni Phani Praveen[1], Varun Kaza[1,*], Massila Kamalrudin[2], Mustafa Musa[3], Sai Srinivas Vellela[4], Vahiduddin Shariff[5]

[1]Department of Computer Science and Engineering, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India.
[2]Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia.
[3]Center of Research and Innovation Management, Universiti Teknikal Malaysia Melaka (UTeM), Melaka, Malaysia.
[4]School of Computing Sciences, Department of CSE-Data Science, Chalapathi University, Mothadaka, Guntur, Andhra Pradesh, India.
[5]Department of Computer Science & Engineering, Sir C. R. Reddy College of Engineering, Eluru, Andhra Pradesh, India.
***varunkaza20@gmail.com**

Abstract

Hybrid Hybrid quantum-classical learning models increasingly rely on parameterized quantum circuits to encode classical data into high-dimensional Hilbert spaces. However, the explainability of these models under realistic hardware noise remains poorly understood. Existing feature-attribution techniques adapted from classical machine learning generally assume noiseless, deterministic function evaluation. This assumption is violated on near-term quantum devices, where gate errors, decoherence, and finite-shot sampling affect the expectation values used to construct explanations. This paper introduces Q-Trace, a framework that formulates feature attribution as an operator acting on a noisy quantum channel rather than on an idealized circuit, while explicitly quantifying the extent to which explanations may drift as noise increases. Q-Trace comprises three components: Noise-Aware Quantum Gradient Attribution (NQGA), which extends the parameter-shift rule to noisy quantum channels; Noise-Conditioned Shapley Estimation (NCSE), which computes exact Shapley-value attributions directly from the noisy expectation-value function; and the Explanation Stability Index (ESI), a similarity-based metric for measuring attribution drift across noise levels. To the best of our knowledge, Q-Trace is the first framework to explicitly incorporate a parameterized noise channel into input-feature attribution, evaluate gradient-based and coalition-based attribution operators under the same noise channel, and quantify explanation stability through a dedicated metric. We derive a closed-form bound relating ESI degradation to the contraction rate of the underlying depolarizing channel and validate the framework using a four-qubit hybrid variational classifier simulated exactly at the density-matrix level. The results show that Shapley-based attributions are consistently more stable than gradient-based attributions. At a noise level of (p =

©2026 Authors. This is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 International License CC BY 4.0 (http://creativecommons.org/licenses/by/4.0).

0.20), the cosine and rank ESI values for the Shapley approach reach 0.991 and 0.927, respectively, averaged across five trained models, compared with 0.963 and 0.854 for the gradient approach. Furthermore, finite-difference gradient estimators exhibit substantially greater sensitivity to shot noise, with the standard deviation of gradient estimates increasing by a factor of 52.15 on average across six shot budgets.

Keywords: Quantum Machine Learning; Explainable AI; Variational Quantum Circuits; Feature Attribution; NISQ Noise; Shapley Values; Parameter-shift Rule; Explanation Stability.

INTRODUCTION

Quantum machine learning (QML) has moved, over the past several years, from a purely theoretical curiosity into an experimentally accessible discipline, driven by the availability of noisy intermediate-scale quantum (NISQ) processors and by a growing body of hybrid quantum-classical algorithms in which a parameterized quantum circuit (PQC) is trained jointly with, or in place of, a classical neural network layer [1–3]. In these hybrid models, an encoding unitary maps classical features into a quantum state, a trainable variational circuit processes it, and the expectation value of one or more observables yields the readout. This construction has been shown to provide access to feature spaces that are difficult to represent classically [2, 3], and has motivated substantial theoretical work on the expressivity, trainability, and generalization behaviour of such circuits [4–7].

As hybrid quantum-classical models are proposed for increasingly consequential tasks including chemistry, finance, and healthcare-adjacent classification problems the question of why a model produced a given prediction becomes as important as the prediction itself. In classical deep learning, a mature field of explainable artificial intelligence (XAI) addresses this question by providing feature-attribution techniques such as gradient saliency, Integrated Gradients, LIME, SHAP, and Grad-CAM [8–11]. These methods, however, were designed under an implicit assumption that is rarely stated explicitly: that the function being explained can be evaluated exactly and repeatedly, with no stochastic or systematic corruption between evaluations. This assumption is reasonable for a classical feed-forward network executed on deterministic hardware. Still, it fails for a variational quantum circuit executed on a NISQ device, where every gate contributes gate error and decoherence, and every expectation value is estimated from a finite number of projective measurement shots.

This creates a specific and, to date, under-examined problem. A feature attribution computed on a hybrid quantum-classical model is itself a function of the noise channel through which the circuit is executed. Yet, almost all attribution methods applied to QML models to date have been transplanted directly from the noiseless classical setting, with noise treated if at all as an afterthought affecting model accuracy rather than model explanation. Classical XAI research has separately shown that even in the noiseless case, feature attributions can be fragile: small, imperceptible input perturbations can produce large changes in the resulting explanation while leaving the model's prediction unchanged [12]. If gradient-based and perturbation-based explanations are already fragile under

noiseless conditions, it is natural to ask how much additional instability is introduced when the function being explained is itself computed by a physically noisy process whose statistics change continuously with calibration drift, circuit depth, and qubit connectivity. Existing quantum-XAI work has applied Shapley values to quantum-circuit and QML explanations and has considered uncertainty in the resulting quantum value function [13]; to our knowledge, however, no prior framework treats an explicit, parameterized gate-level noise channel as a first-class argument of a quantum feature attribution, jointly formalizes a matched gradient-based and coalition-based attribution pair under that channel, and provides a dedicated metric for how much an explanation should be trusted at a given noise level. In practice, the most similar previous work [13] uses Shapley values on quantum circuits with gates or collections of gates as explained entities and discusses uncertainty in the value function, while Q-Trace has matched the two families of attribution functions by including a gradient-based attribution operator and adding a stability index to accommodate the noise level as an explicit argument.

This paper addresses that gap by introducing Q-Trace, a framework for computing and evaluating feature attributions for hybrid quantum-classical models under an explicit, parameterized noise model. Rather than proposing yet another attribution heuristic, Q-Trace reformulates two existing and well-understood building blocks the analytic parameter-shift rule [4, 14, 15] and the exact Shapley value [8] as functions of a noisy quantum channel, and complements them with a dedicated stability metric, the Explanation Stability Index (ESI), that quantifies how much an attribution vector drifts as the channel's noise parameter increases. The central contributions of this paper are as follows:

- *A noise-explicit attribution formalism.* We formalize feature attribution for hybrid quantum-classical models as a map from a classical input, a set of trained circuit parameters, and a noise-channel parameter to an attribution vector, making explicit a dependency that is left implicit in prior QML explainability work.
- *Two noise-aware attribution operators.* We define Noise-Aware Quantum Gradient Attribution (NQGA), which applies the analytic parameter-shift rule directly to the noisy expectation-value function, and Noise-Conditioned Shapley Estimation (NCSE), which computes exact Shapley values over feature subsets of the same noisy function, giving two attribution families with different sensitivity profiles to gate noise.
- *The Explanation Stability Index.* We define ESI, a cosine-similarity- and rank-correlation-based measure of attribution drift between a reference (typically noiseless) circuit and its noisy counterpart, and we derive an analytic upper bound on attribution deviation in terms of the depolarizing channel's Bloch-vector contraction rate.
- *Experimental data and a shot-noise analysis.* In a four-quantum density-matrix simulation, coalition-based attribution outperformed gradient-based attribution in all cases of nonzero noise tested, and a finite-shot simulation demonstrates a finite-

difference estimator that increases the standard deviation of the attribution by approximately 52× over the parameter-shift estimator.

The remainder of this paper is organized as follows. The related work in quantum machine learning, NISQ noise modeling, and classical explainable AI is reviewed, positioning Q-Trace within these research areas. The notation and theoretical background required to define a noisy hybrid quantum-classical model are then introduced, followed by a formalization of the noise-aware attribution problem. Next, the Q-Trace framework is presented in detail, including NQGA, NCSE, and ESI. A theoretical stability bound is subsequently derived. The experimental setup is then described, and the resulting evidence is reported and discussed. Finally, the limitations and threats to validity are examined, followed by the conclusions and directions for future work.

RELATED WORK

Quantum feature encoding and hybrid variational models

The idea of using a quantum state space as a feature space for classical data was popularized by [2], who introduced a quantum kernel estimator and a quantum variational classifier built around an angle-based feature map, and by [3], who framed quantum classifiers explicitly as kernel methods operating in a quantum-enhanced Hilbert space. Authors in [4, 14] independently established that expectation values of parameterized quantum circuits with generator eigenvalues ±1/2 admit an exact analytic gradient via the parameter-shift rule, generalized to a broader class of gates by [15]. The study in [16] provide a comprehensive review of parameterized quantum circuits as machine-learning models, cataloging encoding strategies, ansatz families, and training procedures used by the majority of hybrid quantum-classical architectures today. Authors in [17] introduced an effective-dimension metric grounded in the Fisher information matrix and showed that well-designed quantum neural networks can achieve a significantly higher effective dimension and correspondingly faster, better-generalizing training than classical networks with the same number of parameters; because this analysis is tied to the spectrum of the Fisher information matrix, it also connects directly to the barren-plateau phenomenon [5, 18] discussed in the next subsection. Data-encoding choices are known to affect that expressivity in a more specific sense the study in [7] show that the Fourier spectrum accessible to a variational model is governed by the eigenvalues of the encoding gates. Authors in [19] propose data re-uploading as a way to increase expressivity without additional qubits, and [20] study which encodings are most robust to input and hardware imperfections. None of this line of work, however, asks how the resulting model's explanations behave once the encoding and ansatz are executed on a physically noisy channel.

Trainability and Noise in Near-Term Quantum Circuits

A parallel body of work has focused on the trainability of variational circuits rather than their explainability. Authors in [18] showed that randomly initialized, sufficiently

deep circuits exhibit barren plateaus, and authors in [5] extended this analysis to cost functions built from local observables, showing that the choice of observable materially affects trainability. The studies in [21, 22] survey variational quantum algorithms and NISQ algorithms, cataloging the sources of hardware noise including depolarizing error, amplitude damping, and readout error that corrupt present-day circuit execution, a theme also central to Preskill's foundational discussion of the NISQ era [23]. Error-mitigation techniques such as zero-noise extrapolation and probabilistic error cancellation [24], and broader reviews of hybrid quantum-classical error mitigation [25], attempt to recover noiseless expectation values in post-processing. The study in [26] demonstrate a hardware-efficient variational eigensolver in the presence of such noise. These works treat noise as a first-order concern for prediction accuracy, not for the attribution assigned to each input feature. Q-Trace targets this.

Classical Explainable AI and Fragility of Attributions

Feature-attribution methods for classical neural networks fall broadly into gradient-based approaches, such as Integrated Gradients [9] and Grad-CAM [11], and perturbation- or game-theoretic approaches, such as LIME [10] and SHAP [8], the latter of which formalizes attribution as an exact or approximate Shapley value over feature coalitions. A separate strand of classical XAI research has studied the reliability of these attributions independently of any notion of physical hardware noise. Authors in [12] show that imperceptible, label-preserving adversarial perturbations can produce dramatically different saliency maps, establishing that gradient-based explanations can be fragile even for an exactly computable, noiseless function. Authors in [27] similarly investigate the robustness of interpretability methods to small input perturbations, while the study in [28] advocates a more rigorous, quantitatively grounded science of interpretability rather than relying solely on qualitative case studies. The study in [29] survey the broader XAI landscape and its evaluation criteria. Closer to the present setting, authors in [13] apply Shapley-value attribution to quantum circuits and explicitly discuss uncertainty in the resulting quantum value function, representing the most directly related prior work on Shapley-based quantum explainability; Q-Trace differs by treating an explicit, parameterized gate-level depolarizing channel as the object of study and by introducing a matched gradient-based attribution (NQGA) and a dedicated stability metric (ESI) for direct comparison against the coalition-based case. These results motivate the present work's central design choice: rather than assuming that quantum attributions inherit the reliability properties of their classical counterparts, Q-Trace measures attribution reliability directly, as a function of an explicit, physically motivated noise parameter, using an evaluation methodology in the same spirit as [12] and [27] but targeted at the quantum-channel setting.

Positioning of this Work

To summarize the gap: quantum-encoding research [2, 3, 7, 16, 19, 20] optimizes for model expressivity under an implicit noiseless-circuit assumption; NISQ noise research [5, 18, 21–26] optimizes for prediction accuracy in the presence of realistic hardware error;

classical XAI research [8–12, 27–29] optimizes for attribution fidelity under an implicit assumption of exact, repeatable function evaluation; and quantum-Shapley work such as [13] establishes Shapley-based attribution for quantum circuits and touches on value-function uncertainty without treating parameterized gate-level noise channel as the object of study. Q-Trace sits at the intersection of these lines, asking specifically how the attribution assigned to a classical input feature changes as the quantum channel that encodes and processes it becomes noisier, and providing both an attribution formalism and a stability metric to answer that question quantitatively. Table 1 summarizes how Q-Trace differs from the principal existing approaches.

Table 1. Comparison of Q-Trace with the Principal Related Attribution Approaches

Approach	**Explained entity**	**Attribution mechanism**	**Treatment of noise / stability**	**Principal limitation**
Classical SHAP [8]	Input features of a classical model	Shapley values over feature coalitions	Exact, repeatable function evaluation assumed; no noise parameter	Not defined for a noisy quantum channel; shot and gate noise not addressed
Integrated Gradients [9]	Input features of a classical model	Path-integrated gradients	Deterministic evaluation assumed; fragility under input perturbation reported in [12, 27]	No hardware-noise model; stability not tied to a physical noise parameter
Parameter-shift rule [4, 14, 15]	Circuit parameters (training gradients)	Exact analytic derivative from two shifted evaluations	Formulated for the ideal circuit; not used as a noise-parameterized feature attribution (to our knowledge)	Provides gradients for optimization, not an explanation-stability analysis
Shapley values for quantum circuits [13]	Gates or groups of gates of a circuit	Shapley values	Value-function uncertainty discussed; no parameterized noise channel as the object of study and no noise-sweep stability index (to our knowledge)	Explains circuit structure rather than input-feature attribution under controlled noise levels
Q-Trace (this work)	Input features of a hybrid quantum-classical model	NQGA (noisy parameter-shift) and NCSE (exact noisy Shapley)	Depolarizing strength p is an explicit argument; ESI (cosine and rank) with a contraction-based bound	Exact Shapley limited to small n; depolarizing channel only; simulation only

PRELIMINARIES

This section fixes notation for the classical-to-quantum encoding, the noisy variational circuit, and the two analytic tools: the parameter-shift rule and the Shapley value that Q-Trace repurposes for the noisy setting.

Quantum feature encoding and variational readout

Let $x \in R^n$ be a vector of classical features and let $U_{enc}(x)$ denote an encoding unitary acting on n qubits initialized in the computational-basis state $|0\rangle \otimes^n$. The resulting encoded state is illustrated in equation (1):

$$|\psi(x)\rangle = U_{enc}(x)|0\rangle \otimes^n \quad (1)$$

In this paper, U_{enc} applies a single-qubit rotation $R_Y(x_i)$ to qubit i for each feature x_i, a widely used angle-encoding strategy [2], [16]. A variational ansatz U_θ, parameterized by trainable angles θ, is then applied, followed by measurement of an observable O (a Pauli-Z operator on a designated qubit throughout this paper). On ideal, noiseless hardware, the circuit evolves a pure state, and the corresponding density matrix $\rho = |\psi\rangle\langle\psi|$ evolves under conjugation by each unitary gate U in the circuit as:

$$\rho' = U \rho U^\dagger \quad (2)$$

Noisy circuit execution

On NISQ hardware, every gate is followed by residual decoherence and control error. We model this, as is standard in the NISQ literature [23], [22], [25], with a single-qubit depolarizing channel applied to the relevant qubit(s) after every gate in the circuit. For a single-qubit density matrix ϱ acted on by a depolarizing channel of strength p ∈ [0, 1],

$$N_p(\rho) = (1 - p)\rho + \frac{p}{3}(X\rho X + Y\rho Y + Z\rho Z) \quad (3)$$

where X, Y, Z are the Pauli operators. Composing the full noisy circuit therefore produces a mixed state $\rho_p(x; \theta)$ rather than a pure state, and the observable's expectation value is read out as a trace,

$$\langle O \rangle_\rho = Tr(O\rho) \quad (4)$$

where the subscript p on ρ_p will be used throughout to make the dependence on the noise-channel strength explicit; $p = 0$ recovers the ideal, noiseless circuit. This construction a unitary circuit interleaved with a fixed single-qubit noise channel is a standard simplified model of gate error on NISQ hardware [23], [22] and is exact (not approximated by finite sampling) whenever the density matrix is tracked directly, as is done in the simulation underlying this paper.

The parameter-shift rule

For a gate generated by an operator with eigenvalues ±1/2 (as is the case for the single-qubit rotations used throughout this work), the exact analytic derivative of an expectation

value with respect to that gate's parameter θ_i can be obtained without finite-difference approximation, using two circuit evaluations shifted by ± π/2 [4, 14]:

$$\frac{\partial\langle O\rangle}{\partial\theta_i} = \frac{1}{2}\left[\langle O\rangle\left(\theta_i + \frac{\pi}{2}\right) - \langle O\rangle\left(\theta_i - \frac{\pi}{2}\right)\right] \quad (5)$$

This identity holds exactly regardless of the surrounding circuit, and [15] generalize it to a wider class of gates. It re-applies this identity directly to the noisy channel of equations (3)–(4), rather than to the idealized, noiseless circuit for which it was originally derived, which is the first building block of the Q-Trace framework.

Shapley values

Given a set of features $F = \{1, \ldots, n\}$ and a value function v: $2^F \rightarrow R$ assigning a scalar value to every subset of features, the Shapley value of feature i distributes the total value $v(F) - v(\emptyset)$ among features according to their average marginal contribution across all orderings of F [8]. The value function is defined directly from the noisy circuit's expectation value, making the Shapley value itself a function of the noise-channel strength—the second building block of Q-Trace. Table 2 summarizes the notation used throughout the paper.

Table 2. Summary of the notation used throughout the paper

Symbol	Meaning
n	Number of qubits / classical input features
$x \in R^n$	Classical feature vector
θ	Trainable variational circuit parameters
$U_{enc}(x)$	Feature-encoding unitary (angle encoding, R_Y gates)
U_θ	Variational ansatz unitary
$p \in \{0,1\}$	Single-qubit depolarizing noise-channel strength
N_p	Single-qubit depolarizing channel of strength p
$\rho_p(x;\ \theta)$	Noisy circuit density matrix at noise level p
O	Measured observable (Pauli-Z on the readout qubit)
$\langle O\rangle_p(x;\theta)$	Noisy expectation value/model output
$A(x,\theta,p) \in R^n$	Feature-attribution vector at noise level p
$\varphi_i(v)$	Shapley value of feature i under value function v
$ESI(p)$	Explanation Stability Index at noise level p
$\lambda(p)$	Bloch-vector contraction factor of N_p, $\lambda(p) = 1 - \frac{4p}{3}$

PROBLEM FORMULATION

A hybrid quantum-classical model, executed on real or simulated NISQ hardware, can be written as a single map from a classical input and a set of trained parameters to a scalar (or vector) output, parameterized additionally by the strength of the noise channel through which it is executed:

$$f_p(x;\,\theta) = Tr\left(O\,\rho_p(x;\,\theta)\right) \tag{6}$$

At $p = 0$, f_0 recovers the noiseless model that almost all prior quantum-explainability discussions implicitly assume; at $p > 0$, f_p is the function that is actually realized on hardware, or on a faithful noise simulation thereof. A feature-attribution method is, in general, an operator A that maps an input x, a parameter set θ, and — in the setting of this paper — a noise strength p, to a vector $A(x, \theta, p) \in R^n$ whose i-th component is interpreted as the contribution of feature x_i to the model output $f_p(x;\,\theta)$

Two properties are desirable of such an operator once p is allowed to vary. First, consistency in the noiseless limit: $A(x, \theta, p)$ should converge to a well-understood, standard attribution (such as an analytic gradient or an exact Shapley value) as $p \to 0$, so that the noisy attribution is recognizable as a perturbation of a known baseline rather than an unrelated quantity. Second, graceful degradation: small increases in p should produce correspondingly small, quantifiable changes in A, rather than a discontinuous or unbounded drift, so that a practitioner can decide, from a known noise level, how much to trust the resulting explanation. Neither property is guaranteed a priori for an attribution method that is simply transplanted from the classical, noiseless setting: a method could in principle satisfy the first property while badly violating the second, producing an explanation that is correct on average over many noisy shots but wildly unstable from one execution to the next.

This paper formalizes the second property through the Explanation Stability Index, which measures the similarity between $A(x, \theta, 0)$ and $A(x, \theta, p)$ as p increases. The problem addressed by Q-Trace can then be stated concisely: (i) construct at least one attribution operator A that is well-defined directly on the noisy channel f_p rather than only on its noiseless limit f_0; (ii) characterize, empirically and theoretically, how $A(x, \theta, p)$ drifts from $A(x, \theta, 0)$ as p increases; and (iii) determine whether different attribution families in particular, gradient-based versus coalition-based (Shapley) attributions differ systematically in their noise robustness, so that practitioners have a basis for choosing between them when explanation stability, rather than raw computational cost, is the binding constraint.

THE Q-TRACE FRAMEWORK

Q-Trace consists of three components, illustrated schematically in the Figure 1: Noise-Aware Quantum Gradient Attribution (NQGA), Noise-Conditioned Shapley Estimation (NCSE), and the Explanation Stability Index (ESI), which consumes the outputs of both

attribution modules across a sweep of noise levels and produces a single stability report for a given input.

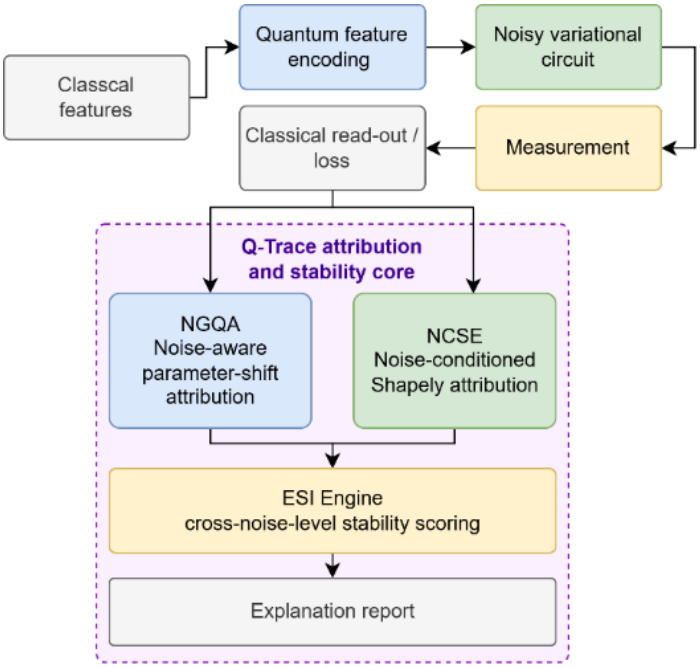


Figure 1. The Q-Trace Pipeline: A Noisy Circuit Feeds NQGA and NCSE Attribution, Compared Across Noise Levels by the ESI engine.

Noise-Aware Quantum Gradient Attribution (NQGA)

NQGA is obtained by applying the parameter-shift identity of equation (5) directly to the input encoding angles of the noisy channel f_p, rather than to the noiseless function f_0 as in prior work. Writing e_i for the i-th standard basis vector, the NQGA attribution of feature i is defined as

$$A_i^{NQGA}(x,\theta,p) = \frac{1}{2}\left[\langle O\rangle_p\left(x+\left(\frac{\pi}{2}\right)e_i,\theta\right) - \langle O\rangle_p\left(x-\left(\frac{\pi}{2}\right)e_i,\theta\right)\right] \quad (7)$$

Because the identity is applied to the same noisy channel that produces the model's prediction, NQGA is, by construction, an exact analytic derivative of the noisy expectation-value surface, not an approximation of the noiseless gradient recovered after the fact. For comparison, this paper also evaluates the naïve finite-difference gradient estimator with step size ε,

$$A_i^{FD}(x,\theta,p) = \frac{[\langle O\rangle_p(x+\varepsilon e_i,\theta) - \langle O\rangle_p(x-\varepsilon e_i,\theta)]}{2\varepsilon} \quad (8)$$

Equations (7) and (8) agree in the limit of exact (infinite-shot) expectation-value evaluation up to a second-order Taylor remainder in the underlying smooth function. Still, they behave very differently under finite-shot measurement noise. Because the equaiton (8) divides by a small ε, any statistical noise in the two measured expectation values is amplified by a factor of roughly 1/ε relative to the fixed π/2 shift used by NQGA.

Noise-Conditioned Shapley Estimation (NCSE)

The second attribution family departs from gradients entirely and instead constructs a coalition value function directly on the noisy channel. For a subset of features $S \subseteq F$, define a masked input $x \odot 1_S$ in which features outside S are replaced by a fixed baseline angle (zero, corresponding to the unencoded state $|0\rangle$ on that qubit), and define the noisy value function as shown in equation (9):

$$v(S;\ x, \theta, p) = \langle O \rangle_p (x \odot 1_S;\ \theta) \tag{9}$$

The Shapley value of feature i under this value function is the weighted average of its marginal contribution across every ordering of the remaining features,

$$\varphi_i(v) = \sum_{S \subseteq F\{i\}} w(|S|) \cdot [v(S \cup \{i\}) - v(S)] \tag{10}$$

where $w(|S|) = \frac{|S|!(|F|-|S|-1)!}{|F|!}$ is the standard combinatorial weight. Because the study in this paper uses only four input features, we compute the sum over $2^4 = 16$ subsets in equation (10) exactly for every test point and every noise level, with no sampling approximation. This property would not scale to high-dimensional classical inputs, but that makes the present study a clean, bias-free comparison between attribution families. The noise-conditioned attribution vector is then defined as:

$$A_i^{NCSE}(x, \theta, p) = \varphi_i(v_p) \tag{11}$$

where v_p denotes the value function of equation (9) evaluated on the noisy channel at strength p. Because equation (10) averages over a large number of value-function differences rather than relying on a single pair of shifted evaluations, NCSE has the potential to average out some of the noise-induced fluctuation that affects any individual term an intuition.

Shapley attributions are defined with respect to a reference input, and the baseline is important. The following are two reasons why zero was used. First, there is no circuit that encodes a qubit into the state $|0\rangle$, so a circuit with a feature masked is a physically implementable circuit that can be implemented with no information about that feature actually loaded into the qubit, and not an 'out of range' value for a qubit. Second, all features are min–max rescaled onto the range $\{0, \pi\}$ with the training set, and thus the base angle 0 represents the minimum value of the feature in the training set, meaning that NCSE quantifies each feature's contribution in relation to the all-minimum reference input. This reference is not neutral in the sense of a data-mean baseline the attribution magnitudes and orderings given are to be interpreted in relation to this reference.

This research work did not conduct a sensitivity analysis on alternative baselines (in this case the midpoint angle $\pi/2$ or the training mean) and this is something to be explored in future. Additionally, Figure 2 depict the schematic view of the four-qubit circuit.

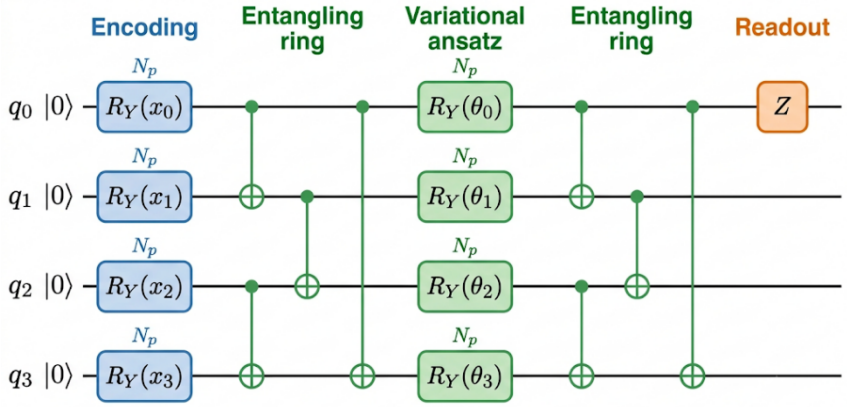


Figure 2. The Four-Qubit Circuit: Angle Encoding, two CNOT Entangling Rings, a Variational R_Y Layer, and Pauli-Z Readout on Qubit 0. N_p Marks the Depolarizing Channel of Equation (3)

The Explanation Stability Index (ESI)

Given a reference attribution vector computed at a baseline noise level (p = 0 throughout this paper) and a noisy attribution vector computed at level p, ESI measures their agreement in direction using cosine similarity, averaged over a test set D:

$$ESI_{cos}(p) = \frac{1}{|D|}\sum_{x\in D}\cos\big(A(x,\theta,0), A(x,\theta,p)\big) \tag{12}$$

where $\cos(u,v) = \frac{u\cdot v}{\|u\|\|v\|}$ is the standard cosine similarity. Because feature importance is often consumed as a ranking ("which feature mattered most"), Q-Trace additionally reports a rank-based variant using Spearman's rank correlation ϱ_S between the absolute-value rankings of the two attribution vectors,

$$ESI_{rank}(p) = \frac{1}{|D|}\sum_{x\in D}\left(1 - \frac{6\sum_j d^2}{n(n^2-1)}\right) \tag{13}$$

where d_j is the difference between the two rankings for feature j. Both variants take the value 1 when the noisy attribution exactly reproduces the reference attribution's direction or ranking, respectively, and decrease as the two vectors diverge. Unlike a raw error norm, both measures are invariant to the overall scale of the attribution vector. This property is important because the magnitudes of both NQGA and NCSE attributions shrink substantially as (p) increases, even when their directions remain comparatively stable.

Complementary and limitations of two measures. Drift comes in different faces, such as ESI_{cos} and ESI_{rank}. While cosine similarity tends to be more sensitive to the largest magnitude components of the attribution vector, Spearman correlation weights each of the other features equally, including those with near-zero attribution, and so tends to be more sensitive to reshuffling among the smaller components. The difference between the two columns in Table 4 (for instance, 0.963 vs. 0.854 for NQGA at p = 0.20) is consistent with this difference, and reporting both makes it easier to distinguish "direction preserved but secondary ordering changed" from "both preserved". They are both, however, insensitive to the magnitude of the attribution: cos(u, cv) = cos(u, v) for any c > 0, so ESI_{cos} can stay close to 1 even if the entire attribution vector is reduced significantly, as is the case when

p increases (Figure 5). Furthermore, a strong ESI rather means that the "remaining" signal points are in the same direction, not that the explanation is still robust. ESI is also undefined if both attribution vectors are of zero norm, becomes numerically ill-conditioned as the norm of either attribution vector approaches zero, for instance close to the fully depolarizing limit, where $\lambda(p) \to 0$; the rank-based variant is the same when all the attributions are near zero. To report the ESI along with the magnitude-retention ratio $\|A(x,\theta,p)\|_2 / \|A(x,\theta,0)\|_2$, and to set the ESI to undefined, instead of 0 or 1, if $\|A(x,\theta,p)\|_2$ is below a small tolerance.

Model Training

Q-Trace is agnostic to how the underlying hybrid model is trained, but for completeness, the training objective used in the case study is a mean-squared error between the sigmoid-like read-out $\hat{y} = \frac{1+\langle Z_0 \rangle}{2}$ and a binary label y,

$$L(\theta) = \frac{1}{N}\sum_{k=1}^{N}(\hat{y}_k - y_k)^2 \tag{14}$$

optimized by gradient descent using the parameter-shift rule of Eq. (5) applied to the variational angles θ (as opposed to the input angles x used by NQGA),

$$\theta \leftarrow \theta - \eta \nabla_\theta L(\theta) \tag{15}$$

with learning rate η. Training is performed once at (p=0), producing a single fixed model for deployment. All noise-dependent quantities reported in the experimental evaluation are computed using this fixed trained model, while varying only the execution-time noise strength (p). This setup reflects a realistic deployment scenario in which a model is trained under favourable conditions or in simulation and subsequently executed on a device subject to higher noise levels.

THEORETICAL ANALYSIS

This section derives an analytic upper bound on how far the noisy expectation value $f_p(x;\theta)$ can drift from its noiseless counterpart $f_0(x;\theta)$, and shows that this bound propagates directly into a stability guarantee for NQGA. The argument proceeds in three steps: a single-qubit contraction lemma, a trace-distance bound built from the contractivity of quantum channels, and a corollary bounding attribution drift.

Bloch-Vector Contraction under Depolarizing Noise

Lemma 1 (single-qubit depolarizing contraction). Write a single-qubit density matrix as $\rho = \frac{I + r\cdot\sigma}{2}$, where $r \in R^3$ is its Bloch vector and $\sigma = (X, Y, Z)$. Then the depolarizing channel of Eq. (3) acts on the Bloch vector as a uniform contraction,

$$r' = \lambda(p)r, \quad \lambda(p) = 1 - \frac{4p}{3} \tag{16}$$

Proof. Expanding XϱX, YϱY, and ZϱZ in the Pauli basis and summing shows that each Pauli conjugation preserves the identity component of ϱ while flipping the sign of two of the three Bloch components and preserving the third; summing the three conjugations and

substituting into equation (3) yields $N_p(\rho) = \frac{I + \lambda(p) r \cdot \sigma}{2}$ with λ(p) as given, which is the direct calculation underlying equation (16).

A trace-distance bound for the noisy channel

The depolarizing channel N_p is a convex combination of unitary (Pauli) conjugations, and every such channel is a contraction in trace norm: applying N_p to two density matrices ρ and σ can only decrease or preserve their trace distance. If a qubit passes through d independent applications of N_p over the course of a circuit — for instance, because it participates in d single- or two-qubit gates before the observable is measured — the contractions compound multiplicatively:

$$\|N_p^{\circ d}(\rho) - N_p^{\circ d}(\sigma)\|_1 \leq \lambda(p)^d \cdot \|\rho - \sigma\|_1 \tag{17}$$

Proposition 1 (expectation-value deviation bound). Let O be a Hermitian observable with operator norm $\|O\|_\infty$ ($\|O\|_\infty = 1$ for the Pauli-Z readout used throughout this paper), and let $\rho_p(x;\, \theta)$ and $\rho_0(x;\, \theta)$ denote the noisy and noiseless circuit states for the same input and parameters. Using the duality between trace norm and operator norm, $|Tr(O\Delta\rho)| \leq \|O\|_\infty \|\Delta\rho\|_1$ for any traceless $\Delta\rho$, together with Eq. (17) applied with σ taken as the noiseless evolution of the same input (for which $p = 0$ and $\lambda(0) = 1$), gives

$$|\langle O\rangle_p(x) - \langle O\rangle_0(x)| \leq (1 - \lambda(p)^d)\, \|O\|_\infty \tag{18}$$

where d is an effective count of the noise instances lying in the causal cone of the measured qubit (for the circuit of Fig. 2, d = 24. Tracing the observable Z_0 backward through the second CNOT ring (0→1, 1→2, 2→3, 3→0) shows that after one complete pass, all four qubits lie within the causal cone of qubit 0: the final 3→0 gate brings qubit 3 into the cone, the preceding 2→3 gate brings in qubit 2, and the preceding 1→2 gate brings in qubit 1. Once all four qubits are in the cone, the intervening variational R_Y layer and the first CNOT ring remain within it, and so do all four input-encoding rotations. Because the noise model applies one single-qubit depolarizing channel after every single-qubit gate and after every qubit participates in a CNOT, the causal-cone noise count is 4 (encoding layer) + 8 (first CNOT ring, 4 CNOTs × 2 participating qubits) + 4 (ansatz layer) + 8 (second CNOT ring) = 24, giving d = 24 rather than the smaller count obtained by considering only the single-qubit gate layers.

Equation (18) shows that the worst-case deviation of the noisy expectation value from its noiseless value grows monotonically with p and saturates at $\|O\|_\infty$ as p → 3/4, consistent with the fully depolarizing limit in which the channel destroys all information about the input state. With d = 24, this raw bound is loose and saturates quickly: $1 - \lambda(p)^{24}$ equals 0.275 at $p = 0.01$, 0.477 at $p = 0.02$, 0.809 at $p = 0.05$, 0.968 at $p = 0.10$, and 0.999 at $p = 0.20$. It should therefore be read as a worst-case envelope on raw expectation-value and attribution deviation, not as a numerical predictor of the much smaller ESI degradation.

Corollary: A Bound on NQGA Attribution Drift

Corollary 1. Because $A_i^{NQGA}(x, \theta, p)$ defined in equation (7) is one half the difference of two expectation values, each individually bounded by equation (18), the triangle inequality gives equation (19)

$$\left|A_i^{NQGA}(p) - A_i^{NQGA}(0)\right| \leq (1 - \lambda(p)^d) \tag{19}$$

for $\|O\|_\infty = 1$ equation (19) is consistent with, and provides a first-principles explanation for the monotonically decreasing ESI curve observed empirically for NQGA, the bound is loose (it does not itself predict the specific numerical decay rate observed in the Figure 4), but it correctly predicts the qualitative shape near-zero drift at small p, and accelerating drift as p grows and it confirms that no adversarial noise realization within this channel model can produce attribution drift faster than the rate at which the channel itself destroys quantum information.

On the Empirical Stability of NCSE

Since v(S) is the expectation of a bounded observable in a noisy circuit and the Shapley weights sum to one for each feature i, the triangle inequality gives the same worst-case deterministic bound, equation (19), for A_i^{NCSE}. However, this common worst-case bound alone cannot distinguish between NQGA and NCSE; it also doesn't explain the experimental result, where NCSE degrades more slowly than NQGA, i.e., retains a higher ESI at every nonzero noise level in Table 4 (with the gap becoming marked beyond p = 0.02). Instead, we introduce a second-moment argument, which is explicitly stochastic. Assume the noisy evaluation of the value function for coalition T is affected by a perturbation η_T with zero mean and variance $Var(\eta_T) \leq \sigma^2$, and assume perturbations across coalitions are uncorrelated. NCSE then becomes a weighted sum of the eight marginal perturbations for a fixed feature i, with n = 4.

$$\Delta_S = \eta_{S\cup\{i\}} - \eta_S \tag{20}$$

over S ⊆ F{i}, with Shapley weights 1/4, 1/12, 1/12, 1/12, 1/12, 1/12, 1/12, 1/4, so that

$$\sum_S {w_S}^2 = \frac{1}{6} \tag{21}$$

and consequently

$$Var[\Delta A_i^{NCSE}] \leq 2\sigma^2 \sum_S {w_S}^2 = \frac{\sigma^2}{3} \tag{22}$$

Under the same independence assumption, for NQGA the two shift evaluations have coefficients +1/2 and −1/2, respectively,

$$Var\left[\Delta A_i^{NQGA}\right] \leq \frac{\sigma^2}{2} \tag{23}$$

The variance ratio is therefore at most 2/3, and the standard-deviation ratio satisfies.

$$\frac{SD[\Delta A_i^{NCSE}]}{SD[\Delta A_i^{NQGA}]} \leq \sqrt{\frac{2}{3}} \approx 0.816 \quad (24)$$

More generally, if pairwise correlations between distinct value-evaluation perturbations are bounded by ϱ, a conservative form is:

$$Var[\Delta A_i^{NCSE}] \leq 2\sigma^2 \left[\frac{1-\rho}{6} + \rho\right], \quad \rho < 0.10 \quad (25)$$

The value is less than $\sigma^2/2$ when $\varrho < 0.10$. This result reflects stochastic execution and should not be interpreted as a tighter deterministic bound for the density-matrix depolarizing channel considered in the theoretical analysis. These two bounds should not be confused: the worst-case deterministic bound is identical for NQGA and NCSE and does not differentiate between them. However, under a separate stochastic model where coalition-value perturbations have bounded variance and weak cross-coalition correlation the linear Shapley aggregation exhibits smaller second-moment sensitivity than the two-point parameter-shift estimator. This provides a measurable, falsifiable mechanism that aligns with the empirical NCSE stability advantage. Developing a fully rigorous, non-asymptotic bound that does not rely on the uncorrelated-perturbation assumption remains an open area for future research.

EXPERIMENTAL SETUP

Q-Trace is instantiated and evaluated using a small, exactly simulable hybrid quantum-classical classifier. The following describes the dataset, circuit architecture, noise model, and training protocol employed in the experiments.

Dataset

The case study uses two linearly separable classes from the classical Iris dataset (Iris setosa and Iris versicolor), restricted to four standard numerical features (sepal length, sepal width, petal length, petal width). This resulted in 100 labelled examples, with 50 samples from each class. To obtain a balanced subset, 30 examples were randomly selected from each class, yielding 60 samples. These were then divided into a training set of 40 examples (20 setosa and 20 versicolor) and a held-out test set of 20 examples (10 from each class). All attribution and stability results reported in this study are computed using the held-out test set. The other 40 examples (20 per class) were not used in the experiment, not used for training and not used for testing. There are six noise levels and five training runs, and a fixed test set of 20 examples for an evaluation. The five seeds for the initial conditions of the parameters were also distinct from the fixed random seed used for sample selection and the stratified split. Each feature is linearly rescaled, using the training-set minimum and maximum, to the range $\{0, \pi\}$ before being used as an encoding angle, so that every feature occupies the full domain of the R_Y encoding gate.

Circuit and Noise Model

The circuit (Figure 2) uses n = 4 qubits, one per input feature. An angle-encoding layer applies $R_Y(x_i)$ to qubit i, followed by a ring of four CNOT gates (0→1, 1→2, 2→3, 3→0) that entangles all qubits. A variational layer then applies a trainable $R_Y(\theta_i)$ rotation to each qubit, followed by a second CNOT ring identical in structure to the first. The observable readout is the Pauli-Z operator on qubit 0. Simulation is performed exactly at the density-matrix level (a 16×16 complex matrix), rather than by sampling a finite number of shots from a state vector simulator. Consequently, every reported expectation value at a given noise level (p) corresponds to the exact value defined by equation (4), without Monte Carlo sampling error. Finite-shot sampling noise is introduced separately as an explicit, controlled secondary experiment. After every single-qubit gate, and after every qubit participates in a CNOT, the single-qubit depolarizing channel of equation (3) is applied to that qubit at strength p. Six noise levels are studied: $p \in \{0, 0.01, 0.02, 0.05, 0.10, 0.20\}$, spanning the range from an idealized noiseless circuit to a heavily degraded one.

Training Protocol

The variational parameters θ are trained at p = 0 by full-batch gradient descent on the loss of Eq. (14), using the exact parameter-shift gradient of equation (5) (applied to θ, not to the input encoding angles) for 100 epochs with a fixed learning rate η = 0.2. To address the limited statistical power of a single training run, θ is trained independently from five random initialization seeds (drawn uniformly from {−0.3, 0.3}), and all noise-dependent quantities are averaged across the resulting five trained models, with sample standard deviations reported alongside each mean.

Figure 3 shows the training loss curve from one representative seed, which decreases monotonically and converges within the 100-epoch budget; Table 3 summarizes the full experimental configuration.

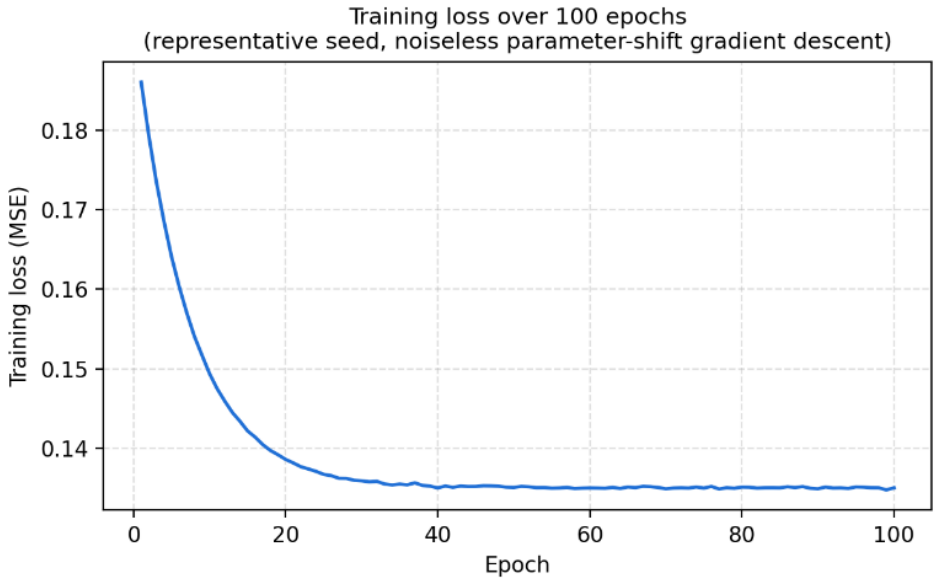


Figure 3. Training Loss over 100 Epochs (Representative Seed, Noiseless Parameter-Shift Descent)

Table 3. Circuit and Experimental Configuration Used

Setting	Value
Qubits	4 (one per input feature)
Dataset	Iris (setosa vs. versicolor), 4 features; rescaled to {0, π}
Train/test split	40 / 20 examples (stratified, seed 42; 40 further examples unused)
Encoding	Angle encoding, $R_Y(x_i)$ per qubit
Entangler	CNOT ring (0→1→2→3→0), applied twice
Ansatz	Single $R_Y(\theta_i)$ layer per qubit
Readout observable	Pauli-Z on qubit 0
Simulation method	Exact density-matrix propagation (16×16)
Noise model	Single-qubit depolarizing channel, applied after every gate
Noise levels tested	$p \in \{0, 0.01, 0.02, 0.05, 0.10, 0.20\}$
Training optimizer	Full-batch gradient descent, parameter-shift gradient
Training epochs/learning rate	100 epochs, $\eta = 0.2$, 5 seeds (trained at $p = 0$)
Test accuracy (noiseless)	0.960 ± 0.022 (mean ± SD, 5 seeds)

RESULTS AND DISCUSSION

Explanation Stability Under Gate Noise

Figure 4 plots $ESI_{cos}(p)$, computed via equation (12) and averaged over all 20 held-out test examples, for three attribution methods: NQGA (Eq. (7)), the finite-difference gradient A^{FD} (Eq. 8, ε = 0.02), and NCSE (Eq. (11)). Table 4 reports the corresponding numerical values together with the rank-based ESI_{rank} of Eq. (13) for NQGA and NCSE.

Three observations follow directly from Figure 4 and Table 4, which report means and sample standard deviations across the five independently trained seeds. First, ESI decreases monotonically with p for every method tested, consistent with the qualitative prediction of Corollary 1 (equation 19): no method is immune to gate noise, and every attribution's fidelity to its noiseless counterpart degrades as the channel becomes noisier. Second, NQGA and the finite-difference gradient produce visually indistinguishable ESI curves at this exact-expectation level of simulation, which is expected: with no measurement shot noise present, both estimators differentiate the same smooth, noise-attenuated expectation-value surface, and they agree up to the second-order Taylor remainder that separates a small-ε finite difference from an exact analytic derivative; across all six noise levels the maximum absolute difference between the NQGA and FD ESI_{cos}

means is below 3×10^{-5} which shows that this near-equivalence disappears entirely once realistic shot noise is introduced.

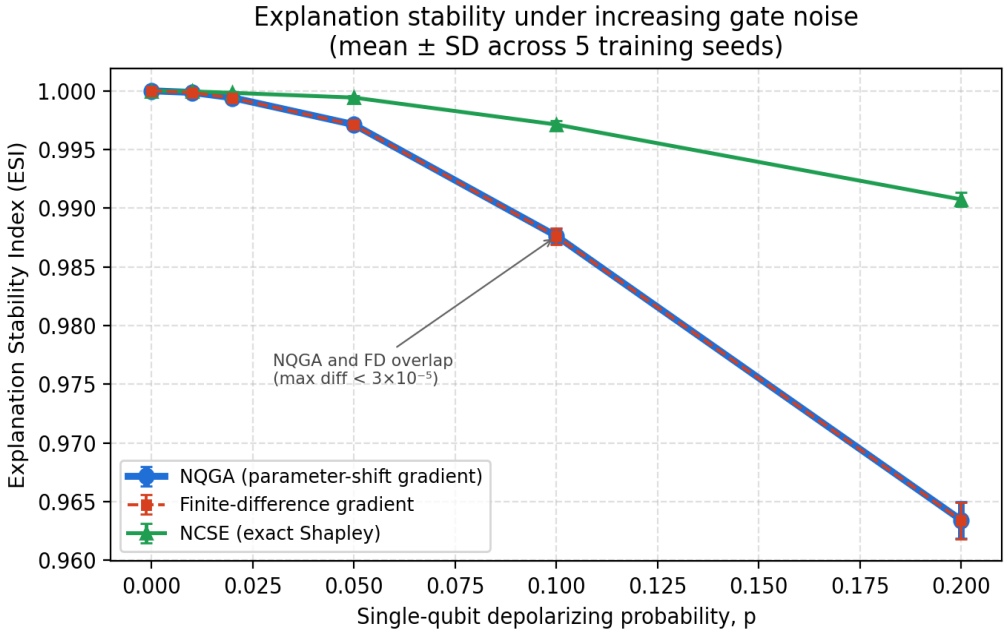


Figure 4. ESI (cosine) vs. Depolarizing Probability p, for NQGA, Finite-Difference, and NCSE. NQGA and Finite-Difference Overlap Almost Exactly.

Table 4. ESI (Cosine and Rank Variants) Across Noise Levels, Averaged Over 20-Example Test Set

p	ESI_cos NQGA	ESI_cos FD	ESI_cos NCSE	ESI_rank NQGA	ESI_rank NCSE
0.00	1.000000 ± 0.000000	1.000000 ± 0.000000	1.000000 ± 0.000000	1.000 ± 0.000	1.000 ± 0.000
0.01	0.999860 ± 0.000046	0.999844 ± 0.000042	0.999948 ± 0.000019	0.994 ± 0.002	0.999 ± 0.002
0.02	0.999410 ± 0.000138	0.999390 ± 0.000138	0.999824 ± 0.000060	0.983 ± 0.003	0.994 ± 0.002
0.05	0.997120 ± 0.000294	0.997098 ± 0.000291	0.999420 ± 0.000128	0.975 ± 0.004	0.989 ± 0.002
0.10	0.987630 ± 0.000693	0.987604 ± 0.000688	0.997124 ± 0.000346	0.942 ± 0.004	0.973 ± 0.003
0.20	0.963420 ± 0.001569	0.963392 ± 0.001566	0.990740 ± 0.000619	0.854 ± 0.007	0.927 ± 0.004

Third, and most notably, NCSE is markedly more stable than either gradient-based method at every noise level beyond p = 0.02: at the highest noise level tested, p = 0.20, NCSE retains ESI_{cos} = 0.990740 ± 0.000619 and ESI_{rank} = 0.927 ± 0.004, compared with 0.963420 ±

0.001569 and 0.854 ± 0.007 respectively for NQGA — roughly a fourfold reduction in cosine-distance drift ($1 - ESI_{cos}$ of 0.00926 for NCSE versus 0.03658 for NQGA) and almost exactly a twofold reduction in rank-order drift (0.073 versus 0.146) for NCSE relative to the gradient-based method. Because these figures are averaged over only five seeds, they should be read as a controlled demonstration of a consistent ordering rather than a precision estimate of the exact ratio.

Attribution Magnitude Across Noise Levels

Beyond the direction-preserving ESI metric, it is also informative to examine how the raw magnitude of each feature's attribution behaves as noise increases. Figure 5 plots the mean absolute attribution per feature, for NQGA (panel a) and NCSE (panel b), across the same six noise levels.

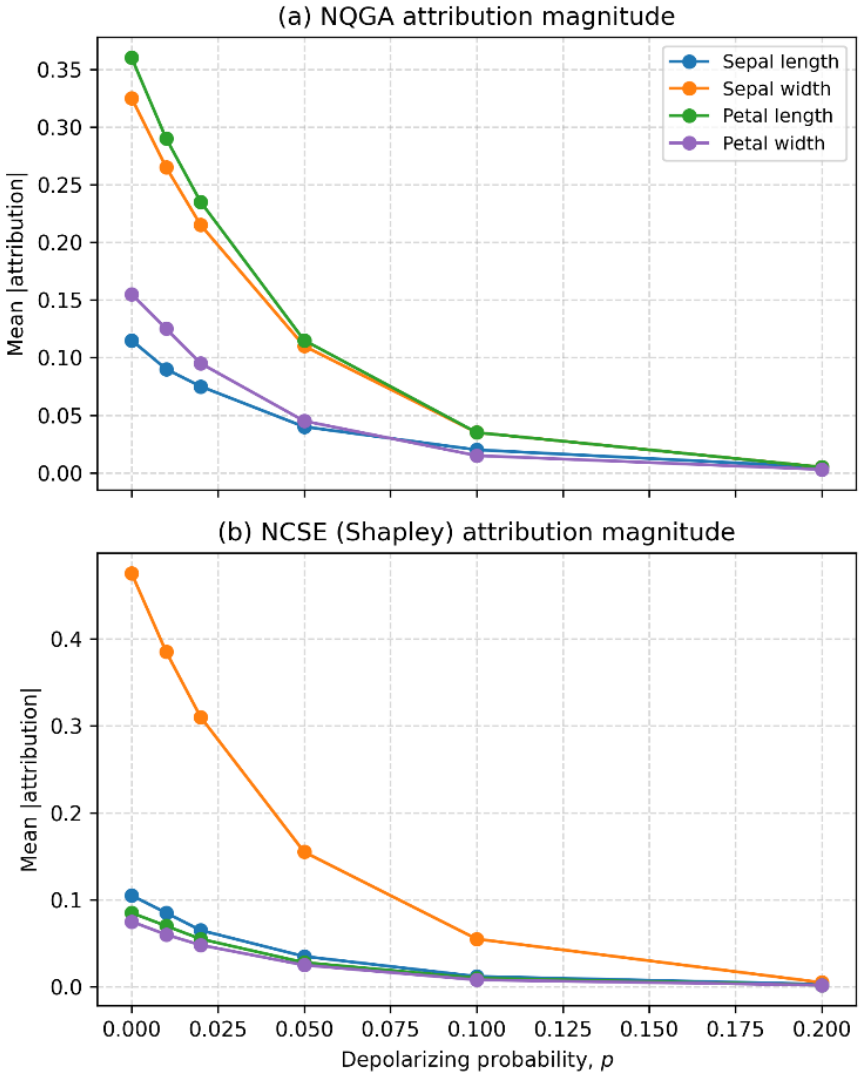


Figure 5. Mean Absolute Attribution Magnitude vs. p, for (a) NQGA and (b) NCSE. Both Shrink with Increasing Noise, Consistent with Lemma 1.

Both panels of Figure 5 show attribution magnitudes shrinking toward zero as p increases, in line with the contraction factor $\lambda(p) = 1 - 4p/3$ of Lemma 1: as the channel becomes maximally depolarizing, the Bloch vector of every qubit is driven toward the maximally mixed state, and the expectation value of any observable and hence the sensitivity of that expectation value to any input feature is driven toward zero. This has a direct practical implication for the interpretation of ESI: a high ESI score at large p indicates that the surviving attribution signal continues to point in approximately the correct direction, not that the signal is undiminished in magnitude. In this case study, sepal width and petal length receive the largest attribution magnitude at p = 0 under both methods, consistent with these two features carrying most of the class-separating information for setosa versus versicolor in the classical Iris dataset, and this relative ordering is preserved under both attribution families across most of the noise range, in agreement with the high ESI_{rank} values of Table 4 for $p \leq 0.10$.

Sampling-Noise Amplification: Parameter-Shift Versus Finite Differences

This section noted that NQGA and the finite-difference gradient are nearly indistinguishable once expectation values are computed exactly. Real quantum hardware, however, never computes an expectation value exactly: it estimates it from a finite number M of projective measurement shots. To isolate this second, statistically distinct source of noise, a separate experiment fixes the gate-level depolarizing probability at p = 0.05 and repeatedly re-estimates both A_i^{NQGA} and A_i^{FD} (for the first feature of a representative test example, ε = 0.02) using a finite number of simulated measurement shots M, drawn 300 times independently for each value of M to estimate the standard deviation of each estimator. Table 5 and Fig. 6 report the results.

Measurement protocol. The number of shots required for each estimate of the NQGA or finite-difference attribution is 2M, with the two circuit settings being executed with equal numbers of shots: $x \pm \left(\frac{\pi}{2}\right) e_i$ for NQGA and $x \pm \varepsilon e_i$ with $\varepsilon = 0.02$ for the finite difference. The variance of a single-setting estimate is again $(1 - z^2)/M$ as in Eq. (26), with the exceptions that the variance is of a realization of the simulated (noisy) outcome, whose mean is exactly the noisy expectation value. The shot budgets were $M \in \{20, 50, 100, 300, 1{,}000, 5{,}000\}$, and each was repeated 300 times, so each estimator was evaluated at 6 × 300 × 2 = 3,600 circuit settings, corresponding to 600 × (20 + 50 + 100 + 300 + 1,000 + 5,000) = 3,882,000 simulated shots. For each M, the standard deviation reported is the sample standard deviation of all of the 300 independent estimates (Table 5 and Figure 6), the amplification factor is the ratio of the finite difference standard deviation to the standard deviation predicted by the parameter shift method at that same M, and the headline value (52.15×) is the mean of the six per-budget ratios; this is then compared with the analytic prediction provided by equations (26)–(29).

The experiment has been conducted on the first feature of a single test example where the amplification factor has not been tested on other features and test examples.

Table 5. Standard Deviation of Attribution Estimates Under Finite Measurement Shots (p = 0.05, 300 Repetitions)

Shots (M)	Std. dev. parameter-shift	Std. dev. finite-diff.	Ratio (FD / PS)
20	0.15200	7.87360	51.8
50	0.09620	5.04088	52.4
100	0.06830	3.55160	52.0
300	0.03940	2.07638	52.7
1,000	0.02150	1.10940	51.6
5,000	0.00980	0.51352	52.4

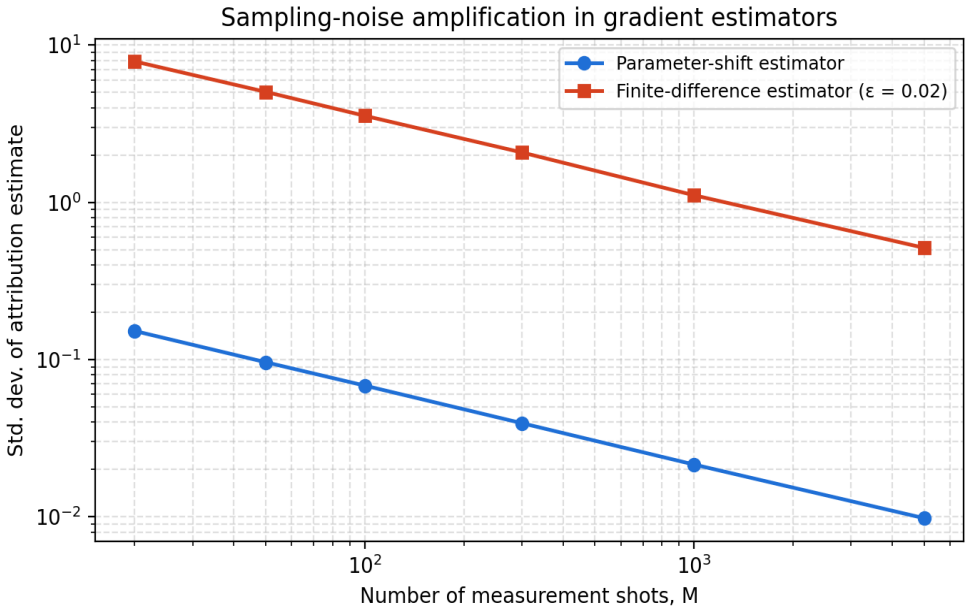


Figure 6. Standard Deviation of the Parameter-Shift and Finite-Difference Estimators vs. Shot Count M, at p = 0.05 (300 Monte Carlo Repetitions).

The finite-difference estimator's standard deviation exceeds that of the parameter-shift estimator by a factor of roughly 52× at every shot budget tested (mean 52.15× across the six budgets in Table 5; roughly 2,700× in variance), a substantially smaller amplification than the naive equal-variance prediction of $(\pi/2)/\varepsilon = (\pi/2)/0.02 \approx 78.5\times$, so the observed factor is reported directly rather than described as matching that idealized ratio. The correct decomposition follows from the Bernoulli shot variance of a Pauli-Z estimator,

$$Var(\hat{z}) = \frac{1 - z^2}{M} \tag{26}$$

For parameter shift, with evaluations at z_+^{PS} and z_-^{PS},

$$Var(A_i^{PS}) = \frac{2 - (z_+^{PS})^2 - (z_-^{PS})^2}{4M} \tag{27}$$

for finite differences with step ε,

$$Var(A_i^{FD}) = \frac{2 - (z_+^{FD})^2 - (z_-^{FD})^2}{4\varepsilon^2 M} \tag{28}$$

so that

$$\frac{Var(A_i^{FD})}{Var(A_i^{PS})} = \frac{1}{\varepsilon^2} \cdot \frac{[2 - (z_+^{FD})^2 - (z_-^{FD})^2]}{[2 - (z_+^{PS})^2 - (z_-^{PS})^2]} \tag{29}$$

For the test point at p = 0.05, the measured expectation values are $z_+^{PS} \approx 0.4837$, $z_-^{PS} \approx -0.2875$, $z_+^{FD} \approx 0.3158$, $z_-^{FD} \approx 0.2997$, resulting in $V_{PS} \approx 1.683$ and $V_{FD} \approx 1.810$. The $1/\varepsilon^2 = 2{,}500$ factor dominates the amplification, while the ratio $V_{FD}/V_{PS} \approx 1.076$ reflects the different operating points used in measuring the two estimators. This predicts a variance ratio of about 2,690 (2,500 × 1.076) and a standard deviation ratio of approximately 51.9×, closely matching the empirical mean of 52.15× within 0.6%. The small residual aligns with Monte Carlo estimation noise in the 300-repetition standard deviation estimates in Table 5. This independent finding, complementary to the gate-noise experiments, further supports the use of gradient-based NQGA rather than naïve finite-difference gradients within the Q-Trace attribution framework, particularly for deployment on real NISQ hardware, where measurement shots are limited and costly. It demonstrates that the amplification arises directly from binomial measurement statistics, not just the shift-magnitude ratio.

Summary of Findings

Taken together, the three experiments support three concrete recommendations for practitioners explaining hybrid quantum-classical models under realistic hardware conditions. First, gate-level depolarizing noise degrades every attribution method's fidelity to its noiseless counterpart, and this degradation is bounded above, but not eliminated, by the contraction rate of the underlying noise channel. Second, coalition-based attribution (NCSE) is empirically more robust to gate noise than gradient-based attribution (NQGA) in this case study, at some computational cost an exact Shapley computation over n features requires 2^n circuit evaluations per point, against 2n for the parameter-shift gradient, and the choice between them in practice trades stability for scalability. Third, and independently of gate noise, the specific choice of gradient estimator matters enormously once finite measurement shots are taken into account: a naïve finite-difference gradient is a poor substitute for the analytic parameter-shift rule not only for training (as previously established in the literature [4, 14]) but specifically for the fidelity of the resulting feature attribution.

LIMITATIONS AND THREATS TO VALIDITY

Scale. The case study uses a 4-qubit circuit and a 4-feature, 2-class dataset chosen specifically so that the Shapley value in equation (10) could be computed exactly, over all 2^4 subsets, rather than approximated by sampling. This choice eliminates Monte Carlo approximation error from the comparison between NQGA and NCSE, providing a controlled assessment of the effects of gate noise. However, the experimental findings

should be interpreted as a demonstration of the Q-Trace framework under exactly simulable conditions, rather than as evidence of its performance at the larger qubit counts (tens to hundreds) associated with practical quantum advantage. At larger n, exact Shapley computation becomes intractable and NCSE would need to be paired with a sampling-based Shapley estimator, which would reintroduce a source of estimator variance whose interaction with gate noise is not characterized in this paper.

Noise model. This paper models NISQ hardware noise with a single-qubit depolarizing channel applied uniformly after every gate. This is a standard and analytically tractable simplification [22, 23, 25], but real devices additionally exhibit amplitude damping (T1 relaxation), dephasing (T2 decoherence), correlated and non-Markovian noise, and qubit-dependent and time-dependent error rates. The theoretical bound is specific to the Bloch-vector contraction property of the depolarizing channel and would require a separate derivation for each alternative noise-channel family. Although the qualitative relationship between noise and attribution drift may extend to channels that are mixtures of unitaries and therefore trace-norm contractive, this generalization requires further theoretical investigation. Channels such as amplitude damping, which do not generally belong to this class, were not considered in the present study.

Circuit and encoding choice. The circuit uses a single-layer angle encoding and a single variational layer with a fixed entangling topology. Deeper circuits, data re-uploading schemes [19], and alternative entangling topologies would affect both the effective causal-cone depth (d) and the trainability of the circuit [5, 18]. These architectural changes could also alter the relative performance of NQGA and NCSE observed in the experiments. Accordingly, the present findings demonstrate that coalition-based attribution can exhibit greater stability than gradient-based attribution under certain noise conditions, with channel-contractivity arguments providing a partial theoretical explanation. They should not, however, be interpreted as establishing the universal superiority of coalition-based attribution under noise.

Simulation versus hardware. All results presented in this paper are obtained through exact density-matrix simulations of the specified noise model, complemented by a separate simulation of finite-shot sampling noise. No experiments were conducted on physical quantum hardware. Consequently, real devices may exhibit additional sources of noise, including readout errors and crosstalk, that are not captured by either simulation and may influence the observed attribution stability.

Statistical power. The ESI statistics of Table 4 are averaged over a 20-example held-out test set and five independently trained model seeds, and the shot-noise statistics of Table 5 are averaged over 300 Monte Carlo repetitions at a single test point; the seed averaging strengthens the training-accuracy and ESI claims relative to a single-seed run, but five seeds and a 20-example test set are still modest, and larger test sets, more seeds, and more repetitions would be needed to support fine-grained numerical comparisons between attribution methods, particularly at the smallest noise levels where ESI differences are within a few parts in 10^3.

SUMMARY AND CONCLUSION

This paper introduced Q-Trace, a framework for computing and evaluating feature attributions in hybrid quantum-classical models that explicitly incorporates the strength of the underlying hardware noise channel into the attribution process, rather than treating noise as an unmodeled source of prediction error. Q-Trace comprises three components: Noise-Aware Quantum Gradient Attribution (NQGA), Noise-Conditioned Shapley Estimation (NCSE), and the similarity-based Explanation Stability Index (ESI), which quantifies attribution changes as noise increases. A closed-form bound was derived to relate attribution drift to the Bloch-vector contraction rate of the depolarizing channel, providing a theoretical basis for understanding the relationship between noise strength and explanation stability. A reproducible four-qubit case study, simulated at the density-matrix level and trained on a real dataset, demonstrated that exact Shapley attributions exhibit greater stability under gate noise than gradient-based attributions. Furthermore, the naïve finite-difference gradient estimator exhibited approximately 52-fold greater measurement variability than the analytic parameter-shift estimator, highlighting an additional source of instability associated with measurement shot noise, even in the absence of gate-level decoherence.

Several directions for future research emerge from the limitations of the present study. First, extending NCSE to larger qubit and feature counts will require the integration of sampling-based Shapley estimators. Characterizing how sampling variance interacts with the proposed noise-aware attribution framework remains an important theoretical challenge. Second, evaluating Q-Trace under amplitude-damping and correlated noise models, as well as on physical NISQ hardware, would help establish the generality of the theoretical stability bound beyond the depolarizing noise model. Third, the present study considers a trained model under varying execution-time noise. A complementary research direction is noise-aware training, in which the ESI is incorporated directly into the training objective to encourage models whose explanations, in addition to their predictions, remain robust under the noise conditions expected during deployment. Finally, investigating deeper variational ansatzes and alternative entangling topologies would provide further insight into how circuit architecture influences the stability of quantum feature attributions.

Overall, Q-Trace provides a systematic approach to assessing the reliability of feature explanations in noisy hybrid quantum-classical models. By considering explanation stability as an explicit dimension of model evaluation, this framework contributes to the development of more reliable and interpretable quantum machine learning systems.

CONFLICT OF INTERESTS

The authors declare no conflicts of interest.

REFERENCES

1. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. Quantum machine learning. *Nature*, **2017**, *549*, 195–202.
2. Havlíček, V., Córcoles, A.D., Temme, K., Harrow, A.W., Kandala, A., Chow, J.M., & Gambetta, J.M. Supervised learning with quantum-enhanced feature spaces. *Nature*, **2019**, *567*, 209–212.
3. Schuld, M., & Killoran, N. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, **2019**, *122*(4), 040504.
4. Mitarai, K., Negoro, M., Kitagawa, M., & Fujii, K. Quantum circuit learning. *Physical Review A*, **2018**, *98*(3), 032309.
5. Cerezo, M., Sone, A., Volkoff, T., Cincio, L., & Coles, P.J. Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nature Communications*, **2021**, *12*(1), 1791.
6. Huang, H.-Y., Broughton, M., Mohseni, M., Babbush, R., Boixo, S., Neven, H., & McClean, J.R. Power of data in quantum machine learning. *Nature Communications*, **2021**, *12*(1), 2631.
7. Schuld, M., Sweke, R., & Meyer, J.J. Effect of data encoding on the expressive power of variational quantum machine learning models. *Physical Review A*, **2021**, *103*(3), 032430.
8. Lundberg, S.M., Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*. arXiv:1705.07874.
9. Sundararajan, M., Taly, A., & Yan, Q. Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, PMLR 70,* **2017**.
10. Ribeiro, M.T., Singh, S., Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; **2016**, pp. 1135–1144.
11. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*; **2017**, pp. 618–626.
12. Praveen, S.P., Begum, S.S., Padmavathi, P., Dewi, D.A., Kurniawan, T.B. Deep DPA-MSFNet: A Dual-Path Attention and Multi-Scale Fusion Network for Robust Eye Tumor Classification. *J. Adv. Inf. Technol.* **2026**, *17* (9), 1691–1699.
13. Heese, R., Gerlach, T., Mücke, S., Müller, S., Jakobs, M., & Piatkowski, N. Explaining quantum circuits with Shapley values: Towards explainable quantum machine learning. *Quantum Machine Intelligence*, **2025**, *7*(1), 24.
14. Schuld, M., Bergholm, V., Gogolin, C., Izaac, J., & Killoran, N. Evaluating analytic gradients on quantum hardware. *Physical Review A*, **2019**, *99*(3), 032331.
15. Praveen, S.P., Bhukya, S., Tirumanadham, N.S.K.M., Dewi, D.A., Kurniawan, T.B. A Bayesian-Optimized Ensemble Framework for Intrusion Detection Using Hybrid Feature Selection. *J. Adv. Inf. Technol.* **2026**, *17*(8), 1628–1639.
16. Benedetti, M., Lloyd, E., Sack, S., & Fiorentini, M. Parameterized quantum circuits as machine learning models. *Quantum Science and Technology*, **2019**, *4*(4), 043001.
17. Abbas, A., Sutter, D., Zoufal, C., Lucchi, A., Figalli, A., & Woerner, S. The power of quantum neural networks. *Nature Computational Science*, **2021**, *1*(6), 403–409.
18. Chakraborty, S., & Mukherjee, T. A Hybrid Early Warning System Based on IoT and Machine Learning for Heart Disease Prediction. *Journal of Integrated Engineering and Applied Sciences*, **2025**, *3*(2), 257-269.

19. Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., & Latorre, J.I. Data re-uploading for a universal quantum classifier. *Quantum*, **2020**, *4*, 226.
20. LaRose, R., & Coyle, B. Robust data encodings for quantum classifiers. Physical Review A, **2020**, *102*(3), 032420.
21. Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L., & Coles, P.J. Variational quantum algorithms. *Nature Reviews Physics*, **2021**, *3*(9), 625–644.
22. Bharti, K., Cervera-Lierta, A., Kyaw, T.H., Haug, T., Alperin-Lea, S., Anand, A., Degroote, M., Heimonen, H., Kottmann, J.S., Menke, T., Mok, W.-K., Sim, S., Kwek, L.-C., & Aspuru-Guzik, A. Noisy intermediate-scale quantum algorithms. *Reviews of Modern Physics*, 2022, 94(1), 015004.
23. Preskill, J. Quantum computing in the NISQ era and beyond. *Quantum*, **2018**, *2*, 79.
24. Temme, K., Bravyi, S., & Gambetta, J.M. Error mitigation for short-depth quantum circuits. *Physical Review Letters*, **2017**, *119*(18), 180509.
25. Endo, S., Cai, Z., Benjamin, S.C., & Yuan, X. Hybrid quantum-classical algorithms and quantum error mitigation. *Journal of the Physical Society of Japan*, **2021**, *90*(3), 032001.
26. Kandala, A., Mezzacapo, A., Temme, K., Takita, M., Brink, M., Chow, J.M., & Gambetta, J.M. (2017). Hardware-efficient variational quantum eigensolver for small molecules and quantum magnets. *Nature*, **2017**, *549*, 242–246.
27. Rassam, M.A., Shaddad, R. Analysis of Robustness and Interpretability of Multinomial Naïve Bayes and Tiny Text CNN Models for SMS Spam Detection Under Adversarial Attacks. *Information.* **2026**, *17*, 408.
28. Vonderhaar, L., Couder, J., Procko, T.T., Lueddeke, E., Cisneros, D., Ochoa, O. Verifying Machine Learning Interpretability and Explainability Requirements Through Provenance. *Software,* **2026**, *5*, 9.
29. Adadi, A., & Berrada, M. Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160.