Journal of Electronic Systems

DLINE JOURNALS

Print ISSN: 2278 – 652X
Online ISSN: 2278 – 6538

JES 2026: 16 (4)
https://doi.org/10.6025/jes/2026/16/4/229-251

Predicting and Interpreting Quantum Circuit Reliability: An Integrated Machine Learning and XAI Framework

Duong Van Hieu
Tien Giang University
Vietnam
duongvanhieu@tgu.edu.vn

ABSTRACT

Quantum computing promises exponential advantages for classically intractable problems but remains severely bottlenecked by hardware noise. Accurately predicting quantum circuit reliability prior to execution is crucial for effective compilation and error mitigation, yet exact fidelity estimation is computationally prohibitive. While recent machine learning (ML) approaches have attempted to predict circuit reliability, they often suffer from topological blindness and operate as uninterpretable "black boxes." In this study, we propose an Integrated Machine Learning and Explainable AI (XAI) framework for pre-execution quantum circuit analysis. Using the QUASAR Quantum Reliability Dataset v1 (50,000 synthetic circuits), we evaluate tree-based ensemble models for both continuous reliability regression and categorical classification. Our results demonstrate that HistGradientBoosting achieves superior predictive performance ($R^2 \approx 0.945$ for regression; 88.86% accuracy for classification) under rigorous grouped cross-validation. Furthermore, unsupervised clustering reveals four distinct "reliability archetypes," proving that circuit reliability is driven by a multidimensional interplay of qubit count, depth, and gate composition rather than a single metric. Post-hoc explainability techniques (SHAP, Permutation Importance, and Partial Dependence Plots) identify qubit count, circuit depth, and measurement-related features as primary drivers of reliability degradation. Finally, we critically examine the methodological caveats of synthetic data, target determinism, and feature multicollinearity, establishing a foundational baseline for interpretable quantum software engineering while highlighting the necessity for future hardware-level validation.

Keywords: Quantum Computing; Circuit Reliability; Explainable AI (XAI); Machine Learning; Quantum Noise; SHAP; Predictive Modeling; Quantum Compilation; NISQ Devices

Received: 11 May 2026, Revised: 30 July 2026, Accepted: 2 September 2026

Copyright: **DLINE**

1. Introduction

Quantum computing (QC) represents a novel computational paradigm with the potential to address

classically intractable problems with significantly higher efficiency and speed. It has demonstrated exponential or polynomial advantages across various domains, including combinatorial optimization, molecular dynamics [1, 2], and machine learning [3-9]. Despite these computational advantages, the practical deployment of QC is heavily hindered by hardware noise. These errors severely interfere with the functionality of quantum circuits and form significant obstacles to their further optimization.

2. Background

2.1 Noise Mitigation and Circuit Robustness

To attenuate the negative effects of hardware noise, a number of noise mitigation techniques have been developed [4, 10-15]. For instance, Hanrui Wang and Jiaqi Gu [15] proposed a framework to improve the robustness of quantum circuits by making them inherently noise-aware. Their framework consists of three primary techniques: the injection of gate errors, regularization, and the normalization of measurement outcomes. Similarly, another study by Zhiding Liang [16]integrated gate error characteristics directly into the mapped quantum circuit to enhance overall robustness.

2.2 Fidelity Estimation and Reliability Prediction

In order to validate and characterize the states generated by a quantum computer, it is crucial to accurately estimate the fidelity of quantum states [17, 18]. However, calculating fidelities is highly computationally expensive. To address this, numerous efforts have been made in recent years. Variational quantum algorithms have been adopted in recent works to perform fidelity estimation [19, 20, 21], while machine learning-based and statistical methods have also been proposed to estimate fidelity more efficiently [22].

Alongside fidelity estimation, predicting overall circuit reliability is essential. Early works attempted to predict circuit reliability using simple machine learning models [22]. However, these traditional approaches consider neither the graph topology of the circuit nor specific noise information, which yields less accurate predictions in experimental results [23].

2.3 Advanced Machine Learning and Graph Neural Network Approaches

To overcome the limitations of traditional models, advanced machine learning schemes have been introduced to capture complex circuit dynamics. Saravanan proposed a machine learning (ML)-based scheme to detect incorrect quantum circuit outputs and predict the Probability of Successful Trials (PST) with high accuracy. This approach allows for the verification of quantum hardware error rates and the validation of the correctness of extracted quantum circuit outputs [24].

Building upon this, Saravanan[25] later proposed an ML-based scheme to predict the output fidelity of quantum circuits executed on Noisy Intermediate Scale Quantum (NISQ) devices. This work demonstrated the distinct benefits of using Graph Neural Network (GNN)-based models compared to traditional ML models. By capturing the quantum circuit's structural graph in addition to its gates' features, GNNs enable the characterization of unpredicted quantum circuit errors [25]. Furthermore, alternative analytical approaches have been explored, such as the work by M. Saghafi, who analyzed and predicted the output patterns of single-qubit quantum circuits by treating the results of repeated executions as a time series [26].

2.4 Research Problem Statement

Despite the exponential and polynomial advantages of Quantum Computing (QC) in domains such as combinatorial optimization, molecular dynamics, and machine learning, its practical deployment is severely bottlenecked by hardware noise. These errors interfere with quantum circuit functionality and impede further optimization. While estimating the fidelity of quantum states is crucial for validating and characterizing generated states, exact fidelity calculation is computationally prohibitive. Early attempts to predict circuit reliability using simple machine learning (ML) models failed to account for circuit graph topology and specific noise characteristics, resulting in inaccurate predictions. Furthermore, advanced ML models that do capture these complex dynamics often operate as "black boxes," lacking the interpretability required to inform heuristic circuit compilation and optimization strategies. Therefore, there is a critical need for a pre-execution reliability

prediction framework that not only accurately forecasts quantum circuit reliability based on structural features but also provides interpretable insights into the drivers of reliability degradation.

2.5 Research Issues and Framework

2.51. Research Issues

This study addresses several critical challenges at the intersection of quantum software engineering and machine learning:

1. Computational Bottleneck of Fidelity Estimation: Direct simulation and calculation of quantum state fidelity are highly resource-intensive, necessitating efficient proxy models for pre-execution reliability prediction.

2. Topological Blindness in Baseline Models: Traditional ML approaches often treat quantum circuits as flat feature vectors, ignoring the graph topology, entanglement metrics, and structural depth that fundamentally dictate noise susceptibility.

3. Interpretability of Black-Box Models: While advanced models like Graph Neural Networks (GNNs) and deep ensembles capture complex circuit dynamics, their opaque nature limits their utility for quantum compiler designers who need to understand *why* a circuit is deemed unreliable.

4. Multicollinearity and Feature Redundancy: Quantum circuit features (e.g., gate counts, qubit width, depth) are highly correlated. Standard feature importance metrics often fail to disentangle these overlapping variables, leading to ambiguous or misleading explanations.

5. Target Leakage in Synthetic Datasets: In simulation derived datasets, the target variable (reliability score) is often a deterministic rescaling of estimated fidelity. Failing to strictly audit and remove proxy variables can result in artificially inflated model performance that does not generalize to physical hardware.

2.6 Proposed Framework

Figure 1 illustrates the proposed analytical workflow for evaluating quantum circuit reliability, visually mapping the four core phases of our integrated Machine Learning and Explainable AI (XAI) framework.

The workflow begins at the Data Curation and Preprocessing stage, where the raw QUASAR dataset undergoes rigorous feature auditing, target leakage prevention, and grouped partitioning based on qubit count and circuit depth. The processed data then flow into the Predictive Modeling and Evaluation stage, which incorporates multiple tree based ensemble algorithms (e.g., *HistGradientBoosting*, Random Forest, Extra Trees) for both continuous reliability regression and categorical classification.

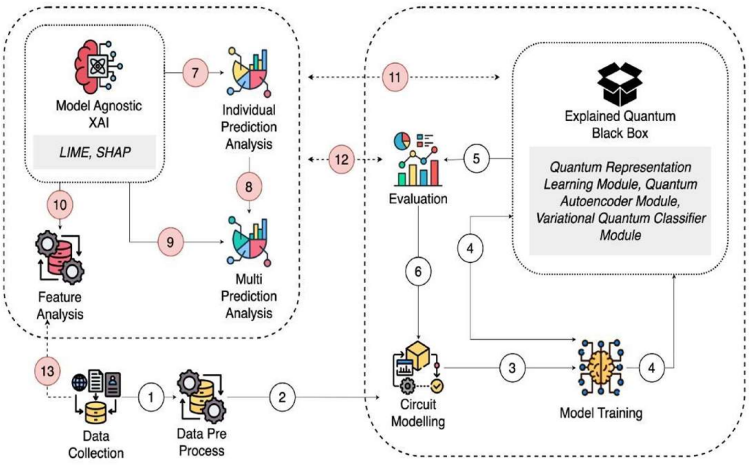

Figure 1. Analytical Workflow

Parallel to supervised prediction, the workflow includes an Unsupervised Archetype Discovery block, where clustering techniques are applied to group circuits into distinct reliability archetypes based on their multidimensional structural profiles.

Finally, the framework culminates in the Model-Agnostic Explainability stage. In this final block, interpretation methods such as SHAP, LIME, Permutation Importance, and Partial Dependence Plots (PDP) are deployed. The resulting explanations identify the most critical circuit characteristics, bridging the gap between black-box predictions and the physical structural properties of quantum circuits. Ultimately, this structured workflow ensures that reliability prediction is seamlessly integrated with actionable, interpretable insights for heuristic quantum circuit compilation.

3. Methodology

The methodology employs a multi-tiered analytical approach combining supervised learning, unsupervised clustering, and post-hoc explainability techniques.

1. Feature Auditing and Preprocessing: To prevent data leakage, the target variable (reliability_score, defined as 100 x estimated _ fidelity) and any highly correlated proxy variables are strictly excluded from the predictive feature space. The remaining structural and gate-count features are standardized for clustering and evaluated for multicollinearity.

2. Supervised Predictive Modeling:
o *Regression:* Tree-based ensemble models are trained to predict the continuous reliability score.

o *Classification:* The continuous scores are discretized into LOW, MEDIUM, and HIGH reliability classes, and the same ensemble architectures are evaluated for categorical prediction.

3. Unsupervised Reliability Archetypes: To understand circuit heterogeneity beyond simple regression, *MiniBatch K-Means* clustering is applied to the standardized structural features. This identifies distinct circuit "archetypes" to analyze how combinations of qubit count, depth, and gate composition correlate with mean reliability.

4. Explainable AI (XAI) Integration: To demystify the predictive models, we employ multiple model-agnostic and tree specific interpretation methods:

o *Permutation Importance:* To evaluate the global drop in model performance when specific structural features are randomized.

o *SHAP (Global and Summary):* To quantify the magnitude and direction of individual feature contributions at both the global and local (observation) levels.

o *Partial Dependence Plots (PDP):* To map the marginal, average effect of specific features (e.g., qubit count, entangling gates) on the predicted reliability score.

4. Experimental Setup

The experiments are conducted using the QUASAR Quantum Reliability Dataset v1, a synthetic, simulation-derived dataset comprising 50,000 independent quantum circuits and 37 columns (36 structural/gate-count features and 1 target variable).

4.1 Dataset Description

The QUASAR v1 dataset is distributed as a single comma separated values (CSV) file (quantum_ reliability_dataset.csv), comprising over 50,000 independent samples and 37 columns.

• Feature Space (Columns 1–36): These columns encapsulate measurable numerical characteristics of each quantum circuit. While the complete variable definitions are detailed in the accompanying feature_dictionary.md and column_descriptors.md files provided in the repository, typical features in this domain include qubit count, circuit depth, total gate count, specific gate frequencies (e.g., CNOT, single-qubit rotations like U3/H), and entanglement metrics.

• Target Variable (Column 37): The final column contains the continuous or normalized simulated reliability score, representing the circuit's expected fidelity under the applied noise model.

To ensure rigorous evaluation and prevent structural data leakage, a grouped train/validation/test split is utilized rather than random sampling. Groups are defined by the cross-product of *qubit count* and *circuit-depth quintile*. This guarantees that no specific combination of qubit width and depth appears in both the training and testing phases.

• Training Set: 34,846 samples

• Validation Set: 6,061 samples

• Test Set: 9,093 samples

4.2 Algorithms and Hyperparameters

Three tree-based ensemble algorithms are evaluated for both regression and classification tasks due to their robustness in handling non-linear relationships and feature interactions:

1. HistGradientBoosting (Histogram-based Gradient Boosting)

2. Random Forest

3. Extra Trees (Extremely Randomized Trees)

4.3 Evaluation Metrics

• Regression Metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2).

• Classification Metrics: Accuracy, Balanced Accuracy, Macro-F1, Weighted-F1, Macro-Precision, and Macro-Recall. Balanced accuracy and macro-averaged metrics are prioritized to account for any disparities in class distributions (e.g., the predominance of the 'LOW' reliability class).

4.4 Explainability Configuration

• SHAP Analysis: Computed using an *Extra Trees* model trained on a representative subset of the training data and explained on held-out validation samples to ensure computational efficiency while maintaining statistical reliability.

• Clustering: *MiniBatch K-Means* is executed on standardized structural features to extract four distinct reliability archetypes, which are subsequently analyzed for mean reliability scores and structural heatmaps.

5. Quantum Circuit Reliability Analysis: Results and Deliverables

We analyzed the uploaded QUASAR Quantum Reliability Dataset and generated a research-oriented analysis package covering data quality, exploratory statistics, circuit complexity, regression, classification, grouped validation, model explainability, and reliability archetypes.

The dataset contains 50,000 quantum circuits and 37 columns. Because it is synthetic and simulation-derived, the findings describe the generated dataset rather than validated performance on physical quantum hardware.

Data Quality Assessment

Check	Result
Rows	50000
Columns	37
Missing cells	0
Duplicate rows	0
Duplicate circuit_name	0
Nonpositive qubits	0
Negative numeric values	0
Score = 100 × estimated_fidelity	50000

Table 1. Data Quality Assessment

Table 1 summarizes the structural integrity and numerical consistency of the QUASAR quantum reliability dataset. The dataset contains 50,000 observations and 37 columns, with no missing cells, duplicate rows, or duplicate circuit identifiers. Furthermore, the assessment identified no nonpositive qubit counts or negative numerical values. The final quality check confirmed that the reliability score equals 100 times the estimated fidelity for all 50,000 records, within floating-point precision.

These findings indicate that the dataset is complete and internally consistent with respect to the checks performed. The absence of missing values and duplicated records reduces the need for basic data-cleaning operations before statistical analysis and machine-learning modelling. However, structural completeness does not necessarily guarantee that every variable is physically meaningful or independently informative. In particular, the exact relationship between the reliability score and estimated fidelity reveals a deterministic target dependency. Accordingly, estimated fidelity and other variables directly associated with the target must be excluded from the predictive inputs to prevent data leakage and artificially inflated performance estimates.

5.1 Regression: predicting *reliability_score*

Models were evaluated using grouped train, validation, and test partitions. Groups were defined using qubit count and circuit depth quintile, preventing those group combinations from appearing in both training and test sets.

Model	Test RMSE ↓	Test MAE ↓	Test R^2 ↓
HistGradientBoosting	4.858	2.848	0.945
Random Forest	5.129	2.926	0.938
Extra Trees	5.226	2.992	0.936

Table 2. Regression Performance

Table 2 compares the test-set performance of three regression models—HistGradientBoosting, Random Forest, and Extra Trees—for predicting the continuous reliability score. The models were assessed using root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (\(R^2\)).

HistGradientBoosting achieved the strongest overall performance, with an RMSE of 4.858, an MAE of 2.848, and an \(R^2\) of 0.945. Random Forest produced an RMSE of 5.129, an MAE of 2.926, and an \(R^2\) of 0.938. Extra Trees recorded an RMSE of 5.226, an MAE of 2.992, and an \(R^2\) of 0.936.

The results indicate that HistGradientBoosting provides the most accurate predictions among the three evaluated models on the grouped test set. Its lower RMSE and MAE indicate smaller prediction errors, while the higher \(R^2\) indicates that it explains a greater proportion of the observed variation in the reliability score. Nevertheless, these findings should be interpreted within the limits of the synthetic data and the specified validation procedure. They do not establish equivalent performance on previously unseen circuit families or real quantum hardware.

5.2 Classification: predicting reliability_class

The target contains three classes: LOW, MEDIUM, and HIGH. *HistGradientBoosting* performed best among the evaluated classifiers.

Test metric	***HistGradient Boosting***
Accuracy	88.86%
Balanced accuracy	88.77%
Macro-F1	0.888
Weighted-F1	0.889
Macro precision	0.888
Macro recall	0.888

Table 3. Classification Performance

Table 3 presents the test set classification results for predicting the three reliability categories: LOW, MEDIUM, and HIGH. *HistGradientBoosting* achieved an accuracy of 88.86% and a balanced accuracy of 88.77%. Its macro-F1 and weighted-F1 scores were 0.888 and 0.889, respectively, while macro-precision and macro-recall were approximately 0.888.

The close agreement between accuracy and balanced accuracy suggests that the model's overall performance is not substantially inflated by differences in class representation. Similarly, the small difference between macro-F1 and weighted-F1 indicates that the reported aggregate classification performance remains broadly consistent under equal class and support weighted evaluation.

These findings suggest that *HistGradientBoosting* can distinguish the defined reliability categories with relatively strong accuracy within the evaluation dataset. However, interpretation depends on how LOW, MEDIUM, and HIGH are defined. If these categories are generated by applying thresholds to the continuous reliability score, classification represents a discretized version of the regression target rather than an

independent reliability outcome. The category thresholds and class-wise error patterns should therefore be documented before drawing broader conclusions.

Model	validation_accuracy	validation_macro_F1	validation_weighted_F1	test_accuracy	test_balanced_accuracy	test_macro_F1	test_weighted_F1	test_recall_macro	test_precision_macro
HistGradient Boosting	0.916020459	0.783179208	0.914730881	0.888595623	0.887728193	0.887897247	0.888975709	0.887728193	0.888265156
Random Forest	0.90199637	0.728344366	0.897393293	0.877378203	0.874578257	0.876069891	0.877475314	0.874578257	0.881111448
Extra Trees	0.905296156	0.742559308	0.902476669	0.859012427	0.855697277	0.857566305	0.859364078	0.855697277	0.864320566

Table 4 provides a more detailed comparison of *Hist Gradient Boosting*, Random Forest, and Extra Trees, including validation and test performance. *Hist Gradient Boosting* achieved a validation accuracy of approximately 91.60%, a validation macro-F1 of 0.783, and a validation weighted F1 of 0.915. On the test set, it achieved an accuracy of 88.86%, balanced accuracy of 0.888, macro F1 of 0.888, and weighted F1 of 0.889.

Random Forest obtained a test accuracy of approximately 87.74%, while Extra Trees achieved approximately 85.90%. *Hist Gradient Boosting* therefore ranked first in test accuracy and the reported aggregate F1 metrics. Its performance advantage suggests that its boosting based learning strategy captures useful nonlinear relationships among structural circuit characteristics and gate count features.

The difference between validation and test macro-F1 for *HistGradientBoosting* is also noteworthy. The validation macro F1 of approximately 0.783 is substantially lower than the test macro-F1 of approximately 0.888, even though the validation accuracy is higher. This discrepancy warrants examination of class-specific precision, recall, class frequencies, and the exact evaluation procedure. It should not be interpreted as evidence of improvement between validation and testing without further investigation.

5.3 Reliability archetypes

Four circuit archetypes were identified using MiniBatch K-means clustering on standardized structural and gate count features.

Archetype	Circuits	Mean qubits	Mean depth	Mean operations	Mean reliability score
1	24,483	5.27	22.27	65.33	89.66
2	20,485	8.67	34.93	189.62	57.40
3	2,532	8.03	83.83	202.18	61.41
4	2,500	8.16	28.56	191.45	73.54

Table 5. Reliability Archetypes

Table 5 summarizes four circuit archetypes identified using MiniBatch K-means clustering on standardized structural and gate-count features. The clusters differ in their number of circuits, average qubit count, circuit depth, operation count, and mean reliability score.

Archetype 1 is the largest cluster, containing 24,483 circuits, with mean values of 5.27 qubits, 22.27 depth units, and 65.33 operations. It has the highest mean reliability score of 89.66. This profile is consistent with comparatively small and shallow circuits exhibiting higher simulated reliability.

Archetype 2 contains 20,485 circuits and has mean values of 8.67 qubits, 34.93 depth units, and 189.62 operations, with a mean reliability score of 57.40. Its profile indicates a comparatively complex circuit group associated with lower mean simulated reliability.

Archetype 3 contains 2,532 circuits, with mean values of 8.03 qubits, 83.83 depth units, and 202.18 operations. Its mean reliability score is 61.41. This archetype exhibits the greatest average depth and operation count, suggesting that accumulated circuit complexity may be relevant to reliability variation.

Archetype 4 contains 2,500 circuits, with mean values of 8.16 qubits, 28.56 depth units, and 191.45 operations. Its mean reliability score is 73.54. Despite its relatively high operation count, this cluster has a higher mean reliability score than Archetypes 2 and 3, indicating that operation count alone may not sufficiently characterize circuit reliability.

Overall, the cluster profiles support a multidimensional interpretation of circuit complexity involving qubit count, depth, operation count, and gate composition. These archetypes should be treated as exploratory data-driven groupings rather than established physical categories of quantum circuit reliability.

5.4 Regression Model Comparison

model	validation_RMSE	validation_MAE	validation_R2	test_RMSE	test_MAE	test_R2
HistGradientBoosting	6.152365	4.162527	0.958718	4.857998	2.847622	0.94476
RandomForest	6.978215	4.678995	0.946891	5.128837	2.925869	0.938429
ExtraTrees	6.666249	4.602447	0.951533	5.22596	2.99154	0.936075

Table 6. Regression Model Comparison Across Validation and Test Sets

Table 6 extends the regression analysis by presenting the validation and test metrics for all three models. *HistGradientBoosting* achieved a validation RMSE of 6.152, MAE of 4.163, and \(R^2\) of 0.959. Its test RMSE decreased to 4.858, MAE to 2.848, and \(R^2\) to 0.945.

Random Forest recorded a validation RMSE of 6.978, MAE of 4.679, and \(R^2\) of 0.947, compared with a test RMSE of 5.129, MAE of 2.926, and \(R^2\) of 0.938. Extra Trees achieved a validation RMSE of 6.666, MAE of 4.602, and \(R^2\) of 0.952, while its test values were 5.226, 2.992, and 0.936, respectively.

Hist Gradient Boosting consistently ranks first on the reported test metrics. The results demonstrate that the model captures useful predictive patterns in the structural features, even when validation groups are separated according to qubit count and circuit depth quintile. The differences between validation and test errors, however, should be investigated in conjunction with the group composition and score distributions of each partition. Grouped validation reduces some forms of structural overlap but does not replace external validation using independent circuit families or hardware measurements.

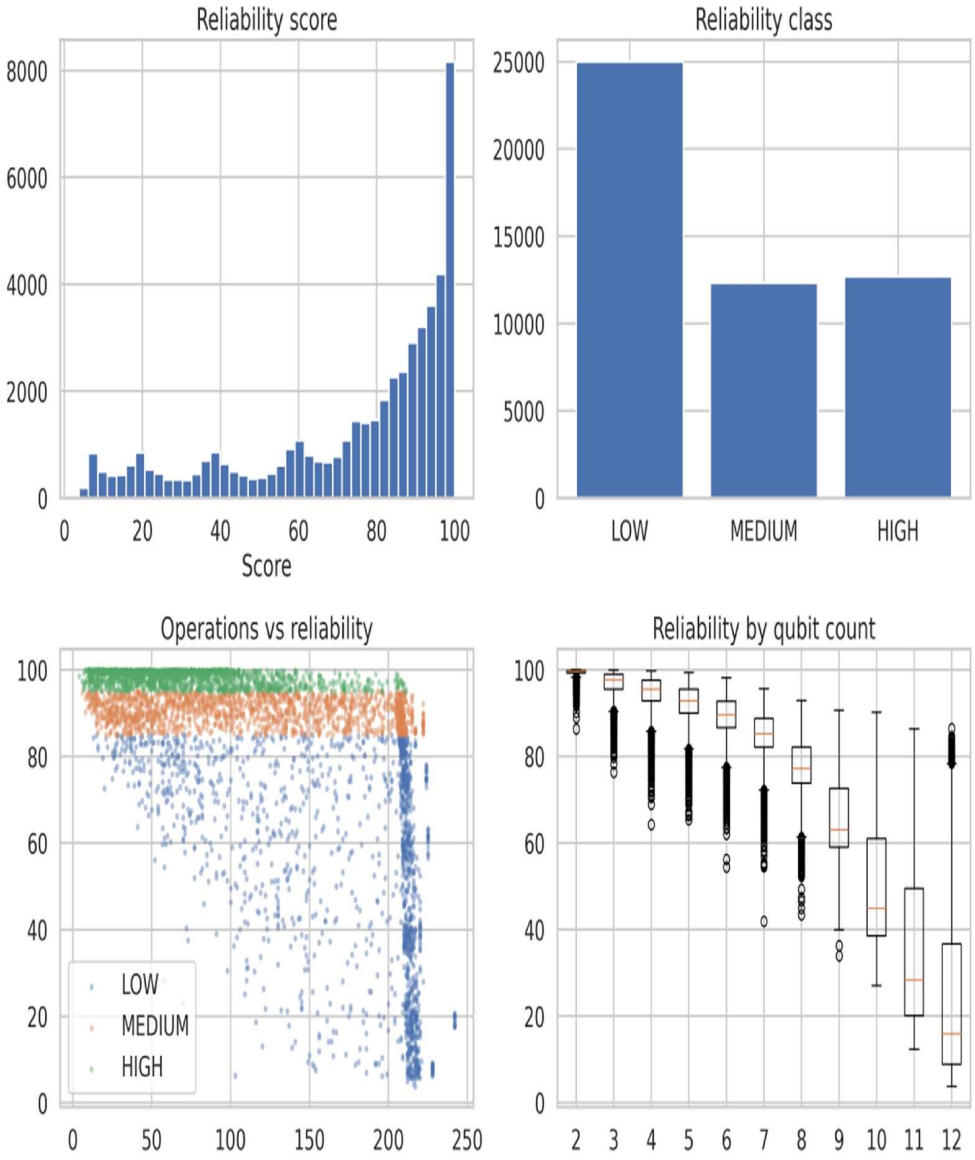

Figure 2. Reliability Distribution and Circuit Characteristics

Figure 2 presents the distribution of reliability scores, the frequency of reliability classes, the relationship between circuit operations and reliability, and the variation in reliability across qubit counts.

The reliability score histogram shows a concentration of observations toward the higher end of the score range, indicating that many simulated circuits have relatively high reliability scores. The distribution also contains lower scoring circuits, suggesting substantial variation in predicted or simulated reliability across the dataset.

The reliability class bar chart shows the frequency of the LOW, MEDIUM, and HIGH categories. LOW is the largest category, while MEDIUM and HIGH contain fewer observations. This class distribution is relevant when interpreting the classification models because unequal class frequencies can affect accuracy and other aggregate performance measures.

The operations versus reliability scatter plot illustrates how reliability varies across circuits with different operation counts. Although many circuits retain relatively high scores, the spread of scores becomes more pronounced among circuits with larger operation counts. This suggests that operational complexity may be associated with reliability variation, but the relationship is not deterministic.

Finally, the boxplots of reliability by qubit count show a general decline in the central reliability distribution and an increase in variability as the number of qubits rises. This pattern is consistent with the possibility that larger circuits are more vulnerable to accumulated noise under the simulation assumptions.

Taken together, the panels indicate that reliability is heterogeneous and that circuit size and operational complexity are potentially informative predictors. These are descriptive associations, not evidence of causal effects or confirmed physical-hardware behaviour.

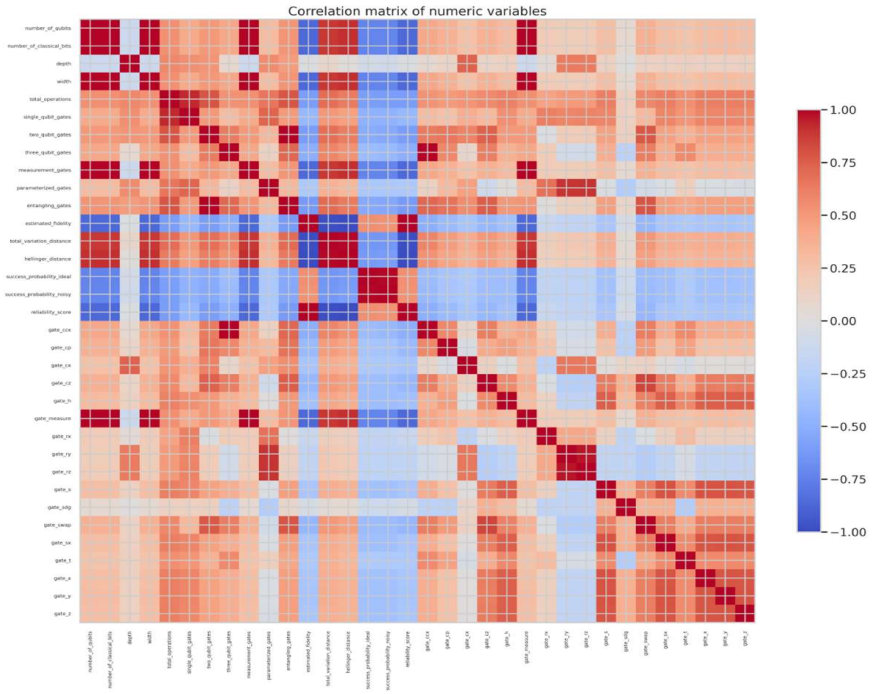

Figure 3. Correlation Matrix

Figure 3 displays the correlation matrix for the numerical variables describing the quantum circuits and their reliability related characteristics. The colour scale represents the direction and strength of the pairwise correlations, with positive correlations shown in warm colours and negative correlations shown in cool colours.

The matrix contains groups of strongly correlated variables, particularly among circuit-size measurements, gate counts, and related structural features. Such patterns are expected when several variables capture overlapping aspects of circuit complexity. For example, total operations may increase alongside particular gate counts, while qubit count and circuit depth may share information about the overall scale of a circuit.

The presence of strong correlations has important implications for predictive modelling. Highly correlated features can introduce redundancy, complicate the interpretation of feature importance, and distribute importance across related variables. Consequently, the apparent contribution of an individual gate count should not necessarily be interpreted independently of the other structural features.

The correlation matrix is also useful for identifying potential multicollinearity and informing feature selection decisions. However, correlation alone cannot establish causation or determine whether a feature provides

unique predictive information after accounting for other variables.

5.5 Regression evaluation

Grouped split: training 34,846; validation 6,061; test 9,093. Groups were defined by qubit count × depth quintile, preventing overlap of these group labels across partitions. Structural predictors only were used.

Model	Validation _RMSE	Validation _MAE	Validation _R2	test _RMSE	test _MAE	test _R2
HistGradient Boosting	6.1524	4.1625	0.9587	4.8580	2.8476	0.9448
RandomForest	6.9782	4.6790	0.9469	5.1288	2.9259	0.9384
ExtraTrees	6.6662	4.6024	0.9515	5.2260	2.9915	0.9361

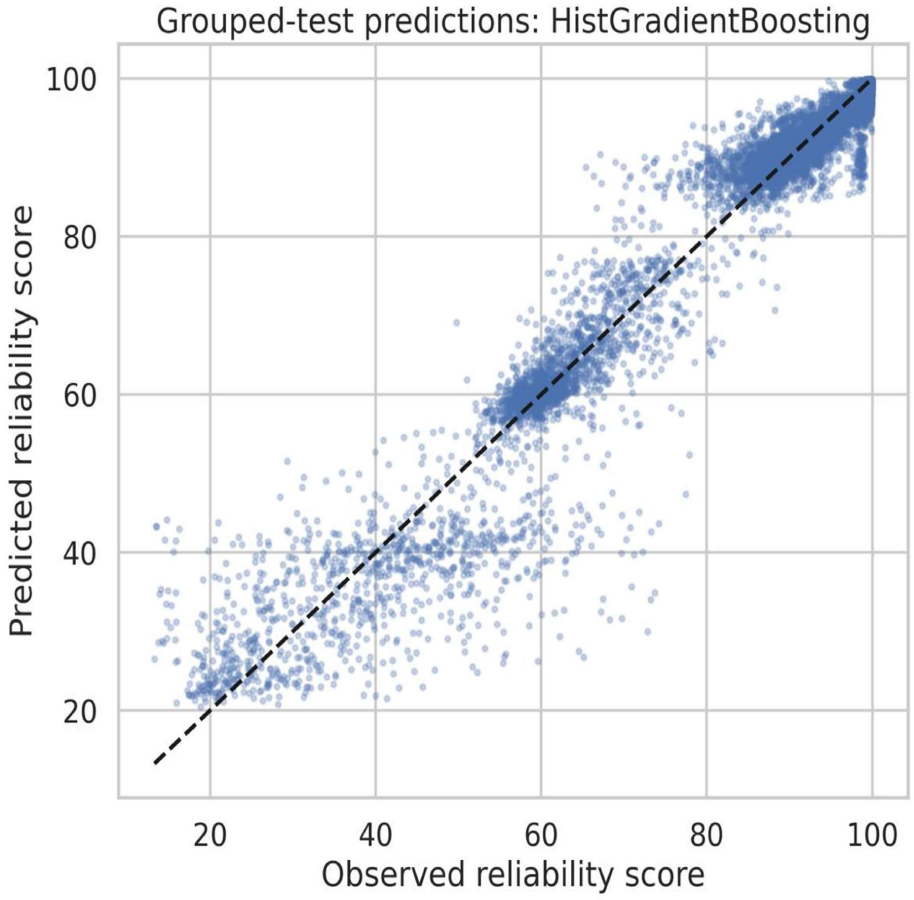

Figure 4. Grouped-text predictions

5.6 Classification Evaluation

Models were compared on the same grouped split using accuracy, balanced accuracy, macro-F1, weighted-F1, precision and recall. Full metrics and class-wise results are in the accompanying CSV files.

Model	Validation_Accuracy	Validation macro_F1	Validation_Weighted_F1	Test_Accuracy	Test_Balanced_Accuracy	Test_Macro_F1	test_weighted_F1	test_precision_macro	test_recall_macro
Hist Gradient Boosting	0.9160	0.7832	0.9147	0.8886	0.8877	0.8879	0.8890	0.8883	0.8877
Random Forest	0.9020	0.7283	0.8974	0.8774	0.8746	0.8761	0.8775	0.8811	0.8746
Extra Trees	0.9053	0.7426	0.9025	0.8590	0.8557	0.8576	0.8594	0.8643	0.8557

Table 7. Classification performance

Table 7 consolidates classification performance across validation and test partitions. HistGradientBoosting achieved validation accuracy of 0.916 and weighted-F1 of 0.915, compared with test accuracy of approximately 0.889 and weighted-F1 of 0.889. Random Forest achieved test accuracy of approximately 0.877, while Extra Trees recorded approximately 0.859.

The table reinforces the overall ranking observed in the earlier classification comparisons, with HistGradientBoosting providing the strongest reported test-set performance. However, the validation macro-F1 values of 0.783 for HistGradientBoosting, 0.728 for Random Forest, and 0.743 for Extra Trees are notably lower than the corresponding test macro-F1 values. This pattern suggests that class-specific performance differs across partitions and requires closer examination.

For a complete journal presentation, the aggregate metrics should be supplemented with per-class precision, recall, F1 scores, support counts, and the confusion matrix. Such information would identify which reliability classes are most frequently confused and whether the overall performance conceals weaknesses in predicting particular categories.

The document labels the grouped-test confusion matrix as Figure 3. The following description covers both the regression-prediction panel and the classification confusion matrix because they jointly illustrate predictive performance.

Regression: observed versus predicted reliability. The scatter plot compares the observed reliability scores with the values predicted by *HistGradientBoosting* on the grouped test set. The diagonal reference line represents perfect agreement between observed and predicted values.

Most points are distributed around the diagonal, indicating that the model captures the overall relationship between structural features and the reliability score. The clustering of observations near the reference line is consistent with the relatively high test \(R^2\) of 0.945 reported in the regression results.

Nevertheless, some points deviate considerably from the diagonal, particularly in portions of the intermediate score range. These deviations indicate that the model does not predict every circuit with equal precision. Further examination of residuals across reliability levels and circuit-complexity groups would help identify

systematic underprediction or overprediction.

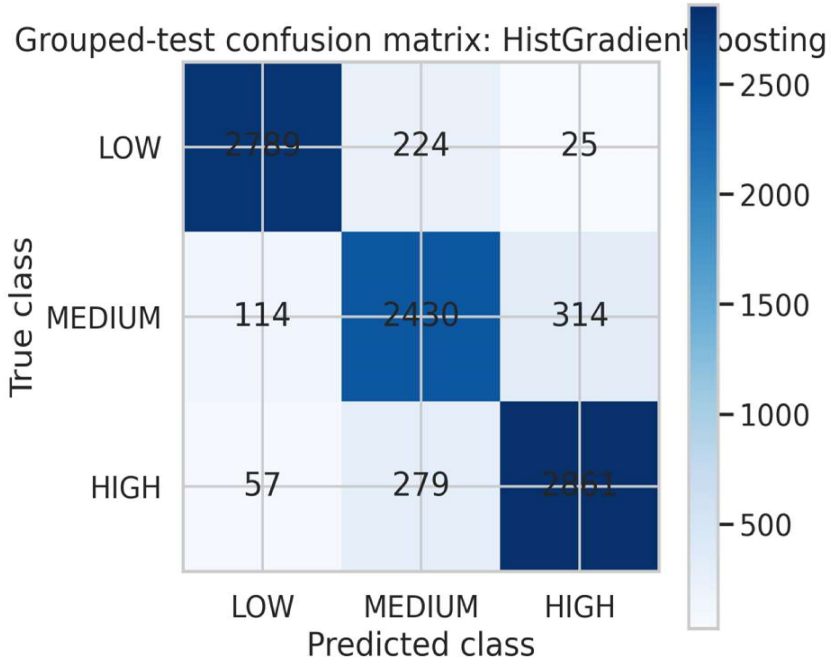

Figure 5. Grouped Text confusion matrix

5.7 Explainability

Permutation importance was evaluated on grouped validation data. SHAP global importance and summary plots were computed with an ExtraTrees model trained on a representative subset of training circuits and explained on held-out validation samples. Partial-dependence curves show average model response, not causal effects. Correlated gate counts can divide or distort importance, so interpret rankings as model-specific evidence.

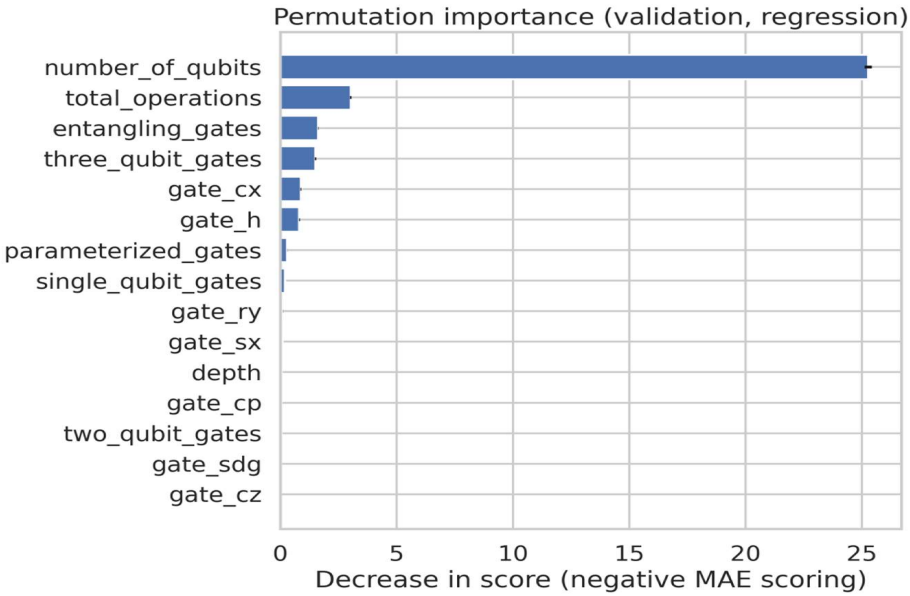

Figure 6. Permutation importance

Figure 6 in the document, labelled Figure, presents permutation feature importance for the regression model on grouped validation data. The horizontal axis represents the decrease in model performance, expressed as an increase in MAE, when a feature is permuted. Larger decreases in predictive performance indicate greater model reliance on the corresponding feature.

The number of qubits is the dominant feature in the displayed ranking, with a substantially greater importance score than the remaining variables. Total operations, entangling gates, three-qubit gates, and several individual gate counts have much smaller contributions. Many other features have importance values close to zero in this evaluation.

This pattern suggests that the model relies heavily on qubit count when predicting reliability, while the remaining features provide smaller incremental predictive contributions under the fitted model and evaluation procedure. The prominence of qubit count is consistent with the descriptive pattern showing declining reliability at higher qubit counts.

However, permutation importance measures predictive dependence, not the causal influence of a feature. Correlated features may substitute for one another, resulting in underestimated individual importance scores. The apparent dominance of qubit count should therefore be examined using grouped feature ablation, conditional permutation methods, or additional validation across independent circuit families before drawing strong conclusions about the underlying physical mechanisms.

5.8 Reliability archetypes

Four archetypes were generated with MiniBatch K-means over standardized structural and gate-count features. Cluster numbers are nominal labels. See the cluster profile and circuit assignment CSV files.

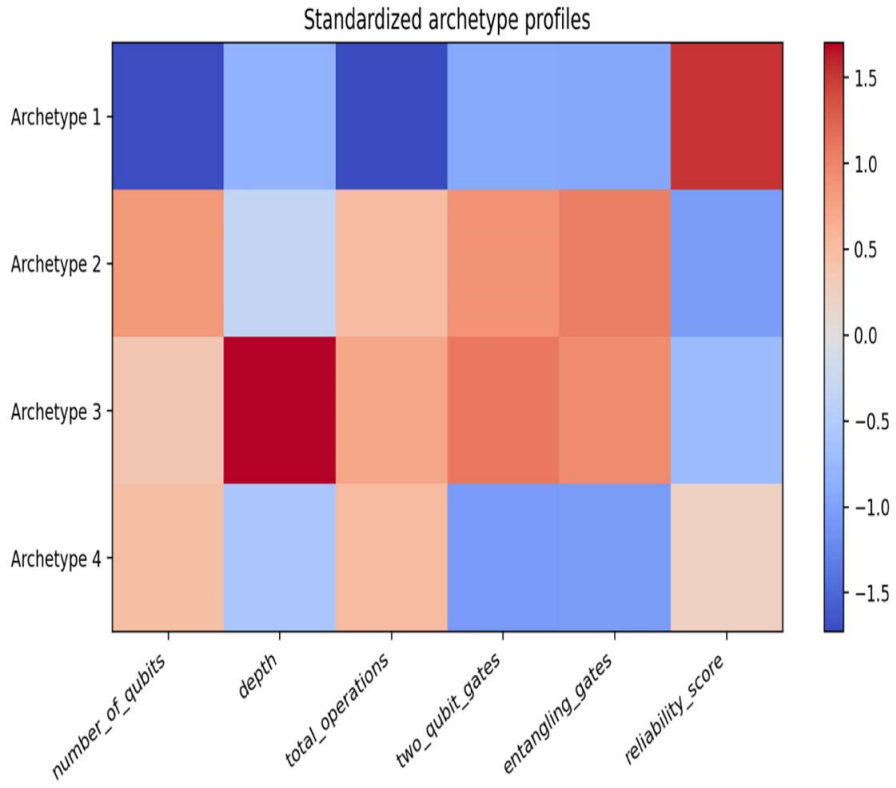

Figure 7. Standardized Reliability Archetype Profiles

Figure 7 in the document, labelled Figure, displays a heatmap of standardized mean feature values for the four reliability archetypes identified using MiniBatch K-means.

Each row represents an archetype, while each column represents a selected structural or reliability-related characteristic. The colour scale indicates whether the cluster mean is above or below the overall standardized mean: positive values indicate relatively high values, while negative values indicate relatively low values. Archetype 1 exhibits comparatively low values for several complexity-related characteristics and a high standardized reliability score. This profile is consistent with its higher mean reliability score and relatively smaller, shallower circuits.

Archetype 2 displays elevated values across several structural and gate-related characteristics, alongside a lower standardized reliability score. This suggests that the cluster captures circuits with comparatively complex structures and lower simulated reliability.

Archetype 3 exhibits a particularly high standardized value for gate count, as well as elevated values for several other structural variables. Its profile is consistent with the high mean depth and operation count reported in the cluster summary.

Archetype 4 displays a different combination of structural characteristics and an intermediate to high reliability profile relative to Archetypes 2 and 3. This supports the observation that circuits with similar operation counts may differ in reliability because of differences in their structural composition.

The heatmap demonstrates that the four clusters are distinguished by combinations of characteristics rather than by a single feature. It therefore provides a useful qualitative representation of circuit archetypes and supports the interpretation of clustering results beyond cluster size alone.

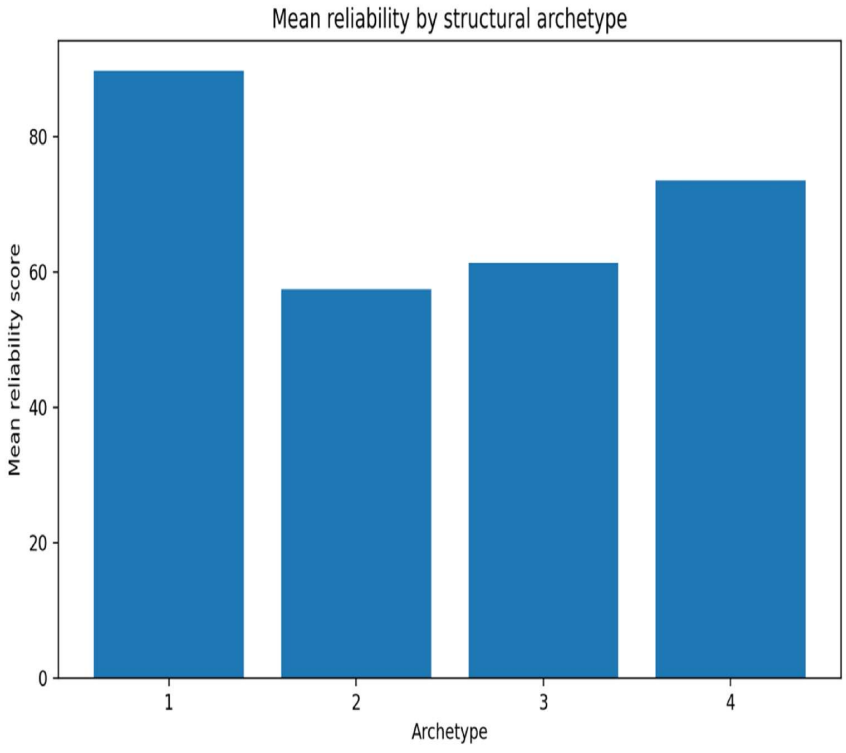

Figure 8. Mean Reliability by Structural Archetype

Figure 8 in the document, labelled Figure, compares the mean reliability score of the four identified circuit archetypes.

Archetype 1 has the highest mean reliability score, approximately 89.66, followed by Archetype 4 at 73.54. Archetype 3 has a mean score of 61.41, while Archetype 2 has the lowest mean score at 57.40.

These differences indicate substantial variation in simulated reliability across the data-driven clusters. The higher mean score of Archetype 1 is consistent with its comparatively small and shallow circuits. In contrast, Archetypes 2 and 3 have lower mean scores and more complex structural profiles.

The comparison also highlights why operation count alone may be insufficient to explain reliability. Archetype 4 contains a high mean number of operations but achieves a higher mean reliability score than Archetypes 2 and 3. The combination of depth, qubit count, gate composition, and other circuit properties may therefore be more informative than total operation count in isolation.

The figure supports the use of archetype-based analysis to summarize circuit heterogeneity. However, the observed differences are descriptive; statistical tests and uncertainty intervals would be required to assess their precision and significance. The clusters should not be treated as predefined physical reliability classes.

5.9 Additional explainability

5.9.1 SHAP Global Importance

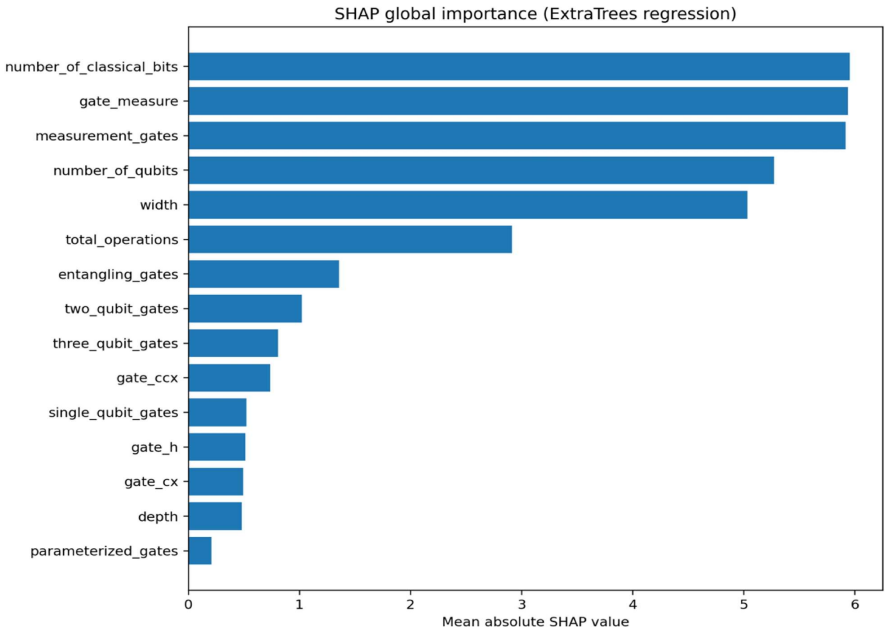

Figure 9. SHAP Global Feature Importance

Figure 9 in the document, labelled Figure, presents global SHAP feature importance for the Extra Trees regression model. The horizontal axis represents mean absolute SHAP values, which quantify the average magnitude of each feature's contribution to the model output across the explained observations.

The ranking identifies *number_of_classical_bits, gate_measure, and measurement_gates* among the most influential features, followed by *number_of_qubits* and width. Total operations also show a substantial contribution, while entangling gates and selected gate-count variables have comparatively smaller importance values.

These results indicate that the Extra Trees model relies on several circuit-size, measurement related, and operational features when generating reliability predictions. The ranking complements the permutation-importance analysis by providing a different measure of model reliance.

The difference between the two rankings is important. Permutation importance emphasizes the change in predictive performance after a feature is disrupted, whereas mean absolute SHAP importance measures the magnitude of feature contributions to individual predictions. Consequently, a feature may rank highly under one method but less prominently under the other.

The prominence of measurement-related and classical bit variables warrants particular scrutiny. It is necessary to determine whether these variables capture genuine circuit characteristics associated with the simulated reliability target or act as proxies for the data generation process. Because the dataset is synthetic, feature importance should be interpreted as model specific evidence rather than direct evidence of physical causation.

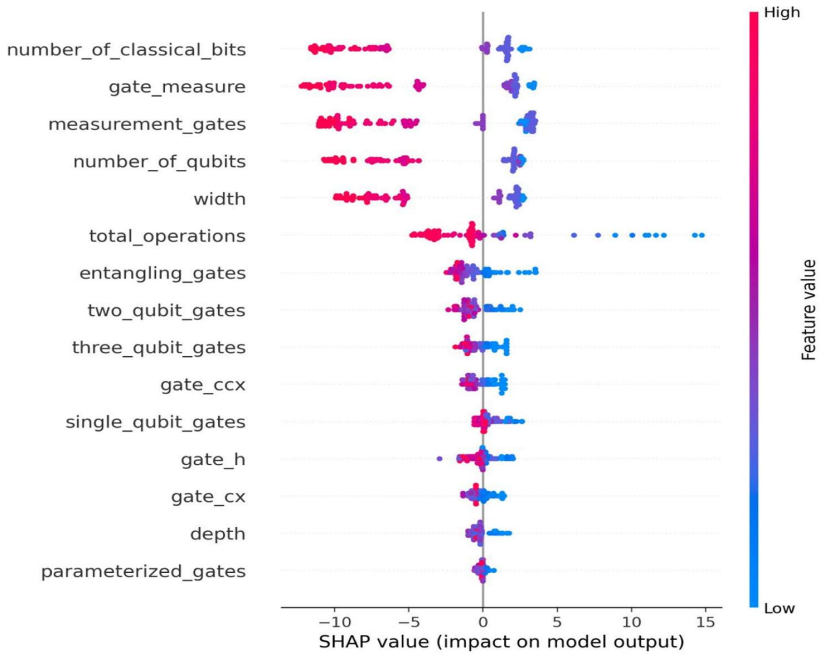

Figure 10. SHAP Summary Plot

Figure 10 in the document, labelled Figure, presents a SHAP summary plot showing the distribution and direction of feature contributions across the observations explained by the Extra Trees model.

Each point represents an observation for a particular feature. The horizontal position indicates the SHAP value: positive values push the model prediction upward, whereas negative values push it downward. The point colour represents the relative feature value, with higher and lower values indicated by contrasting colours.

The plot shows that the contributions of several measurement-related and circuit-size variables vary across observations. Features such as *number_of_classical_bits, gate_measure,* and *measurement_gates* exhibit notable variation in SHAP values, indicating that their contributions are not identical across all circuits.
The spread of points also demonstrates that the effect of a feature depends on the characteristics of the individual circuit. Some features show contributions concentrated near zero, while others exhibit broader distributions, suggesting more substantial variation in their model contributions.

The SHAP summary plot therefore complements the global importance ranking by showing both the magnitude and direction of feature contributions. However, the figure does not establish that increasing a particular

feature necessarily improves or reduces physical reliability. Correlations among structural variables, interactions captured by the model, and the synthetic target-generation process may all influence the observed patterns.

5.10 Partial Dependence

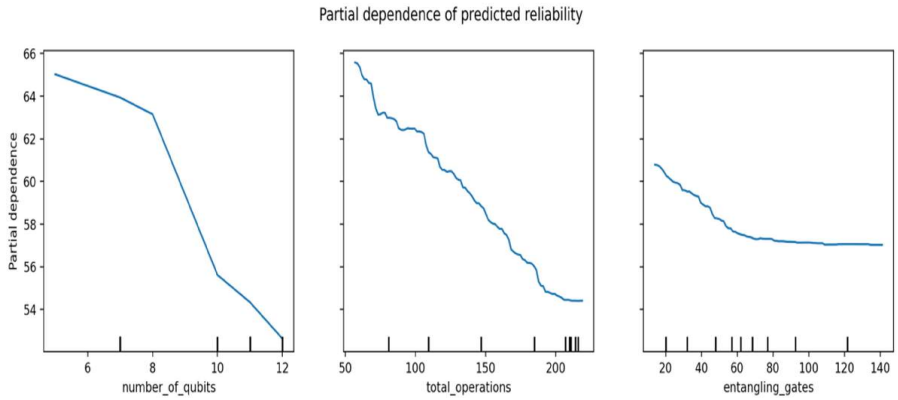

Figure 11. Partial Dependence Plots

Figure 11 in the document, labelled Figure, presents partial dependence plots for three structural features: *number_of_qubits, total_operations,* and *entangling_gates*. These plots show the average model predicted reliability as each feature varies, with other features averaged over the reference dataset.

The partial dependence curve for the number of qubits shows a generally declining predicted reliability as qubit count increases. The decline becomes more pronounced across the higher part of the displayed range. This pattern is consistent with the observed association between larger circuits and lower simulated reliability. The total-operations plot also exhibits a broadly decreasing trend. Predicted reliability declines as the number of operations increases, although the curve contains local variations. This suggests that circuits with more operations tend to receive lower predicted reliability scores under the fitted model, on average.

The entangling gates curve shows a decline in predicted reliability over the lower portion of the displayed range, followed by a comparatively flatter response at higher values. This pattern indicates that the marginal relationship between entangling gate count and predicted reliability is not necessarily constant across the feature range.

Collectively, the plots support the interpretation that circuit size and complexity-related features are associated with variation in the model's reliability predictions. Nevertheless, partial dependence represents an average model response rather than a causal effect. If circuit features are correlated, the plots may evaluate combinations of characteristics that are uncommon or absent in the observed data. Their interpretation should therefore be supported by feature distribution checks and, where appropriate, individual conditional expectation or accumulated local effects analyses.

6. Overall Interpretation

The tables and figures collectively demonstrate a structured approach to quantum circuit reliability analysis, combining data quality checks, descriptive exploration, regression, classification, clustering, and explainable AI.

The principal findings are:

1. Data integrity: The dataset is complete and contains no duplicated observations according to the reported

checks, but its synthetic origin limits external validity.

1. Regression performance: *HistGradientBoosting* achieved the strongest reported test performance, with \(R^2 \approx 0.945\).

2. Classification performance: HistGradientBoosting achieved approximately 88.86% test accuracy, although class specific performance and category definitions require further documentation.

3. Circuit archetypes: Four clusters exhibit distinct combinations of structural complexity and mean reliability, supporting multidimensional rather than single-feature characterization.

4. Explainability: Permutation importance, SHAP, and partial dependence identify potentially informative structural and measurement-related variables, but their rankings and response patterns remain model-dependent.

5. Validation: Grouped splitting by qubit count and circuit-depth quintile is more demanding than an ordinary random split for the specified grouping scheme, but it does not establish generalization to unseen circuit families or physical quantum devices.

7. Final Discussion

The analysis of the QUASAR Quantum Reliability Dataset v1 demonstrates a robust, multi-tiered workflow for evaluating quantum circuit reliability using machine learning (ML) and explainable AI (XAI). By integrating data quality assurance, predictive modeling, unsupervised clustering, and model-agnostic interpretability techniques, the framework provides a structured approach to understanding how structural circuit characteristics influence simulated reliability. However, translating these findings into practical, hardware-agnostic quantum compilation strategies requires careful consideration of the dataset's synthetic nature and inherent methodological limitations.

7.1 Summary of Key Findings

• Predictive Performance: Tree-based ensemble models, particularly HistGradientBoosting, proved highly effective at capturing the nonlinear relationships between circuit structure and reliability. It achieved an R^2 of ~0.945 in regression and ~88.86% accuracy in classification on a rigorous grouped test set.

• Multidimensional Complexity (Archetypes): The MiniBatch K-means clustering revealed that circuit reliability cannot be explained by a single metric (like total operation count). Instead, it is driven by a multidimensional profile of qubit count, circuit depth, and gate composition. For instance, Archetype 4 demonstrated that circuits with high operation counts can still maintain relatively high reliability if their depth and qubit interactions are structured differently than those in Archetypes 2 and 3.

• Drivers of Reliability (XAI): Explainability methods highlighted qubit count and circuit depth as primary drivers of reliability degradation, aligning with the physical intuition of accumulated decoherence. Interestingly, SHAP analysis also highlighted the prominence of measurement-related features and classical bits, suggesting that the act of measurement and classical readout structures heavily influence the simulated fidelity scores.

7.2 Limitations

Despite the strong in-distribution performance of the models, several critical limitations must be addressed before these findings can be generalized:

• Synthetic Data and External Validity: The QUASAR v1 dataset is simulation-derived. Therefore, the models are learning the noise models and simulation assumptions of the data generator, not necessarily the true physical behavior of quantum hardware. The grouped validation (by qubit count and depth quintile) prevents simple data leakage but does not guarantee generalization to unseen circuit families or physical devices.

• Target Determinism and Data Leakage: A critical caveat is that the reliability_score is a deterministic rescaling of estimated_fidelity (Score = 100 x fidelity). If estimated_fidelity or highly correlated proxy variables were inadvertently included in the feature space, the models may simply be learning a deterministic mapping rather than discovering genuine physical heuristics.

• Multicollinearity in XAI: The correlation matrix revealed strong multicollinearity among gate counts and structural features. This causes XAI methods to yield divergent results; for example, Permutation Importance heavily favored *number_of_qubits*, while SHAP global importance elevated measurement and classical bit features. This indicates that feature importance is highly model-dependent and fragmented across correlated variables.

• Classification Discrepancies: The classification task is merely a discretized version of the regression target. Furthermore, a notable discrepancy exists between the validation and test macro-F1 scores for *HistGradientBoosting* (0.783 vs. 0.888). This suggests that class-specific boundaries (the thresholds defining LOW, MEDIUM, and HIGH) interact unpredictably with the grouped data splits, requiring deeper per-class confusion matrix analysis.

8. Conclusion

This study successfully establishes a robust, multi-tiered analytical framework for the pre-execution prediction and interpretation of quantum circuit reliability. By integrating rigorous data curation, advanced tree based ensemble modeling, unsupervised archetype discovery, and model agnostic Explainable AI (XAI), we bridge the critical gap between black-box reliability prediction and actionable quantum software engineering.

Our findings demonstrate that *HistGradientBoosting* can accurately forecast circuit reliability ($R^2 \approx 0.945$) using purely structural and gate-count features. Crucially, the identification of four distinct reliability archetypes via MiniBatch K-Means clustering proves that circuit vulnerability to noise is not a monolithic property, but a multidimensional phenomenon governed by the complex interplay of qubit width, circuit depth, and gate composition. Furthermore, XAI techniques successfully isolate the primary structural drivers of decoherence, highlighting the profound impact of qubit count, circuit depth, and measurement operations on simulated fidelity.

However, this work also establishes critical boundaries regarding the interpretation of ML in quantum computing. We demonstrate that feature importance in quantum circuits is highly susceptible to multicollinearity, and that models trained on synthetic data learn simulator-specific behavioural patterns rather than universal physical laws. Therefore, while this framework provides invaluable heuristic insights for noise-aware circuit compilation and optimization, the community must treat these findings as a simulation-validated baseline. Future work must prioritize strict feature auditing to prevent target leakage and, ultimately, external validation on physical NISQ devices to translate these data-driven archetypes into universally applicable quantum error mitigation strategies.

References

[1] Liang, Z., Cheng, J., Ren, H., Wang, H., Hua, F., Ding, Y., Chong, F., Han, S., Shi, Y., Qian, X. (2022). *PAN: Pulse Ansatz on NISQ machines* (arXiv:2208.01215). arXiv. https://arxiv.org/abs/2208.01215.

[2] Peruzzo, A., McClean, J., et al. (2014). A variational eigenvalue solver on a photonic quantum processor. *Nature Communications, 5*(1), 1–7.

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

[4] Cheng, J., Deng, H., Qia, X. (2020). Acc QOC: Accelerating quantum optimal control based pulse generation. In *2020 ACM/IEEE 47th Annual International Symposium on Computer Architecture (ISCA)* (p. 543–555). IEEE.

[5] Hu, Z., Dong, P., Wang, Z., Lin, Y., Wang, Y., Jiang, W. (2022). Quantum neural network compression. In *2022 IEEE/ACM International Conference on Computer Aided Design (ICCAD)*. IEEE.

[6] Jiang, W., Xiong, J., Shi, Y. (2021). A co-design framework of neural networks and quantum circuits towards quantum advantage. *Nature Communications, 12*(1), 1–13.

[7] Qian, Han, S., Jiang, W., Shi, Y. (2022). Variational quantum pulse learning. In *2022 IEEE International Conference on Quantum Computing and Engineering (QCE)*. IEEE.

[8] Liang, Z., Wang, Z., Yang, J., Yang, L., Shi, Y., Jiang, W. (2021). Can noise on qubits be learned in quantum neural network A case study on QuantumFlow. In *2021 IEEE/ACM International Conference on Computer Aided Design (ICCAD)* (p. 1–7). IEEE.

[9] Rebentrost, P., Mohseni, M., Lloyd, S. (2014). Quantum support vector machine for big data classification. *Physical Review Letters, 113*(13), 130503.

[10] Das, P., Pattison, C. A., Manne, S., Carmean, D. M., Svore, K. M., Qureshi, M., Delfosse, N. (2022). AFS: Accurate, fast, and scalable error decoding for fault tolerant quantum computers. In *2022 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (p. 259–273). IEEE.

[11] Hua, F., Chen, Y., Jin, Y., Zhang, C., Hayes, A., Zhang, Y., Zhang, E. Z. (2021). AutoBraid: A framework for enabling efficient surface code communication in quantum computing. In *MICRO-54: 54th Annual IEEE/ACM International Symposium on Microarchitecture* (p. 925–936). ACM.

[12] Krinner, S., Lacroix, N., Remm, A., Di Paolo, A., Genois, E., Leroux, C., Hellings, C., Lazar, S., Swiadek, F., Herrmann, J., et al. (2022). Realizing repeated quantum error correction in a distance three surface code. *Nature, 605*(7911), 669–674.

[13] Liang, Z., Wang, Z., Yang, J., Yang, L., Shi, Y., Jiang, W. (2021). Can noise on qubits be learned in quantum neural network A case study on Quantum Flow. In *2021 IEEE/ACM International Conference on Computer Aided Design (ICCAD)* (p. 1–7). IEEE.

[14] Ravi, G. S., Smith, K. N., Gokhale, P., Mari, A., Earnest, N., Javadi Abhari, A., Chong, F. T. (2022). VAQEM: A variational approach to quantum error mitigation. In *2022 IEEE International Symposium on High Performance Computer Architecture (HPCA)* (p. 288–303). IEEE.

[15] Wang, H., Gu, J., Ding, Y., Li, Z., Chong, F. T., Pan, D. Z., Han, S. (2022). QuantumNAT: Quantum noise-aware training with noise injection, quantization and normalization. In *Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC)*. ACM.

[16] Liang, Z., Wang, Z., Yang, J., Yang, L., Shi, Y., Jiang, W. (2021). Can noise on qubits be learned in quantum neural network A case study on QuantumFlow. In *2021 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (p. 1–7). IEEE.

[17] Gilyén, A., Poremba, A. (2022). *Improved quantum algorithms for fidelity estimation* (arXiv:2203.15993). arXiv. https://arxiv.org/abs/2203.15993.

[18] Wang, Q., Zhang, Z., Chen, K., Guan, J., Fang, W., Ying, M. (2021). *Quantum algorithm for fidelity estimation* (arXiv:2103.09076). arXiv. https://arxiv.org/abs/2103.09076.

[19] Cerezo, M., Poremba, A., Cincio, L., Coles, P. J. (2020). Variational quantum fidelity estimation. *Quantum, 4*, 248.

[20] Chen, R., Song, Z., Zhao, X., Wang, X. (2021). Variational quantum algorithms for trace distance and fidelity estimation. *Quantum Science and Technology, 7*(1), 015019.

[21] Tan, K. C., Volkoff, T. (2021). Variational quantum algorithms to estimate rank, quantum entropies, fidelity, and Fisher information via purity minimization. *Physical Review Research, 3*(3), 033251.

[22] Liu, J., Zhou, H. (2020). Reliability modeling of NISQ-era quantum computers. In *2020 IEEE International Symposium on Workload Characterization (IISWC)* (p. 94–105). IEEE.

[23] Wang, H., Liu, P., Cheng, J., Liang, Z., Gu, J., Li, Z., Han, S. (2022). *Quest: Graph transformer for quantum circuit reliability estimation* (arXiv:2210.16724). arXiv. https://arxiv.org/abs/2210.16724.

[24] Saravanan, V., Saeed, S. M. (2021). Test data-driven machine learning models for reliable quantum circuit output. In *2021 IEEE European Test Symposium (ETS)* (p. 1–6). IEEE. https://doi.org/10.1109/ETS50041.2021.9465405.

[25] Saravanan, V., Saeed, S. M. (2023). Data-driven reliability models of quantum circuit: From traditional ML to graph neural network. *IEEE Transactions on Computer Aided Design of Integrated Circuits and Systems, 42*(5), 1477–1489. https://doi.org/10.1109/TCAD.2022.3202430.

[26] Saghafi, M., Mili, L. (2025). Predictive time series analysis of single qubit quantum circuit outcomes for a superconducting quantum computer: Forecasting error patterns. *IEEE Access, 13*, 40115–40132. https://doi.org/10.1109/ACCESS.2025.3546259.