computation

MDPI

Article

Quantum vs. Classical Machine Learning for Diabetes Prediction: A Leakage-Controlled Benchmarking Study with Repeated Evaluation

Zia Ud Din [1], Hufsa Mohsin [1], Ahmad Sami Al-Shamayleh [2], Ghulam Mustafa [1,*], Saif Ur Rehman Khan [1], Naveed Ahmad [1], Isabel De la Torre Díez [3], Roger Sarín Rivas Suárez [4,5,6,7] and Leonardo Hernández Cruz [6,7,8,9]

1 Faculty of Computing, Shifa Tameer-e-Millat University, Islamabad 44000, Pakistan; zia20004@gmail.com (Z.U.D.); hufsa.ssc@stmu.edu.pk (H.M.); saif_rehman.ssc@stmu.edu.pk (S.U.R.K.); naveed_ahmad.ssc@stmu.edu.pk (N.A.)
2 Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19111, Jordan; a.alshamayleh@ammanu.edu.jo
3 eHealth and Telemedicine Group, University of Valladolid, 47011 Valladolid, Spain; isator@uva.es
4 Fundación Universitaria Internacional de Colombia, Bogotá 111321, Colombia; roger.rivas@unic.co.ao
5 Departamento de Ciencias de la Computación, Universidad de La Romana, La Romana 22000, Dominican Republic
6 Departamento de Soluciones Tecnológicas y Sistemas, Escuela Politécnica Superior, Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain; leonardo.hernandez@unic.co.ao
7 Department of Computer Science, Universidad Internacional Iberoamericana, Arecibo 00613, Puerto Rico
8 Departamento de Ciências da Computação, Universidade Internacional do Cuanza, Estrada Nacional 250, Bairro Kaluapanda, Cuito, Bié, Angola
9 Departamento de Ciencias de la Computación, Universidad Internacional Iberoamericana, Campeche 24560, Mexico
* Correspondence: ghulam.mustafa.ssc@stmu.edu.pk

Abstract

This study re-evaluates classical and simulated quantum machine-learning approaches for diabetes prediction under a leakage-controlled and reproducible protocol. The original experimental code was audited in response to peer review, revealing that several preprocessing operations had previously been fitted before the train/test split and that some reported metrics were not computed consistently across models. The revised protocol therefore performs stratified splitting before preprocessing; fits outlier handling, scaling, and PCA only on the training partition; applies SMOTE only to training data; evaluates five independent random seeds; and reports positive-class precision, recall, F1, balanced accuracy, ROC-AUC, PR-AUC, MCC, and 95% confidence intervals. Tuned classical baselines (Logistic Regression, SVC, KNN, Random Forest, and Gradient Boosting) were compared with 8-qubit fidelity-kernel QSVC and QKNN models on a common 2000-record BRFSS subset, while PIMA was evaluated as a second benchmark. Under training-only SMOTE on BRFSS, Logistic Regression achieved mean ROC-AUC 0.811 and positive-class recall 0.748, compared with 0.737 and 0.326 for QSVC; on PIMA, SVC achieved ROC-AUC 0.836 and recall 0.741, compared with 0.761 and 0.607 for QSVC. Thus, the corrected experiments do not support a general quantum predictive advantage. The principal value of the quantum experiments is instead the controlled assessment of representation behavior and computational cost. Measured classical scalability experiments from 10,000 to 55,000 records further show stable ROC-AUC but a strong trade-off between raw accuracy and minority-class sensitivity, whereas exact kernel-memory analysis shows that a dense float64 QSVC kernel would require approximately 23.04 GB at 60,000 records. PCA is therefore interpreted strictly as an unsupervised variance representation, while supervised permutation importance is used separately for predictive attribution. These results provide a more cautious and

Check for updates

Academic Editor: Rafael Lahoz-Beltra

Received: 15 October 2026
Revised: 15 November 2026
Accepted: 17 November 2026
Published: 20 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

reproducible assessment of when simulated quantum kernels are competitive and where their current computational limitations remain substantial.

Keywords: quantum machine learning; quantum support vector classifier; quantum K-nearest neighbor; feature identification; empirical analysis; diabetes

1. Introduction

Diabetes is a chronic metabolic disease that affects millions of people worldwide, characterized by consistently high blood glucose levels resulting from insufficient insulin production or reduced cellular response to insulin [1]. The prevalence of the disease has grown significantly over recent decades, particularly in developing countries, where urbanization and lifestyle changes have contributed to rising obesity rates and declining physical activity [2]. According to current estimates, the number of affected individuals worldwide is projected to rise sharply in the coming decades, with a disproportionate impact expected in low- and middle-income countries where healthcare resources remain limited [1,2], as illustrated in Figure 1.

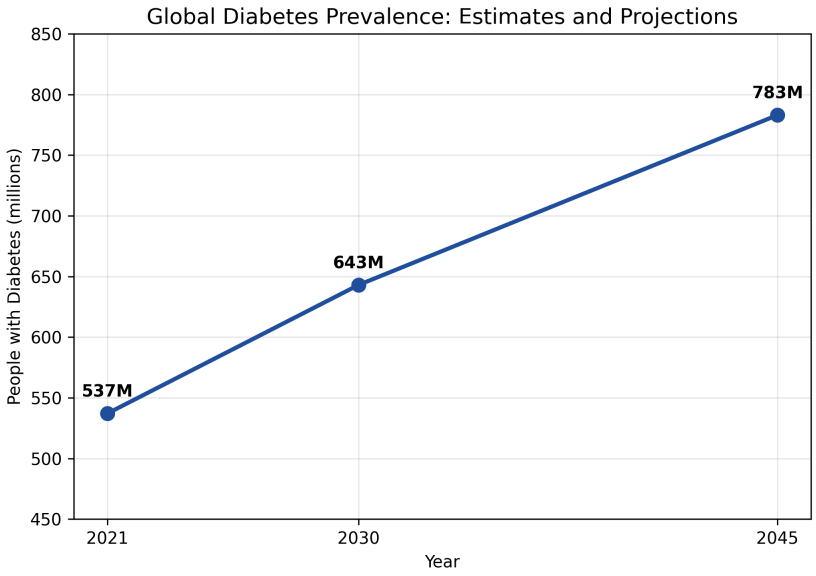

Figure 1. Diabetic patients estimation.

Given this growing burden, the field of diabetes classification and risk prediction has seen increasing interest in the integration of artificial intelligence (AI) and machine learning (ML) approaches to support earlier and more accurate diagnosis [3]. Classical Machine Learning (CML) algorithms, including Support Vector Classifier (SVC), Random Forest (RF), and K-Nearest Neighbor (KNN), have achieved considerable success in identifying patterns within complex medical data that would otherwise go unnoticed [4] and have been used to optimize treatment planning and predict early signs and symptoms of diabetes [5]. However, CML algorithms frequently suffer from the "curse of dimensionality," which increases computational complexity and the likelihood of overfitting [4]. As the number of features grows, CML models often struggle to fully capture the inherent variability present in healthcare data, leading to unpredictable or unreliable outcomes [6]. This limitation restricts their practical utility in increasingly data-rich and evolving clinical environments, motivating the search for alternative computational paradigms [7,8].

Quantum computing has emerged as one such alternative, offering the potential to mitigate overfitting and produce more consistent predictive outcomes. By leveraging the principles of quantum mechanics, quantum computing enables computations that go beyond the capabilities of classical systems [9]. Through the use of quantum bits (qubits), it supports parallelism and superposition, offering significant speed-ups for certain classes of computation and optimization problems. Nonetheless, challenges such as maintaining coherence and implementing effective error correction remain key obstacles to achieving scalable and reliable quantum systems [10]. As classical computing approaches its practical limits in processing increasingly large and complex medical datasets, researchers have turned growing attention to Quantum Machine Learning (QML) as a promising direction. QML incorporates quantum properties such as superposition and entanglement, offering potential advantages over CML in terms of speed and problem-solving capability [11]. Its inherently probabilistic nature is well suited to modeling the uncertainty present in medical data, and it has shown particular promise in handling high-dimensional data more efficiently, thereby addressing the curse of dimensionality that limits classical approaches [12].

This potential has drawn considerable research interest in QML applications for healthcare, particularly in complex disease classification and prediction tasks [13]. Quantum algorithms such as Quantum Support Vector Classifier (QSVC) and Quantum K-Nearest Neighbors (QKNN) are increasingly being investigated for their ability to process high-dimensional data and detect subtle patterns that may remain invisible to classical algorithms [14]. Recent advances in quantum kernel methods have further demonstrated promise in identifying complex molecular patterns associated with diabetes progression [15], particularly in capturing intricate relationships between genetic markers and disease manifestation, with potential advantages reported in selected high-dimensional genomic settings, although such gains are task-, representation-, simulator-, and hardware-dependent and should not be interpreted as a universal quantum advantage.

Building on this motivation, the present research empirically examines whether quantum representations based on superposition and entanglement provide measurable benefits over CML in processing complex medical data while also accounting for current technological constraints and noise-related challenges inherent to present-day quantum systems. This empirical evaluation aims to provide practical insight into the applicability and future potential of both CML and QML approaches in the diagnosis and management of diabetes, ultimately contributing to the broader goal of advancing precision medicine in diabetic care. In pursuit of this aim, the study makes the following contributions:

- **RC1:** An empirical analysis framework to predict diabetic and non-diabetic patients by implementing different CML and QML algorithms.
- **RC2:** An empirical evaluation to assess the limitations and strengths of CML and QML for the prediction of diabetes across different datasets.
- **RC3:** An exploration of the limitations faced by existing CML and QML algorithms in processing and analyzing diabetic data.
- **RC4:** A system capable of recommending suitable features for diabetes prediction, along with the most appropriate CML or QML algorithm, based on various performance parameters.

Taken together, these contributions provide a controlled basis for testing whether simulated QML offers a reproducible predictive advantage over tuned classical baselines while explicitly accounting for class imbalance, preprocessing leakage, statistical uncertainty, and computational cost. The revised study does not assume a quantum advantage a priori; instead, any advantage must be demonstrated under identical partitions and evaluation rules.

The remainder of this paper is organized as follows: Section 2 reviews related work on QML applications in healthcare and diabetes prediction; Section 3 presents the proposed

framework, including data preprocessing, feature mapping, and model design; Section 4 reports experimental results and discussion; and Section 5 concludes the paper and outlines directions for future work.

2. Related Work

The application of Quantum Machine Learning (QML) to healthcare classification problems, particularly diabetes prediction, has attracted growing research interest over the past few years. This section reviews the most relevant studies in this domain, organized thematically according to their methodological focus, before presenting a comparative summary and identifying the gaps that motivate the present work.

2.1. Diabetes-Specific QML Frameworks

Several studies have directly targeted diabetes classification using hybrid quantum-classical pipelines. Hancco-Quispe et al. [16] applied dimensionality reduction techniques, namely, Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA), to the PIMA Diabetes Database in a hybrid approach designed to enhance the performance of the Quantum Support Vector Classifier (QSVC) and the Variational Quantum Classifier (VQC). Their results highlighted that, although QML offers notable advantages over classical methods, the limited availability of qubits on current hardware remains a significant constraint on realizing its full potential. Similarly, Maheshwari et al. [17] compared classical models such as Decision Tree and Random Forest against the quantum-boosting algorithm Qboost for classifying diabetic patients with and without acute conditions, reporting that the optimized quantum model achieved 0.69 across precision, recall, F1 score, specificity, and accuracy, alongside a 55-fold increase in computational speed relative to classical approaches.

Building on this direction, Sierra-Sosa et al. [18] applied a hybrid VQC model to classify Type 2 Diabetes Mellitus with acute comorbidities, using a preprocessing pipeline based on Stokes parameters to improve data mapping and state preparation for Noisy Intermediate-Scale Quantum (NISQ) devices. Using IBM's Qiskit framework, their approach achieved classification accuracies of 0.70 and 0.72 with two and three qubits, respectively, underscoring the promise of VQC for real-world medical data despite current hardware limitations. Kumbargeri et al. [19] instead employed the Quantum Kitchen Sinks (QKS) algorithm, transforming classical input data into high-dimensional features to train a Logistic Regression model on a diabetes dataset; their two-qubit quantum simulator implementation demonstrated improved feature extraction and classification accuracy over conventional baselines. In a closely related comparative study, Vimal [20] evaluated the PIMA diabetes dataset using both classical models (SVM, Logistic Regression) and quantum models (QSVC, VQC), finding that classical SVM (0.76 accuracy) outperformed QSVC (0.74 accuracy) in that particular setting, suggesting that while QML shows long-term promise, its practical advantage over CML is not yet consistent across all diabetes datasets and configurations. Singh and Patel [21] further examined quantum simulation scenarios and disease classification behavior specifically on diabetes data, reinforcing the observation that quantum model performance is highly sensitive to encoding strategy and dataset characteristics.

2.2. QML in Related Healthcare Domains

Beyond diabetes, QML has been explored across a range of adjacent healthcare classification problems, offering useful methodological parallels. Bali et al. [22] proposed QuantumNet, a hybrid classical deep learning-quantum transfer learning model for diabetic retinopathy detection, while De et al. [23] introduced a federated quantum neural network

(FQPDR) for privacy-preserving early detection of diabetic retinopathy across distributed clinical sites. In the domain of cardiovascular disease, Maheshwari et al. [24] applied QML to electronic healthcare records for ischemic heart disease classification, and Babu et al. [25] demonstrated quantum-enhanced machine learning for heart disease prediction, both reporting measurable gains over classical baselines. Munshi et al. [26] extended this line of work to heart failure detection within a Healthcare 4.0 framework, while Maouaki et al. [27] applied a Quantum Support Vector Machine to prostate cancer detection, reporting a detailed performance analysis relative to classical SVM formulations.

Collectively, these studies indicate a consistent pattern: QML models, particularly quantum kernel-based classifiers such as QSVC, tend to perform competitively or favorably against classical counterparts on small-to-moderate healthcare datasets, but reported gains vary considerably depending on the disease domain, encoding strategy, and dataset size, with few studies systematically examining performance across varying data scales or reporting the computational trade-offs (memory, training/inference time) that accompany quantum advantages.

Table 1 summarizes ten representative studies most closely related to the present work, comparing their target disease, datasets, classical and quantum models employed, and key reported outcomes.

Table 1. Comparative summary of related QML studies in healthcare classification.

Ref.	Disease/Domain	Dataset	Models Compared	Key Reported Findings
[16]	Diabetes	PIMA Diabetes	LDA/PCA + QSVC, VQC	Hybrid dimensionality reduction improved QSVC/VQC performance; qubit scarcity remains a limiting factor.
[17]	Diabetes (acute conditions)	Diabetes dataset	Decision Tree, Random Forest vs. Qboost	Qboost achieved 0.69 across accuracy, precision, recall, F1, specificity; 55-fold speed-up over classical models.
[18]	Type 2 Diabetes with comorbidities	Clinical dataset (Stokes parameter encoding)	Classical baseline vs. VQC (Qiskit)	VQC achieved 0.70–0.72 accuracy using 2–3 qubits on NISQ devices.
[19]	Diabetes	Diabetes dataset	Logistic Regression + QKS	QKS-based feature transformation improved classification accuracy over conventional Logistic Regression.
[20]	Diabetes	PIMA Diabetes	SVM, LR vs. QSVC, VQC	Classical SVM (0.76) outperformed QSVC (0.74); QML competitive but not yet superior in this setting.
[21]	Diabetes	Diabetes dataset	Classical vs. quantum simulation	Highlighted sensitivity of quantum classification behavior to simulation scenario and data characteristics.
[22]	Diabetic Retinopathy	Retinal image dataset	Deep learning vs. quantum transfer learning (QuantumNet)	Hybrid classical-quantum transfer learning enhanced detection performance over classical deep learning alone.
[23]	Diabetic Retinopathy	Distributed clinical data	Federated quantum neural network (FQPDR)	Enabled privacy-preserving detection with competitive accuracy across distributed institutions.
[24]	Ischemic Heart Disease	Electronic Health Records	Classical ML vs. QML	QML models demonstrated improved classification of cardiac risk from EHR data.
[27]	Prostate Cancer	Clinical dataset	SVM vs. QSVM	QSVM provided a competitive performance analysis relative to classical SVM for cancer detection.

Recent work also motivates a more careful treatment of interpretability, feature representation, entanglement design, and robustness. Mohasel Arjomandi and Jabbarvand proposed interpretable on-the-fly repair of uncertain neural classifications, emphasizing that post hoc reliability mechanisms should be distinguished from feature-level interpretability [28]. Mehrnia and Elbaz showed that entanglement topology can materially alter QML performance, reinforcing that circuit connectivity itself is an experimental factor rather than a neutral implementation detail [29]. Kieslich et al. demonstrated how PCA-derived representations can support biomedical prediction while still requiring supervised validation of predictive relevance [30]. In a healthcare classification setting, Shalmashi and Kabiri further highlighted the importance of training-only balancing and robustness-oriented evaluation for imbalanced ECG data [31]. These studies directly motivate the leakage-controlled preprocessing, feature-map ablation, supervised attribution, and robustness reporting adopted in the revised experiments.

2.3. Critical Analysis and Research Gap

The reviewed literature confirms that QML, and quantum kernel methods in particular, hold measurable promise for healthcare classification tasks, including diabetes prediction. However, a critical examination of these studies reveals several recurring limitations that the present work seeks to address.

First, the majority of existing diabetes-focused QML studies [16–21] rely on small or moderately sized datasets, typically limited to a few hundred or a few thousand records (e.g., the PIMA dataset), and few, if any, evaluate how QML performance scales as dataset size grows into the tens of thousands of records. This leaves an open question regarding the practical viability of QML for large-scale, real-world clinical deployment, which the present study directly addresses through its dataset-scaling experiments (Section 4.4).

Second, reported outcomes across studies are inconsistent: while some works report clear quantum advantages [17–19], others find that classical models remain competitive or superior [20,21], and no clear consensus has emerged regarding the conditions under which QML reliably outperforms CML. This inconsistency suggests a need for controlled, side-by-side benchmarking using identical preprocessing, feature selection, and evaluation protocols across both paradigms, which the present study provides.

Third, existing studies rarely report the computational costs of QML in terms of training time, inference time, and memory consumption, focusing predominantly on classification accuracy and related metrics. Given the well-documented resource constraints of simulating quantum circuits classically, this omission limits the practical usefulness of prior findings for real-world deployment decisions. The present work explicitly quantifies these trade-offs (Section 4.5), providing evidence-based guidance for when QML is computationally justified.

Fourth, few of the reviewed studies incorporate systematic feature identification or interpretability analysis alongside model comparison; most treat feature selection as a preprocessing step rather than an object of empirical investigation in its own right. This study addresses that gap by combining PCA-based feature analysis with algorithm benchmarking to identify the specific clinical and lifestyle-related predictors most relevant to diabetes risk.

Finally, related work in adjacent domains such as diabetic retinopathy [22,23] and cardiovascular disease [24–27] demonstrates that QML's applicability extends beyond diabetes classification alone, but cross-domain comparisons of QML scalability and computational behavior remain scarce. By combining empirical benchmarking, dataset-scale analysis, computational cost profiling, and feature-level interpretability within a single unified framework, this study addresses these gaps collectively, offering a more complete empirical basis

for evaluating the practical trade-offs between QML and CML in diabetes prediction than has been provided by prior work.

3. Proposed Framework

This section presents the proposed framework for evaluating Classical Machine Learning (CML) and Quantum Machine Learning (QML) approaches in diabetes prediction. The framework begins with a formal problem statement that unifies the classification logic underlying all four models considered in this study (Section 3.1), followed by an overview of the end-to-end pipeline (Section 3.2). It then details the data preprocessing procedures (Section 3.3), introduces the foundational quantum computing concepts required to interpret the quantum models (Section 3.4), and finally describes the classical and quantum classification models themselves (Section 3.5).

3.1. Problem Formulation

Let $x_i \in \mathbb{R}^d$ denote a patient feature vector and $y_i \in \{0, 1\}$ the diabetes label. The objective is to learn a model f from the training data only and estimate $\hat{y}_i = f(x_i)$ for previously unseen observations. The revised study does not use a hand-crafted rule of the form $\max(F) \geq T$ as a unifying classifier because SVC, KNN, QSVC, and QKNN employ fundamentally different decision mechanisms. Instead, Algorithm 1 describes the common experimental workflow while each classifier retains its own decision rule.

Algorithm 1 Leakage-Controlled Experimental Workflow

1: **Input:** dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ and random seed s
2: Create a stratified training/test split using seed s
3: Estimate admissible outlier bounds on training data only
4: Fit scaler on training data and transform training/test data
5: Fit PCA on training data and retain eight components; transform test data
6: **if** balanced-training experiment **then**
7: Apply SMOTE to the transformed training partition only
8: **end if**
9: Tune classical hyperparameters using training data only
10: Encode the eight PCA components into eight qubits for QML models
11: Train SVC, KNN, classical baselines, QSVC, and QKNN on common partitions
12: Evaluate on the untouched test partition using positive-class and threshold-independent metrics
13: Repeat for seeds $\{17, 42, 73, 101, 202\}$ and summarize mean, SD, 95% CI, and paired tests

This formulation makes the comparison reproducible and prevents the preprocessing and metric inconsistencies identified during code audit. In particular, the test set is never used to estimate PCA loadings, scaling parameters, SMOTE samples, or tuning decisions.

3.2. Framework Overview

Having established the underlying classification problem in Section 3.1, this subsection presents the overall pipeline used to implement and evaluate both the CML and QML approaches. Figure 2 illustrates the end-to-end workflow adopted in this study, spanning data collection, preprocessing, model selection, evaluation, and empirical analysis.

The pipeline begins with Data Collection , in which an open-source diabetes dataset comprising over 253,000 records is obtained from Kaggle. This is followed by the Preprocessing stage, which encompasses data cleaning, feature selection and dimensionality reduction using PCA, and, specific to the quantum pipeline, feature mapping to prepare classical data for encoding into quantum states. These preprocessing steps are described in detail in Section 3.3.

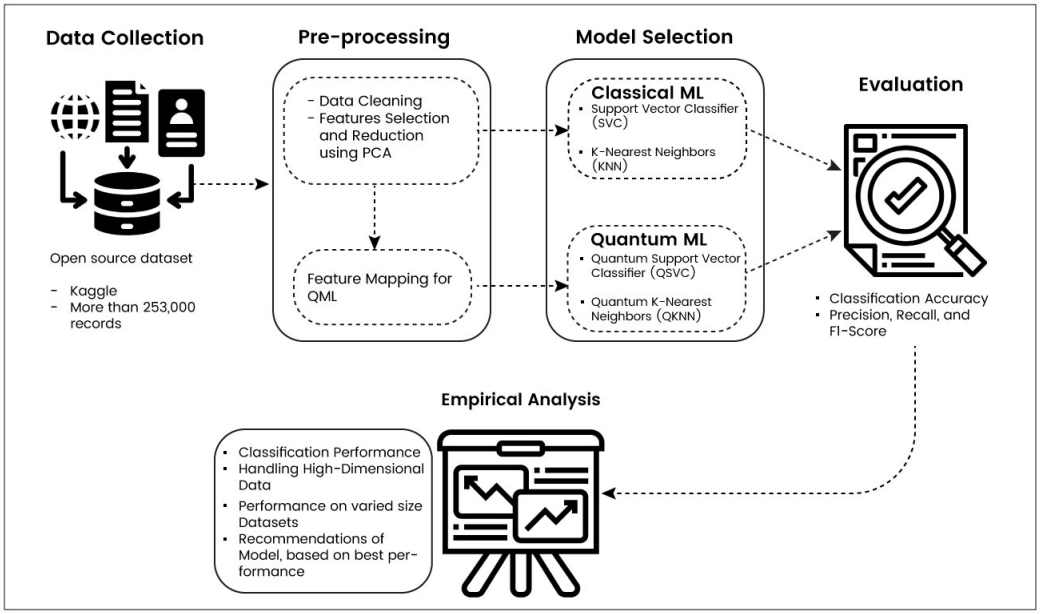

Figure 2. Proposed framework.

The prepared data is then passed to the Model Selection stage, in which two classical models, the Support Vector Classifier (SVC) and K-Nearest Neighbors (KNN), and their quantum-enhanced counterparts, the Quantum Support Vector Classifier (QSVC) and Quantum K-Nearest Neighbors (QKNN), are trained independently on the processed data. The theoretical background necessary to interpret the quantum models is presented in Section 3.4, while the models themselves, both classical and quantum, are described in detail in Section 3.5.

Following model training, the Evaluation stage assesses each model using classification accuracy, precision, recall, and F1 score, providing a consistent basis for comparing CML and QML performance (detailed in Section 4.3). Finally, the Empirical Analysis stage synthesizes these results to assess classification performance, evaluate each approach's ability to handle high-dimensional data, examine performance across varying dataset sizes, and generate recommendations regarding the most suitable model based on observed performance characteristics. The outcomes of this final stage are reported and discussed in full in Section 4.

Taken together, this pipeline ensures that the comparison between CML and QML models is conducted under identical preprocessing, feature selection, and evaluation conditions, thereby supporting a fair and reproducible empirical assessment of both paradigms.

3.3. Data Preprocessing

The revised preprocessing pipeline was designed explicitly to prevent information leakage. For every random seed, a stratified 80/20 train/test split was created first. All subsequent preprocessing parameters were estimated from the training partition only and then applied unchanged to the test partition.

3.3.1. Data Cleaning

Class imbalance was handled in two explicitly separated regimes: (i) the original imbalanced training distribution and (ii) a training-only SMOTE regime. Synthetic samples were never introduced into the test set. The IQR rule was restricted to genuinely continuous/count-like variables (BMI, MentHlth, and PhysHlth) because applying IQR filtering to binary survey indicators can incorrectly treat valid minority responses as out-

liers. IQR bounds were estimated on the training partition and used only for clipping the corresponding training and test variables.

3.3.2. Feature Reduction Using PCA

After training-only standardization, PCA was fitted only on the training partition and the first eight components were retained so that the quantum representation used eight qubits. If X_{train} denotes the standardized training matrix, PCA learns an orthogonal projection W_8 from X_{train} and transforms observations as $z = W_8^\top x$. The same fitted W_8 is then applied to the untouched test partition. Importantly, PCA is unsupervised: large loadings identify variables contributing strongly to variance in a component, not variables that are necessarily predictive of diabetes. Supervised permutation importance is therefore reported separately in Section 4.4.

3.3.3. Classical-to-Quantum Feature Encoding

For QML, the eight PCA coordinates were mapped to $[0, \pi]$ using a min–max transformation fitted on the training partition only. For an encoded vector $z = (z_1, \ldots, z_8)$, the state preparation can be written as

$$|\phi(z)\rangle = U_\phi(z)|0\rangle^{\otimes 8}, \tag{1}$$

where U_ϕ is an eight-qubit ZZ-style feature map containing Hadamard preparation, data-dependent phase rotations, and pairwise ZZ interactions. Thus, the classical values are not "stored in the qubits" directly; they parameterize the gates that prepare a data-dependent quantum state. Figure 3 should be read in this order: the eight scaled PCA values enter as classical parameters, one coordinate per qubit, and the circuit transforms the initial state into $|\phi(z)\rangle$.

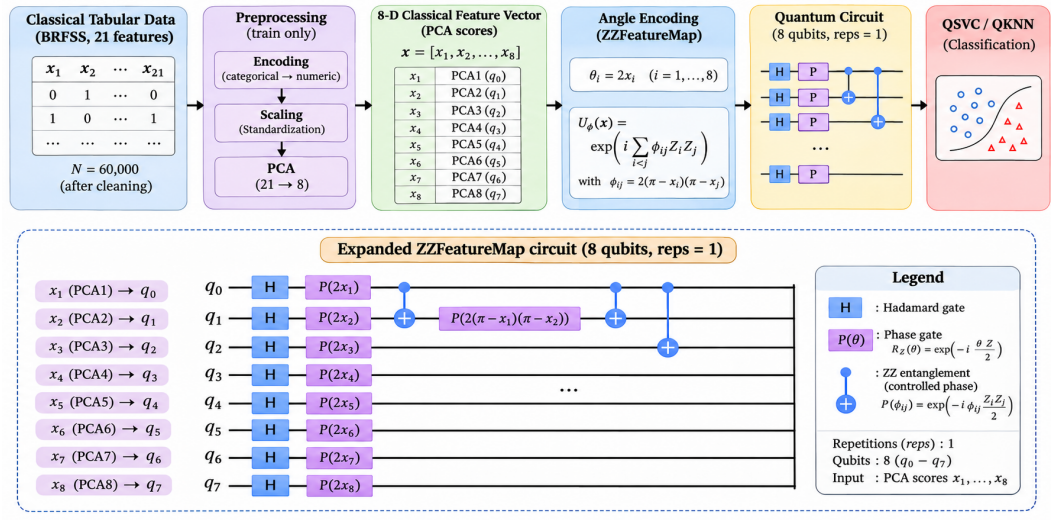

Figure 3. Revised classical-to-quantum encoding pipeline. Twenty-one BRFSS predictors are processed using training-fitted encoding/scaling and PCA, producing eight classical PCA scores $x_1, \ldots, x_8$. These scores parameterize an eight-qubit ZZFeatureMap with the explicit mapping $x_1 \to q_0, \ldots, x_8 \to q_7$; the resulting fidelity representation is used by QSVC/QKNN. Qubits therefore represent data-dependent quantum states rather than raw classical feature values.

The primary quantum benchmark used one full-entanglement repetition. A controlled ablation compared full and linear entanglement with one and two repetitions (Section 4.6.3). No physical quantum processor or calibrated hardware-noise model was used; all quantum states and fidelity kernels were evaluated by classical statevector simulation.

3.4. Preliminaries on Quantum Computing

Only the quantum concepts required to interpret the implemented kernels are summarized here. An eight-qubit register is initialized in $|0\rangle^{\otimes 8}$. Hadamard gates create superposition, while data-dependent phase operations and ZZ interactions encode the PCA coordinates and introduce correlations among qubits. The resulting state $|\phi(x)\rangle$ defines a fidelity kernel

$$K_Q(x_i, x_j) = |\langle \phi(x_i) | \phi(x_j) \rangle|^2. \quad (2)$$

This quantity is computed for pairs of observations and supplied either to an SVC as a precomputed kernel or to the QKNN similarity rule. The present experiments use classical statevector simulation; therefore, runtime and memory results characterize simulated quantum kernels and must not be interpreted as evidence of hardware-level quantum speed-up. Finite-shot sampling, gate noise, readout errors, and decoherence on NISQ hardware would introduce additional uncertainty and can reduce kernel fidelity. Moreover, as shown by stochastic-entanglement studies [29], the chosen entanglement topology can materially affect QML performance; this motivates the feature-map ablation reported later.

3.5. Classification Models

Model selection involves identifying the most effective CML or QML technique for a given problem, which requires careful consideration of the dataset characteristics, the problem domain, and the evaluation metrics used. This study investigates the relative advantages and limitations of four models: two classical models, SVC and KNN, and their quantum-enhanced counterparts, QSVC and QKNN. Each model instantiates the shared decision-rule formalism introduced in Algorithm 1 (Section 3.1) in a model-specific way, as detailed below.

3.5.1. Classical Machine Learning Models

Classical Machine Learning (CML) models evaluate data, detect patterns, and generate predictions using traditional computational approaches. CML models typically follow a workflow of data preparation, model training, evaluation, and prediction, relying on algorithms that learn directly from structured datasets.

Support Vector Classifier (SVC)

SVC was selected for its robustness in high-dimensional spaces. By constructing hyperplanes to separate data classes, SVC provides a strong classical baseline for comparison with its quantum-enhanced counterpart, QSVC. SVC is a supervised learning algorithm derived from the Support Vector Machine (SVM) family, adapted specifically for classification tasks. While SVM more broadly addresses both classification and regression by identifying an optimal separating hyperplane, SVC focuses exclusively on classification and supports a range of kernels (e.g., linear, polynomial, and Radial Basis Function) to handle complex decision boundaries, as illustrated in Figure 4.

In this study, SVC classifies individuals into two groups, Diabetic (D_{p_i}) and Non-Diabetic (ND_{p_i}), based on a set of health-related parameters $F = \{f_1, f_2, f_3, \ldots, f_n\}$, such as blood pressure, BMI, and glucose level, represented as a feature vector x_i for each patient p_i. The classification model $M(x_i)$ determines class membership relative to a learned decision boundary T, obtained by solving the margin-maximization problem in Equation (3):

$$\min_{w,b} \|w\|^2 + C \sum_{i=1}^{n} \max\left(0, 1 - y_i(w^T x_i + b)\right) \quad (3)$$

where w is the weight vector, b is the bias term, C is the regularization parameter, and y_i is the class label for patient p_i (1 for diabetic, 0 for non-diabetic). Consistent with the

decision-rule formalism introduced in Algorithm 1, the resulting classification is expressed in Equation (4):

$$M(x_i) = \begin{cases} D_{p_i}, & \text{if } \max(F_{x_i}) \geq T \\ ND_{p_i}, & \text{otherwise} \end{cases} \quad (4)$$

where $\max(F_{x_i})$ denotes the most relevant feature value for patient x_i and T is the threshold decision limit learned during training. This formulation allows SVC to both classify patients and identify the feature contributing most strongly to each prediction, offering a transparent and clinically interpretable approach to diabetes classification.

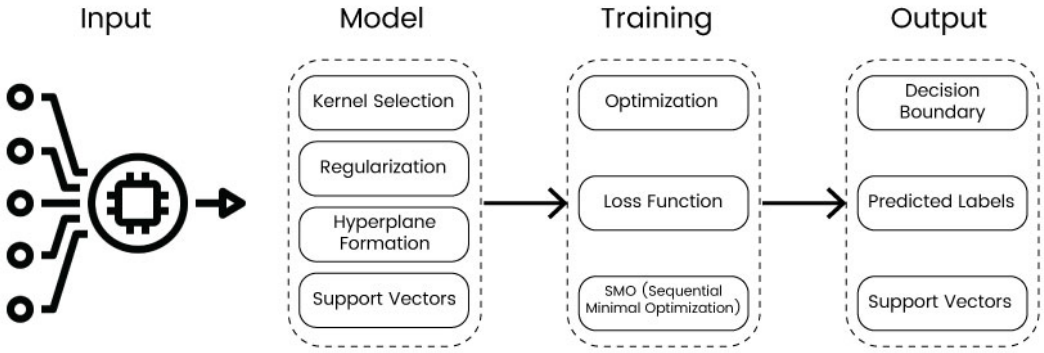

Figure 4. SVC architecture.

K-Nearest Neighbors (KNN)

As a non-parametric method, KNN offers simplicity and interpretability, with performance that depends heavily on data structure and feature selection, making it a useful classical benchmark against QKNN. KNN classifies a data point according to the majority class among its k-Nearest Neighbors in feature space, with distance typically computed using the Euclidean metric, as shown in Equation (5):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (5)$$

where x and y are two data points and x_i, y_i are their respective feature values. Depending on the application, alternative distance measures such as Manhattan or Minkowski distance may also be used. The number of neighbors k is a key hyperparameter: a large k smooths the decision boundary and risks underfitting, while a small k increases sensitivity to noise and risks overfitting. Selecting an appropriate value of k (Figure 5) is therefore typically performed via cross-validation or other hyperparameter tuning techniques.

In this study, KNN categorizes patients into Diabetic (D_{p_i}) and Non-Diabetic (ND_{p_i}) groups based on the same set of health-related parameters $F = \{f_1, f_2, f_3, \ldots, f_n\}$ used by SVC. The classification model $M(x_i)$ assigns a class based on the majority vote among the k-Nearest Neighbors of the patient's feature vector F_{x_i}, as defined in Equation (6):

$$M(x_i) = majority_{group(k-nearest\ neighbors\ x_i)} \quad (6)$$

Here, the algorithm first computes the distance between x_i and every other point in the dataset using the chosen distance metric and then selects the k-Nearest Neighbors points and assigns the most common class label among them. This procedure provides clear insight into how individual data points influence a new patient's classification, making KNN highly interpretable. Furthermore, KNN's intrinsic non-linearity allows it to capture complex decision boundaries that linear classifiers such as SVC may struggle to represent.

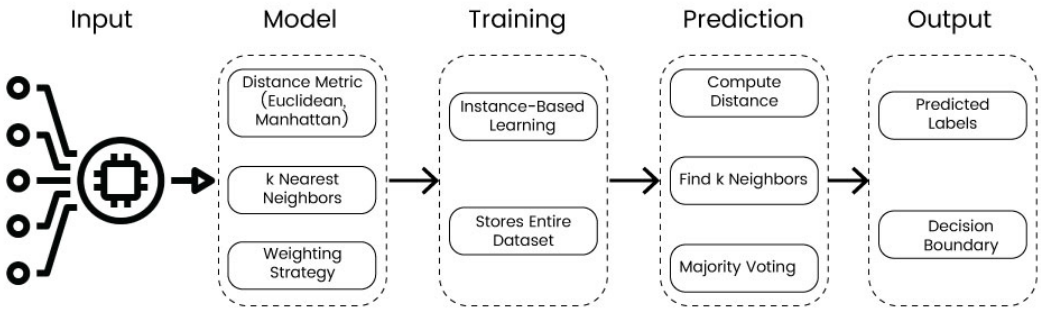

Figure 5. KNN architecture.

3.5.2. Quantum Machine Learning Models

Quantum Support Vector Classifier (QSVC)

QSVC (presented in Figure 6) differs from classical SVC in the kernel construction, not in a hand-crafted threshold rule. Given the quantum fidelity kernel in Equation (2), the SVC decision function is

$$f(x) = \sum_{i \in SV} \alpha_i y_i K_Q(x_i, x) + b, \qquad \hat{y} = \mathbb{I}[f(x) \geq 0], \tag{7}$$

where SV denotes support vectors, α_i are learned dual coefficients, and b is the intercept. Equation (7) clarifies the distinction from the classical RBF-SVC: both optimize an SVM decision boundary, but QSVC replaces the classical kernel by fidelity between quantum-encoded states.

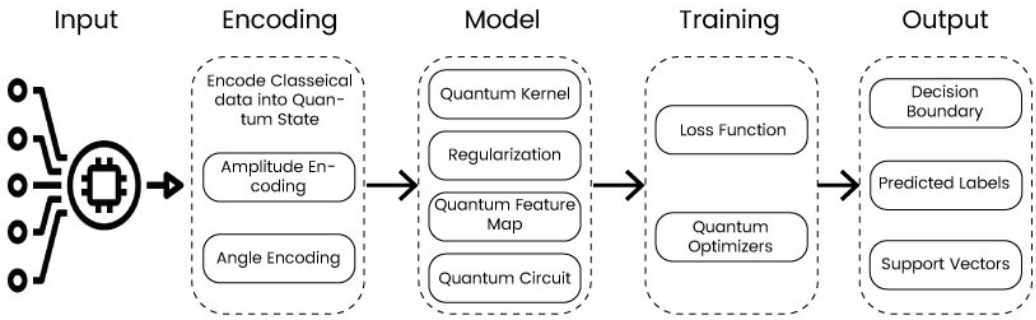

Figure 6. QSVC architecture based on an eight-qubit fidelity kernel followed by a classical SVC optimization step.

Quantum K-Nearest Neighbors (QKNN)

QKNN (presented in Figure 7) uses the same encoded states but replaces Euclidean distance with quantum-state fidelity. For query x, let $\mathcal{N}_k(x)$ denote the k training observations having the largest $K_Q(x_i, x)$. A fidelity-weighted positive-class score is

$$s(x) = \frac{\sum_{i \in \mathcal{N}_k(x)} K_Q(x_i, x) y_i}{\sum_{i \in \mathcal{N}_k(x)} K_Q(x_i, x)}, \qquad \hat{y} = \mathbb{I}[s(x) \geq 0.5]. \tag{8}$$

Unlike the earlier implementation, ROC-AUC and PR-AUC are calculated from the continuous score $s(x)$ rather than from hard class labels. The principal QKNN benchmark used $k = 15$, matching the tuned classical KNN neighborhood size on the common benchmark.

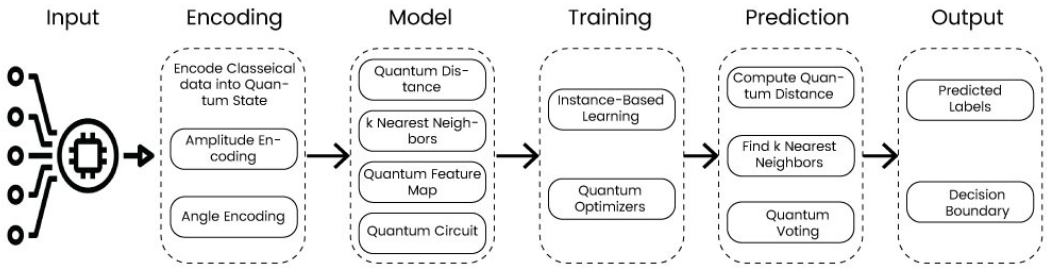

Figure 7. QKNN architecture using quantum-state fidelity as the neighbor-similarity measure.

4. Results and Discussion

This section presents a comprehensive empirical evaluation of the CML and QML approaches described in Section 3. The evaluation is designed to address the research contributions outlined in Section 1 (RC1–RC4) by measuring classification performance, examining scalability across varying dataset sizes, quantifying computational efficiency in terms of runtime and memory usage, and identifying the most influential features for diabetes prediction.

The results are organized as follows. Section 4.1 describes the two benchmark datasets used for evaluation. Section 4.2 details the experimental setup and implementation environment, and Section 4.3 defines the evaluation metrics adopted throughout this study. Section 4.4 presents the feature identification results obtained via PCA, while Section 4.5 examines model scalability across dataset sizes ranging from 10,000 to 55,000 records. Section 4.6 reports detailed performance comparisons on both benchmark datasets, and Section 4.7 quantifies the computational costs of each approach in terms of training time, inference time, and memory consumption. Finally, Section 4.8 synthesizes these findings into a broader empirical discussion, and Section 4.9 translates the results into practical recommendations for model selection under different deployment scenarios.

Taken together, this structure allows the relative strengths and limitations of CML and QML to be assessed not only in terms of raw predictive performance but also in terms of the practical trade-offs, scalability, resource demands, and interpretability that determine their suitability for real-world deployment in diabetes prediction pipelines.

4.1. Datasets Used

The large dataset is the 2015 Behavioral Risk Factor Surveillance System (BRFSS)-derived Diabetes Health Indicators dataset. The supplied binary working file contains 60,000 records and 21 predictors plus the target; 51,839 records (86.40%) are non-diabetic and 8161 (13.60%) are diabetic. The full supplied BRFSS-derived file contains 253,680 records. The corrected preprocessing clips training-derived bounds for appropriate continuous/count-like variables rather than deleting observations, so the final working size remains $N = 60{,}000$. The dataset and reproducibility materials used for this revision are available from the permanent project repository: https://github.com/ziauddin55/Quantum-Computing (accessed on 24 October 2025). The second benchmark is the PIMA Indians Diabetes dataset with 768 observations, of which 500 are non-diabetic and 268 are diabetic. Binary/ordinal survey variables were retained numerically; continuous/count-like variables were standardized as described in Section 3.3.

To evaluate the CML and QML approaches described in Section 3, two diabetes-related benchmark datasets were selected: the Diabetes Binary dataset (Table 2) and the PIMA Indians Diabetes dataset (Table 3). Both datasets were chosen for their real-world medical relevance [32] and because they jointly present a range of practical challenges relevant to this study, including imbalanced class distributions, diverse feature types (binary, ordinal, and

continuous), and complex interdependencies among clinical and lifestyle-related variables, making them well suited for assessing the comparative performance of diabetes prediction models.

The Diabetes Binary dataset, summarized in Table 2, comprises 21 health-related predictors spanning clinical indicators (e.g., HighBP, HighChol, Stroke), lifestyle behaviors (e.g., Smoker, PhysActivity, Fruits, Veggies), and self-reported health status (e.g., GenHlth, MentHlth, PhysHlth), alongside demographic variables (Age, Sex), with a binary diabetes outcome label.

The PIMA Indians Diabetes dataset, summarized in Table 3, instead comprises a smaller set of clinically measured, continuous features, including glucose concentration, blood pressure, skin thickness, insulin level, and BMI, together with a diabetes pedigree function reflecting family history, and serves as a widely used benchmark in diabetes prediction research due to its moderate size and diagnostic relevance to a specific patient population (women of Pima Indian descent).

Using two datasets that differ in scale, feature composition, and target population allows the comparative evaluation in the following subsections to assess not only raw model performance but also the robustness of each approach, CML and QML alike, across differing data conditions, an important consideration given the scalability concerns discussed later in Section 4.5.

Table 2. Diabetes binary dataset description.

Feature	Description	Definition
HighBP	Indicates if the patient has high blood pressure.	(0 = No, 1 = Yes)
HighChol	Indicates if the patient has high cholesterol levels.	(0 = No, 1 = Yes)
BMI	Body Mass Index, a measure of body fat.	Numerical
Stroke	Indicates if the patient has a history of stroke.	(0 = No, 1 = Yes)
HeartDiseaseOrAttack	Indicates heart disease or heart attack history.	(0 = No, 1 = Yes)
CholCheck	Cholesterol checked within last 5 years.	(0 = No, 1 = Yes)
Smoker	Smoked at least 100 cigarettes.	(0 = No, 1 = Yes)
PhysActivity	Physically active in the past 30 days.	(0 = No, 1 = Yes)
Fruits	Daily fruit consumption.	(0 = No, 1 = Yes)
Veggies	Daily vegetable consumption.	(0 = No, 1 = Yes)
HvyAlcoholConsump	Heavy alcohol consumer.	(0 = No, 1 = Yes)
AnyHealthcare	Has any healthcare coverage.	(0 = No, 1 = Yes)
NoDocbcCost	Could not see a doctor due to cost.	(0 = No, 1 = Yes)
GenHlth	Self-rated general health.	1 (Excellent)–5 (Poor)
MentHlth	Days mental health was not good (past 30 days).	Range 0–30
PhysHlth	Days physical health was not good (past 30 days).	Range 0–30
DiffWalk	Difficulty walking or climbing stairs.	(0 = No, 1 = Yes)
Age	Ordinal BRFSS age category.	Categories 1–13
Sex	Biological sex.	(0 = Female, 1 = Male)
Education	Highest education category.	Ordinal categories 1–6
Income	Household income category.	Ordinal categories 1–8
Diabetes_Binary	Diabetes status.	(0 = No, 1 = Yes)

Table 3. PIMA Indians diabetes dataset description.

Feature	Description	Definition
Pregnancies	Number of times the patient has been pregnant.	Numerical
Glucose	Plasma glucose concentration (2 h oral glucose tolerance test).	Numerical
BloodPressure	Diastolic blood pressure (mm Hg).	Numerical
SkinThickness	Triceps skinfold thickness (mm).	Numerical
Insulin	2 h serum insulin (mu U/mL).	Numerical
BMI	Body Mass Index calculated as weight (kg)/height $(m)^2$.	Numerical
DiabetesPedigreeFunction	Function score related to diabetes likelihood based on family history.	Numerical
Age	Patient's age (years).	Numerical
Outcome	Diabetes diagnosis outcome.	(0 = No diabetes, 1 = Diabetes)

4.2. Experimental Setup

All corrected comparisons use common stratified partitions and five seeds {17, 42, 73, 101, 202}. For the BRFSS direct CML–QML benchmark, a fixed stratified 2000-record subset was used because dense statevector fidelity kernels grow quadratically with the number of training observations; the same records and partitions were used for every compared model. Larger BRFSS subsets from 10,000 to 55,000 records were used for measured classical scalability and analytical quantum-kernel memory requirements. PIMA was evaluated in full.

Classical baselines comprised Logistic Regression, RBF-SVC, KNN, Random Forest, and Gradient Boosting. Hyperparameter selection was performed by stratified five-fold cross-validation using training data only; the untouched test partition was never used for model selection. The search considered alternative SVC kernels and regularization/kernel-width settings, and alternative KNN neighbor counts, weighting rules, and distance metrics. The selected direct-benchmark settings were $C = 1$ and $\gamma = 0.1$ for RBF-SVC and $k = 15$ with distance weighting for KNN. The exact candidate grids and selection scripts are provided in the permanent GitHub repository. QML used an eight-qubit ZZ-style fidelity feature map with one repetition for the primary comparison. Both imbalanced and training-only SMOTE regimes were evaluated. Final uncertainty was assessed separately using five independent stratified 80/20 train/test splits with seeds {17, 42, 73, 101, 202}; thus, cross-validation was used for tuning, whereas the reported five-seed results are repeated locked-holdout evaluations.

The quantum experiments were executed through deterministic statevector simulation of the explicitly defined feature-map unitary and fidelity kernel. This produces the same mathematical object used by a precomputed-kernel QSVC while avoiding dependence on a particular cloud device. No claim of real-hardware acceleration is made. Runtime measurements therefore reflect the present classical implementation and are intended only as implementation-cost indicators.

The corrected experiments were executed on a Linux x86_64 environment with five available CPU cores on an Intel Xeon Platinum 8573C processor, approximately 5.8 GiB RAM, Python 3.13.5, and no GPU or quantum processing unit. No HPC cluster was used for the corrected benchmark. Because the quantum component is statevector-simulated, its timing and memory values characterize this classical execution environment rather than physical-quantum hardware performance.

4.3. Evaluation Metrics

A robust comparison between Classical Machine Learning (CML) and Quantum Machine Learning (QML) requires metrics that capture both predictive performance and practical applicability. This study adopts four standard classification metrics, accuracy, precision, recall, and F1 score, to assess how well each model's predictions align with actual patient outcomes and how effectively each model generalizes to unseen data. These metrics are reported for all four models (SVC, KNN, QSVC, QKNN) across both benchmark datasets in Section 4.6. In addition, computational efficiency is assessed separately in Section 4.7 using training time, inference time, and memory usage, providing a complete picture of the practical trade-offs between CML and QML beyond raw classification performance alone.

4.3.1. Accuracy

Accuracy measures the overall proportion of correct predictions among all predictions made, as defined in Equation (9):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

While straightforward and effective for balanced datasets, accuracy can be misleading on imbalanced datasets, such as the diabetes datasets used in this study, where the minority (diabetic) class is underrepresented; a model that consistently predicts the majority class can still achieve a high accuracy score despite performing poorly on the minority class. For this reason, accuracy is interpreted alongside precision, recall, and F1 score throughout the evaluation in Section 4.6, rather than in isolation.

4.3.2. Precision

Precision measures the proportion of true positive predictions among all positive predictions, both correct and incorrect, as defined in Equation (10):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

High precision indicates a low rate of false positives, which is particularly important in clinical contexts where incorrectly classifying a healthy patient as diabetic could lead to unnecessary anxiety, follow-up testing, or treatment.

4.3.3. Recall (Sensitivity)

Recall measures the proportion of true positive predictions among all actual positive cases, as defined in Equation (11):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{11}$$

High recall ensures that the majority of actual diabetic cases are correctly identified, which is critical in healthcare applications where failing to detect a true positive (a false negative) can delay diagnosis and treatment, with potentially serious consequences for patient outcomes.

4.3.4. F1 Score

The F1 score is the harmonic mean of precision and recall, providing a single balanced measure that remains informative even for imbalanced datasets, as defined in Equation (12):

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{12}$$

Because the F1 score penalizes models that achieve high precision at the expense of recall (or vice versa), it is used throughout Section 4.6 as the primary metric for summarizing each model's overall balance between correctly identifying diabetic patients and avoiding false positive diagnoses.

In the revised experiments, precision, recall, and F1 refer to the positive (diabetic) class unless explicitly stated otherwise. Balanced accuracy, specificity, ROC-AUC, PR-AUC, and Matthews correlation coefficient (MCC) are also reported. This distinction is important because, in single-label classification, weighted recall is numerically equal to accuracy and can conceal poor minority-class sensitivity on an imbalanced test set. All test sets preserve their natural class distribution and contain no synthetic SMOTE observations. Five-seed means, standard deviations, and 95% confidence intervals are reported; paired two-sided Wilcoxon signed-rank tests are used as a non-parametric sensitivity analysis, with the limited power of $n = 5$ explicitly acknowledged.

4.4. Feature Representation and Predictive Attribution

PCA was retained for dimensionality reduction, but its interpretation was corrected (Shown in Table 4). PCA loadings indicate how strongly original variables contribute to

the variance represented by each principal component; they do not establish association with the diabetes target. Explained-variance ratios and loading matrices were computed independently for each seed, using training data only.

Table 4. Five-seed mean PCA explained-variance ratios for the first eight BRFSS principal components. The complete 21×8 loading matrices are provided in the GitHub reproducibility package.

PC	Explained Variance	Cumulative Variance
PC1	0.1683	0.1683
PC2	0.0881	0.2564
PC3	0.0648	0.3212
PC4	0.0587	0.3799
PC5	0.0561	0.4360
PC6	0.0517	0.4877
PC7	0.0491	0.5368
PC8	0.0452	0.5819

To evaluate target-specific predictive relevance, permutation importance was calculated on the untouched test data using a supervised classical model. The highest-ranked original variables were HighBP, BMI, GenHlth, HighChol, Age, HvyAlcoholConsump, Stroke, and DiffWalk. These rankings are therefore discussed as predictive attribution, whereas PCA loadings are discussed only as representation structure. This separation prevents the earlier over-interpretation of variables such as MentHlth or PhysHlth solely because they had large PCA coefficients and aligns with broader work showing that dimensionality reduction and supervised feature relevance answer different questions [30].

4.5. Scalability Analysis Across Dataset Sizes

The previous scalability figure used fixed values and therefore could not support statistical claims about scaling behavior. It has been replaced by measured classical experiments on stratified BRFSS subsets of 10,000, 15,000, ..., 55,000 records and by an exact memory calculation for dense QSVC kernels. For each size, Logistic Regression and KNN were evaluated under both the original imbalanced training distribution and training-only SMOTE (Shown in Table 5).

Table 5. Selected measured scalability results from corrected BRFSS experiments. Recall is for the diabetic class.

N	Model	Regime	Accuracy	Recall+	ROC-AUC
10,000	Logistic Regression	Imbalanced	0.842	0.165	0.808
10,000	Logistic Regression	SMOTE	0.723	0.717	0.806
30,000	Logistic Regression	Imbalanced	0.868	0.149	0.817
30,000	Logistic Regression	SMOTE	0.738	0.754	0.816
55,000	Logistic Regression	Imbalanced	0.867	0.159	0.807
55,000	Logistic Regression	SMOTE	0.727	0.728	0.806

The main pattern is a class-imbalance trade-off rather than a monotonic loss of classical scalability: imbalanced training preserves high raw accuracy (approximately 0.84–0.87) but yields low diabetic-class recall, whereas training-only SMOTE raises recall to approximately 0.72–0.77 while reducing raw accuracy. ROC-AUC remains comparatively stable across sample sizes. Thus, raw accuracy alone is not an appropriate basis for a scalability claim in this imbalanced clinical problem (Shown in Table 6).

Table 6. Dense float64 QSVC kernel memory requirement under an 80/20 split.

Dataset N	Train N	Test N	Total Kernel Memory (GB)
10,000	8000	2000	0.64
20,000	16,000	4000	2.56
30,000	24,000	6000	5.76
40,000	32,000	8000	10.24
50,000	40,000	10,000	16.00
55,000	44,000	11,000	19.36
60,000	48,000	12,000	23.04

For the 60,000-record working dataset, the dense training and test kernel matrices alone require approximately 23.04 GB in float64 representation. This quadratic $O(N^2)$ storage requirement explains the observed allocation failures in the original simulator-scale workflow and contradicts any claim that the present simulated QSVC "scales better" computationally than classical baselines. The result should instead be interpreted as a practical limitation of dense simulated quantum kernels.

4.6. Performance Comparison on Benchmark Datasets

The primary comparison uses mean test performance across five seeds. To avoid conflating majority-class accuracy with clinical sensitivity, the tables emphasize balanced accuracy, diabetic-class recall and F1, ROC-AUC, and PR-AUC. Both natural-imbalance and training-only SMOTE experiments were performed; the SMOTE regime is shown below because it provides a more clinically informative sensitivity comparison while preserving an untouched test set.

4.6.1. BRFSS Common 2000-Record Benchmark

QSVC attained the highest raw accuracy in this balanced-training experiment, but its diabetic-class recall was only 0.326 and its balanced accuracy was 0.607. Logistic Regression achieved substantially higher recall (0.748; 95% CI 0.703–0.793) and ROC-AUC (0.811; 95% CI 0.800–0.821). QKNN produced higher recall than QSVC (0.667) but lower raw accuracy and ROC-AUC. Under the natural imbalanced training distribution, QSVC predicted almost entirely the majority class (mean diabetic recall 0.000), demonstrating why the earlier accuracy-only interpretation was misleading. These corrected results do not establish a general predictive quantum advantage(Shown in Table 7).

Table 7. Five-seed mean performance on the common 2000-record BRFSS benchmark with training-only SMOTE.

Model	Accuracy	Balanced Acc.	Recall+	F1+	ROC-AUC	PR-AUC
Majority baseline	0.865	0.500	0.000	0.000	0.500	0.135
Logistic Regression	0.738	0.742	0.748	0.435	0.811	0.379
SVC	0.738	0.691	0.626	0.393	0.770	0.327
KNN	0.692	0.689	0.685	0.376	0.759	0.317
Random Forest	0.794	0.650	0.452	0.372	0.781	0.320
Gradient Boosting	0.738	0.698	0.644	0.398	0.779	0.329
QSVC-ZZ	0.812	0.607	0.326	0.319	0.737	0.286
QKNN-ZZ	0.653	0.658	0.667	0.341	0.725	0.298

4.6.2. PIMA Indian Diabetes Dataset

On PIMA, tuned SVC provided the highest mean ROC-AUC (0.836; 95% CI 0.780–0.892) and diabetic-class recall (0.741; 95% CI 0.629–0.852), while Logistic Regression achieved

the highest mean accuracy (0.770) and F1 (0.689). Both quantum-kernel models were competitive in some thresholded metrics but did not outperform the strongest tuned classical baselines. Across five seeds, paired two-sided Wilcoxon comparisons between QSVC and the strongest classical models generally produced $p \geq 0.0625$; given only five paired runs, these tests are low powered and are treated as sensitivity analyses rather than definitive evidence of equivalence or superiority (Shown in Table 8).

Table 8. Five-seed mean performance on PIMA with training-only SMOTE.

Model	Accuracy	Balanced Acc.	Recall+	F1+	ROC-AUC	PR-AUC
Majority baseline	0.649	0.500	0.000	0.000	0.500	0.351
Logistic Regression	0.770	0.760	0.726	**0.689**	0.833	0.717
SVC	0.757	**0.753**	**0.741**	0.681	**0.836**	**0.736**
KNN	0.696	0.699	0.707	0.621	0.779	0.655
Random Forest	0.758	0.742	0.689	0.666	0.815	0.686
Gradient Boosting	0.749	0.734	0.681	0.655	0.807	0.685
QSVC-ZZ	0.723	0.697	0.607	0.607	0.761	0.638
QKNN-ZZ	0.677	0.668	0.641	0.579	0.738	0.607

4.6.3. Feature-Map and QKNN Ablation

A controlled BRFSS seed-42 ablation compared full versus linear ZZ entanglement with one and two repetitions. Full entanglement with one repetition produced ROC-AUC 0.731, while two repetitions reduced ROC-AUC to 0.627. Linear entanglement produced ROC-AUC 0.675 with one repetition and 0.644 with two repetitions. Thus, increasing circuit depth did not improve performance in this setting, and the entanglement topology materially affected the resulting kernel. This observation is consistent with recent evidence that constructive entanglement patterns are task dependent [29]. QKNN used fidelity rather than Euclidean distance and continuous fidelity-weighted scores for ROC-AUC/PR-AUC; its performance remained below the strongest classical baseline on both corrected benchmarks.

This ablation shows that the QKNN results depend on the fidelity-based similarity rather than on the classical Euclidean/Minkowski neighbor geometry. The quantum similarity did not improve ROC-AUC or PR-AUC on either corrected benchmark. The full-versus-linear ZZ comparison above is used as the lower-resource feature-map comparison: linear entanglement reduces the number of pairwise interactions while preserving the same eight-qubit input dimension (Shown in Table 9).

Table 9. Numerical neighbor-similarity ablation under training-only SMOTE. Classical KNN uses the selected conventional distance-based neighbor rule, whereas QKNN replaces that similarity with quantum-state fidelity.

Dataset	Method	Recall+	ROC-AUC	PR-AUC
BRFSS ($N = 2000$)	Classical KNN	0.685	0.759	0.317
BRFSS ($N = 2000$)	QKNN (fidelity)	0.667	0.725	0.298
PIMA ($N = 768$)	Classical KNN	0.707	0.779	0.655
PIMA ($N = 768$)	QKNN (fidelity)	0.641	0.738	0.607

4.7. Computational Efficiency Analysis

The revised runtime analysis separates state preparation, pairwise quantum-kernel construction, and classifier fitting. Because all quantum operations were simulated classically, these measurements characterize simulator overhead and must not be interpreted as a hardware quantum speed advantage. On the common 2000-record benchmark, the simulated QSVC pipeline required substantially more computation than lightweight classical

KNN/Logistic Regression. Across the recorded five-seed BRFSS runs, state preparation and pairwise kernel construction were timed separately from final precomputed-SVC fitting; representative seed-17 SMOTE timings were approximately 0.116 s for state preparation, 0.185 s for kernel construction, and 0.466 s for classifier fitting on this reduced benchmark. These timings are implementation-specific and are reported only to show where simulator cost is incurred; at larger sample sizes, dense kernel storage becomes the dominant constraint.

The key scalability result is the quadratic memory requirement in Table 6. At 60,000 records with an 80/20 split, the dense float64 training and test quantum-kernel matrices require approximately 23.04 GB before accounting for statevectors, intermediate arrays, Python/Qiskit objects, or operating-system overhead. Consequently, the previously reported approximately 24-s "training time" cannot represent the complete quantum pipeline when kernel generation itself is included; the revised manuscript therefore avoids comparing only the final SVC fit against end-to-end classical training.

Real quantum hardware could change the cost model, but finite shots, noise, calibration drift, readout error, and communication overhead would also affect the observed kernel. No physical device or calibrated noise model was used here, so the present study makes no empirical claim about wall-clock performance on quantum hardware.

4.8. Empirical Analysis and Discussion

Three conclusions follow from the corrected experiments. First, class imbalance is a dominant factor: high raw accuracy can coexist with very poor diabetic-class recall, as seen in the natural-imbalance BRFSS runs. Second, tuned classical baselines remain highly competitive and, on the corrected BRFSS and PIMA benchmarks, generally outperform the evaluated fidelity-kernel QSVC/QKNN configurations in ROC-AUC and clinically relevant sensitivity. Third, simulated quantum kernels impose a severe $O(N^2)$ storage burden, limiting direct large-scale evaluation even when classification behavior is interesting.

Interpretability is also treated more narrowly in the revision. PCA explains representation variance but is not a target-aware attribution method. Supervised permutation importance is therefore used separately, and the manuscript no longer describes large PCA loadings as proof of diabetes predictive signal. This distinction is especially important in safety-critical settings, where interpretable model-repair and reliability methods address different questions from feature-space compression [28].

The results should therefore be interpreted as a controlled benchmark of simulated quantum representations rather than evidence of computational quantum supremacy or universal QML superiority. The biomedical prediction literature similarly shows that performance depends strongly on representation design, balancing, validation, and robustness procedures [30,31].

4.9. Recommendations

For the present tabular diabetes tasks, tuned classical models should remain the default choice when deployment cost, reproducibility, and minority-class sensitivity are priorities. Simulated quantum kernels are most useful as research tools for studying alternative similarity representations on computationally manageable subsets. Future quantum studies should pre-register common data partitions, report positive-class and threshold-independent metrics, include tuned classical baselines, test multiple feature maps/entanglement topologies, and evaluate noisy finite-shot hardware before making claims of practical advantage.

Threats to Validity

The corrected study still has limitations. The BRFSS direct quantum comparison uses a stratified 2000-record common subset because dense fidelity kernels are computationally

expensive; therefore, quantum predictive results should not be extrapolated to the full 60,000-record dataset. Statevector simulation is deterministic and omits finite-shot and hardware noise. Only five random seeds were used, limiting statistical-test power. Hyperparameter searches were deliberately bounded to maintain reproducibility and computational feasibility. Finally, BRFSS is survey-derived rather than a prospective clinical diagnostic cohort, and PIMA represents a specific population, so external clinical validity requires independent datasets and prospective evaluation, consistent with robustness concerns raised in healthcare classification studies [31].

5. Conclusions and Future Work

This revised study provides a leakage-controlled reassessment of classical and simulated Quantum Machine Learning for diabetes prediction. The experimental code and protocol were corrected so that splitting precedes all fitted preprocessing, SMOTE is restricted to training data, metrics are defined consistently for the diabetic class, tuned classical baselines are included, repeated seeds and confidence intervals are reported, and PCA is separated from supervised feature attribution.

Under these controls, the earlier claim of a general "quantum advantage" is not supported. On the common BRFSS benchmark with training-only SMOTE, QSVC achieved higher raw accuracy than several models but substantially lower diabetic-class recall and ROC-AUC than Logistic Regression; on PIMA, tuned SVC and Logistic Regression were stronger than the evaluated QSVC/QKNN configurations on the main discrimination metrics. The feature-map ablation further showed that additional repetitions did not improve the quantum kernel and that entanglement topology affects results.

The most robust contribution is therefore a reproducible characterization of the trade-offs among class imbalance, predictive discrimination, feature representation, and simulator cost. Measured classical scalability experiments remain stable through 55,000 records, while dense QSVC kernel storage grows quadratically and reaches approximately 23.04 GB at 60,000 records. Future work should evaluate approximate/low-rank quantum kernels, finite-shot noisy simulation, and real quantum hardware using the same locked data partitions and statistically powered repeated comparisons. External validation on independent clinical cohorts is also required before any diagnostic deployment claim can be made.

Author Contributions: Conceptualization, Z.U.D., H.M., A.S.A.-S. and L.H.C.; methodology, Z.U.D., I.D.l.T.D. and R.S.R.S.; software, Z.U.D.; validation, H.M., G.M., S.U.R.K. and A.S.A.-S.; formal analysis, Z.U.D. and R.S.R.S.; investigation, Z.U.D. and I.D.l.T.D.; resources, N.A.; data curation, Z.U.D.; writing—original draft preparation, Z.U.D.; writing—review and editing, H.M., G.M., S.U.R.K., N.A., A.S.A.-S., I.D.l.T.D., R.S.R.S. and L.H.C.; visualization, Z.U.D.; supervision, H.M., G.M. and L.H.C.; project administration, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset files used for the revised experiments, together with the corrected experimental scripts, fixed random seeds, data-partition logic, hyperparameter-selection code, repeated-run outputs, PCA/loading outputs, scalability tables, and supporting reproducibility materials, are permanently available in the project GitHub repository: https://github.com/ziauddin55/Quantum-Computing (accessed on 24 October 2025). The BRFSS-derived file originates from the 2015 Behavioral Risk Factor Surveillance System Diabetes Health Indicators data, and the PIMA Indians Diabetes dataset is used as the second public benchmark.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CML	Classical Machine Learning
QML	Quantum Machine Learning
SVC	Support Vector Classifier
KNN	K-Nearest Neighbor
QSVC	Quantum Support Vector Classifier
QKNN	Quantum K-Nearest Neighbor
RF	Random Forest
PCA	Principal Component Analysis
NISQ	Noisy Intermediate-Scale Quantum
BMI	Body Mass Index

References

1. Ogurtsova, K.; Guariguata, L.; Barengo, N.C.; Lopez-Doriga Ruiz, P.; Sacre, J.W.; Karuranga, S.; Sun, H.; Boyko, E.J.; Magliano, D.J. IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Res. Clin. Pract.* **2022**, *183*, 109118. [CrossRef]
2. Nanditha, A.; Ma, R.C.W.; Ramachandran, A.; Snehalatha, C.; Chan, J.C.N.; Chia, K.S.; Shaw, J.E.; Zimmet, P.Z. Diabetes in Asia and the Pacific: Implications for the Global Epidemic. *Diabetes Care* **2016**, *39*, 472–485. [CrossRef]
3. Battineni, G.; Sagaro, G.G.; Chinatalapudi, N.; Amenta, F. Applications of Machine Learning Predictive Models in the Chronic Disease Diagnosis. *J. Pers. Med.* **2020**, *10*, 21. [CrossRef]
4. Kasten, J.E. Big Data Applications in Healthcare Administration. *Int. J. Big Data Anal. Healthc.* **2020**, *5*, 12–37. [CrossRef]
5. Deberneh, H.M.; Kim, I. Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *Int. J. Environ. Res. Public Health* **2021**, *18*, 3317. [CrossRef]
6. Jha, S.K.; Adhikari, N.B. Analysis of Ensembled Learning using Classical and Quantum Boosting Approaches for Diabetes Mellitus Prediction. *Res. Sq.* 2024. [CrossRef]
7. Rustam, F.; Al-Shamayleh, A.S.; Shafique, R.; Obregon, S.A.; Iglesias, R.C.; Gonzalez, J.P.M.; Ashraf, I. Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model. *Sci. Rep.* **2024**, *14*, 23274. [CrossRef]
8. Hailat, M.; Hailat, Z.; Ifraitekh, M.; Zakaraya, Z.; Shalash, M.; Al-Ani, I.; Abu Dayyih, W. Bisoprolol and Amlodipine Co-Administration with Glimepiride in a Diabetic Rat Model: A Statistical and Machine Learning Analysis. *Pharmaceuticals* **2026**, *19*, 1064. [CrossRef]
9. Thanasilp, S.; Wang, S.; Cerezo, M.; Holmes, Z. Exponential concentration in quantum kernel methods. *Nat. Commun.* **2024**, *15*, 5200. [CrossRef]
10. Kanamori, Y.; Yoo, S.M. Quantum Computing: Principles and Applications. *J. Int. Technol. Inf. Manag.* **2020**, *29*, 43–71. [CrossRef]
11. Huang, H.L.; Du, Y.; Gong, M.; Zhao, Y.; Wu, Y.; Wang, C.; Li, S.; Liang, F.; Lin, J.; Xu, Y.; et al. Experimental Quantum Generative Adversarial Networks for Image Generation. *Phys. Rev. Appl.* **2021**, *16*, 024051. [CrossRef]
12. Upama, P.B.; Kolli, A.; Kolli, H.; Alam, S.; Syam, M.; Shahriar, H.; Ahamed, S.I. Quantum Machine Learning in Disease Detection and Prediction: A survey of applications and future possibilities. In Proceedings of the 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, 26–30 June 2023; pp. 1545–1551. [CrossRef]
13. Ethan, A. Machine Learning for Medical Diagnostics: A Review in Bioinformatics. *OSF Prepr.* 2023. [CrossRef]
14. Abbas, A.; Sutter, D.; Zoufal, C.; Lucchi, A.; Figalli, A.; Woerner, S. The power of quantum neural networks. *Nat. Comput. Sci.* **2021**, *1*, 403–409. [CrossRef]
15. Wu, Y.; Wu, B.; Wang, J.; Yuan, X. Quantum Phase Recognition via Quantum Kernel Methods. *Quantum* **2023**, *7*, 981. [CrossRef]
16. Hancco-Quispe, J.K.; Borda-Colque, J.P.; Torres-Cruz, F. Quantum Machine Learning Applied to the Classification of Diabetes. *arXiv* **2022**, arXiv:2301.00109. [CrossRef]
17. Maheshwari, D.; Garcia-Zapirain, B.; Sierra-Soso, D. Machine learning applied to diabetes dataset using Quantum versus Classical computation. In Proceedings of the 2020 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, 9–11 December 2020; pp. 1–6. [CrossRef]
18. Sierra-Sosa, D.; Arcila-Moreno, J.D.; Garcia-Zapirain, B.; Elmaghraby, A. Diabetes Type 2: Poincaré Data Preprocessing for Quantum Machine Learning. *Comput. Mater. Contin.* **2021**, *67*, 1849–1861. [CrossRef]
19. Kumbargeri, V.; Sinha, N.; Chickerur, S. A Quantum Machine Learning Classifier Model for Diabetes. In *Innovative Data Communication Technologies and Application*; Raj, J.S.; Iliyasu, A.M.; Bestak, R.; Baig, Z.A., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer: Singapore, 2021; Volume 59, pp. 603–613. [CrossRef]

20. Vimal, D. PIMA Diabetes Prediction Using Machine Learning and Quantum Machine Learning Techniques. In Proceedings of the 2024 ITU Kaleidoscope: Innovation and Digital Transformation for a Sustainable World (ITU K), New Delhi, India, 21–23 October 2024; pp. 1–7. [CrossRef]
21. Singh, A.; Patel, N.D. Quantum simulation scenarios and disease classification behaviour on diabetes data. *Int. J. Ad Hoc Ubiquitous Comput.* **2023**, *44*, 109–123. [CrossRef]
22. Bali, M.; Mishra, V.P.; Yenkikar, A.; Chikmurge, D. QuantumNet: An enhanced diabetic retinopathy detection model using classical deep learning-quantum transfer learning. *MethodsX* **2025**, *14*, 103185. [CrossRef]
23. De, D.; Nandy Pal, M.; Hazra, D. FQPDR: Federated quantum neural network for privacy-preserving early detection of diabetic retinopathy. *Evol. Intell.* **2024**, *17*, 4047–4068. [CrossRef]
24. Maheshwari, D.; Ullah, U.; Osorio Marulanda, P.A.; García-Olea Jurado, A.; Diez Gonzalez, I.; Ormaetxe Merodio, J.M.; Garcia-Zapirain, B. Quantum Machine Learning Applied to Electronic Healthcare Records for Ischemic Heart Disease Classification. *Hum.-Centric Comput. Inf. Sci.* **2023**, *13*, 6. [CrossRef]
25. Babu, S.V.; Ramya, P.; Gracewell, J. Revolutionizing heart disease prediction with quantum-enhanced machine learning. *Sci. Rep.* **2024**, *14*, 7453. [CrossRef]
26. Munshi, M.; Gupta, R.; Jadav, N.K.; Polkowski, Z.; Tanwar, S.; Alqahtani, F.; Said, W. Quantum machine learning-based framework to detect heart failures in Healthcare 4.0. *Softw. Pract. Exp.* **2024**, *54*, 168–185. [CrossRef]
27. El Maouaki, W.; Said, T.; Bennai, M. Quantum Support Vector Machine for Prostate Cancer Detection: A Performance Analysis. *arXiv* **2024**, arXiv:2403.07856. [CrossRef]
28. Mohasel Arjomandi, H.; Jabbarvand, R. Interpretable On-the-Fly Repair of Deep Neural Classifiers. In *Proceedings of the 1st International Workshop on Dependability and Trustworthiness of Safety-Critical Systems with Machine Learned Components (SE4SafeML 2023)*; ACM: New York, NY, USA, 2023, pp. 14–17. [CrossRef]
29. Mehrnia, M.; Elbaz, M.S.M. Stochastic Entanglement Configuration for Constructive Entanglement Topologies in Quantum Machine Learning with Application to Cardiac MRI. *arXiv* **2025**, arXiv:2507.11401.
30. Kieslich, C.A.; Alimirzaei, F.; Song, H.; Do, M.; Hall, P. Data-driven prediction of antiviral peptides based on periodicities of amino acid properties. *Comput. Aided Chem. Eng.* **2021**, *50*, 2019–2024. [CrossRef] [PubMed Central]
31. Shalmashi, P.; Kabiri, M. Deep Neural Network Architectures for Robust ECG Classification in Healthcare Systems. In Proceedings of the IISE Annual Conference 2026, Health Systems/Data Analytics and Information Systems, Arlington, TX, USA, 16–19 May 2026.
32. Chang, V.; Ganatra, M.A.; Hall, K.; Golightly, L.; Xu, Q.A. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthc. Anal.* **2022**, *2*, 100118. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.