



A Transparent Hybrid Quantum-Classical Framework for Automated Multi-Class Diabetic Retinopathy Grading

Bilal Babayigit¹, Mustafa S. M. Alazzawi², Ahmed alkhateb³

^{1,2}Department of Computer Engineering, Faculty of Engineering, Erciyes University Kayseri, Türkiye.

³Ibn Sina University for Medical Sciences, Medical College, Baghdad, Iraq.

To Cite this Article: Bilal Babayigit¹, Mustafa S. M. Alazzawi², Ahmed alkhateb³, “A Transparent Hybrid Quantum-Classical Framework for Automated Multi-Class Diabetic Retinopathy Grading”, Indian Journal of Computer Science and Technology, Volume 05, Issue 02 (May-August 2026), PP: 1066-1078.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: Diabetic Retinopathy (DR) is a leading cause of preventable blindness worldwide, requiring early and accurate multi-class grading to prevent vision loss. Conventional screening faces challenges due to specialist shortages and manual evaluation variability, while classical deep learning models struggle with subtle non-linear inter-class variations and lack clinical transparency. To address this, we propose an interpretable hybrid Quantum Convolutional Neural Network (QCNN) framework. The pipeline uses a fine-tuned DenseNet121 backbone to extract 1024-dimensional spatial features, which are compressed via a multi-layer classical bottleneck to a 4-dimensional latent vector. These features are mapped into a 4-qubit Variational Quantum Circuit (VQC) using parallel R_x angle embedding and processed through parameterized R_y/R_z rotations and circular CNOT entanglement gates to enhance representation capacity. Evaluated on 1,506 retinal fundus images mapped across three severity classes (No/Mild, Moderate, and Severe DR), an independent 20% holdout test set ($N=302$) yielded a certified multi-class accuracy of 84.10%, an F1-score of 0.83, an AUC of 0.95 for severe cases, and a real-time inference latency of 0.085 seconds per image. To ensure clinical trust, SHapley Additive exPlanations (SHAP) was integrated to backmap quantum expectations into pixel-level heatmaps, accurately highlighting localized microaneurysms and exudates. This framework demonstrates the feasibility of transparent, quantum-enhanced AI for real-time medical decision support.

Key Word: Quantum Machine Learning, Diabetic Retinopathy, Hybrid QCNN, Variational Quantum Circuits, Explainable AI, SHAP.

I. INTRODUCTION

Diabetes mellitus represents one of the most critical global health emergencies of the twenty-first century, with its worldwide prevalence continuing to escalate rapidly and placing an unprecedented burden on healthcare infrastructure [1]. Among its various chronic microvascular complications, Diabetic Retinopathy (DR) stands out as a primary cause of preventable and irreversible visual impairment and blindness among the working-age population globally [2, 3]. Persistent hyperglycemia induces systemic microvascular damage within the retina, leading to progressive structural alterations such as capillary occlusion, vascular leakage, microaneurysms, intraretinal hemorrhages, and eventual abnormal neovascularization [3, 4]. Early detection, precise risk stratification, and timely clinical intervention are paramount to halting disease progression and preventing catastrophic vision loss [3, 5].

Retinal fundus photography remains the primary non-invasive modality for population-scale DR screening and clinical evaluation [5, 6]. Standard clinical guidelines, such as the International Clinical Diabetic Retinopathy severity scale, grade the disease across multiple stages ranging from non-proliferative manifestations to advanced proliferative DR [5]. However, routine manual interpretation of fundus images by certified ophthalmologists is severely constrained by its labor-intensive nature, inherent subjectivity, and substantial inter-observer variability [8]. These diagnostic bottlenecks are further aggravated by a critical worldwide shortage of trained retinal specialists, particularly in rural and low-resource medical settings [9, 10].

To address these screening limitations, automated Computer-Aided Diagnosis (CAD) systems driven by deep learning most notably Convolutional Neural Networks (CNNs) have emerged as a transformative technology in medical image analysis [11, 12]. Pre-trained architectures such as DenseNet121 leverage dense inter-layer connectivity to facilitate feature reuse and mitigate vanishing gradient issues, demonstrating state-of-the-art capabilities in extracting complex hierarchical retinal representations from fundus photographs [13, 14]. Nevertheless, classical CNN architectures encounter fundamental barriers when modeling highly complex, non-linear feature interactions, often struggling to reliably discriminate between adjacent, visually overlapping DR severity stages [15, 16]. Furthermore, the inherent "black-box" nature of deep neural networks poses a major obstacle to clinical adoption, as medical practitioners require auditable and transparent reasoning to ensure safety, accountability, and compliance with stringent regulatory frameworks such as FDA guidelines and GDPR mandates [17, 18].

Explainable Artificial Intelligence (XAI) frameworks, particularly SHapley Additive exPlanations (SHAP) based on cooperative game theory, have been integrated into medical AI to produce pixel-level attribution maps that highlight pathology-sensitive regions like microaneurysms and exudates [19-21]. Although SHAP significantly enhances visual transparency and builds clinical trust, it serves primarily as an interpretive tool and does not inherently expand the mathematical feature representation

capacity of the underlying classical neural network [15, 20]. Concurrently, Quantum Machine Learning (QML) utilizing Variational Quantum Circuits (VQCs) has introduced a novel computational paradigm [29, 30]. By projecting classical features into multi-dimensional Hilbert spaces using parameterized quantum gates and entangling operations, VQCs can capture complex non-linear feature correlations beyond the reach of classical architectures [22-24].

Despite extensive research in classical CNN-based DR screening and the growing adoption of standalone XAI tools, existing solutions remain compartmentalized. There is a distinct lack of unified multi-class DR grading frameworks that simultaneously leverage the superior non-linear feature representation capacity of Variational Quantum Circuits (VQCs) beyond classical limits, while seamlessly integrating pixel-level SHAP explainability to deliver transparent, reliable, and clinically deployable diagnostic decisions. To bridge this critical gap, the main objective of this study is to design, implement, and evaluate a novel hybrid Quantum Convolutional Neural Network (QCNN) architecture coupled with a post-hoc SHAP interpretability layer. By integrating DenseNet121 feature extraction, progressive classical dimensionality reduction, a 4-qubit Variational Quantum Circuit (VQC), and pixel-wise attribution backmapping, the proposed framework aims to achieve high-accuracy multi-class DR severity grading while providing visually grounded explanations that align with established clinical diagnostic criteria.

II. MATERIALS AND METHODS

2.1. Overview of the Proposed Framework

The paradigm shift in clinical computer-aided diagnosis requires simultaneous progress in the representational capacity of deep learning architectures and in the transparency of their decision-making processes. This thesis systematically engineers a novel unified hybrid quantum-classical pipeline coupled with a post-hoc Explainable Artificial Intelligence (XAI) interpretability layer to automate the multiclass severity grading of Diabetic Retinopathy (DR). The main objective of the proposed framework is to overcome the dimensionality constraints of classical deep Convolutional Neural Networks (CNNs) by embedding high-density classical feature representations in the multi-dimensional Hilbert space of a Variational Quantum Circuit (VQC) while maintaining clinical accountability through pixel-level attribution mapping.

The global operational architecture and algorithmic life cycle of the proposed methodology is systematically structured in seven sequential, interconnected phases. Each phase represents a separate computational stage in terms of functionality, but all phases are part and parcel of a unified end-to-end learning infrastructure.

The holistic pipeline progresses through the following algorithmic intervals (Figure 1):

- **Phase 1: Dataset Acquisition and Partitioning:** The input layer ingests a real-world clinical volume consisting of 1,506 digital fundus photographs. To shield the learning parameters from data leakage and optimistic generalization bias, a strict hierarchical partitioning protocol is enforced. At the macro-level, a 20% holdout split isolates an independent Testing Hub (302 images) for final inference. The remaining 80% development core (1,204 images) is further partitioned into a pristine Training Set (903 images) and an operational Validation Set (301 images) via a 75/25 internal split.
- **Phase 2: Comprehensive Preprocessing Protocol:** The ingested images intrinsically reflect the heterogeneous nature of routine clinical screening environments, with pronounced non-ideal artefacts, illumination exposure variations, and focus irregularities. In this phase, a robust digital standardisation pipeline was used, which included spatial resizing to $224 \times 224 \times 3$, multi-class ICDR balancing, and pixel intensity scaling to provide a unified computational base before feature extraction.
- **Phase 3: Deep Classical Feature Extraction:** Standardised retinal images are fed into a fine-tuned, pre-trained DenseNet121 convolutional backbone. A targeted domain adaptation transfer learning protocol is enacted by selectively unfreezing the high-level representational blocks, specifically Denseblock4 and the terminating Norm5 layer, to isolate a continuous, high-density spatial feature representation vector of 1,024 dimensions (1,024-D).
- **Phase 4: Multi-Layer Dimensionality Reduction:** Current Noisy Intermediate-Scale Quantum (NISQ) simulation devices exhibit rigorous physical restrictions regarding active qubit registers. To bridge this vast dimensional gap, a classical multi-layer linear bottleneck layer is designed to sequentially compress the 1024D feature space down to 256, then 64, and ultimately into a 4-dimensional (4D) dense latent vector bounded non-linearly via a Hyperbolic Tangent (Tanh) activation matrix.
- **Phase 5: Variational Quantum Optimization:** The compressed classical coordinates ($\theta_1, \theta_2, \theta_3, \theta_4$) are embedded into the quantum domain through Parallel RX Angle Embedding gates on a 4-qubit computational register. The initialised quantum state $|\psi\rangle$ is iteratively optimised by trainable parameterised Ansatz layers (RY, RZ rotation gates) and a circular Strongly Entangling Wall of Controlled-NOT (CNOT) gates to capture non-local pathological feature correlations in the Hilbert space.
- **Phase 6: Post-Hoc Explain Ability Layer:** To solve the inherent black-box problem in hybrid quantum neural architectures, a local model-agnostic interpretability framework based on SHAP is systematically integrated into the output interface. The Kernel Explainer algorithm receives continuous quantum expectation logits as input and performs a mathematical calculation to determine the marginal contribution of each individual spatial feature element's Shapley value to the final diagnostic prediction.
- **Phase 7: Multiclass Diagnostic Classification:** The calculated Shapley attribution weights are back mapped onto the original two-dimensional spatial geometry of the fundus image, producing 224×224 Spatial Pixel Attribution Maps (Clinical Heatmaps). Concurrently, the final classification scoring is scaled probabilistically by a categorical SoftMax activation function, which diagnoses each retinal sample into one of three clinical severity stages: No/Mild DR, Moderate DR, or Severe/Proliferative DR.

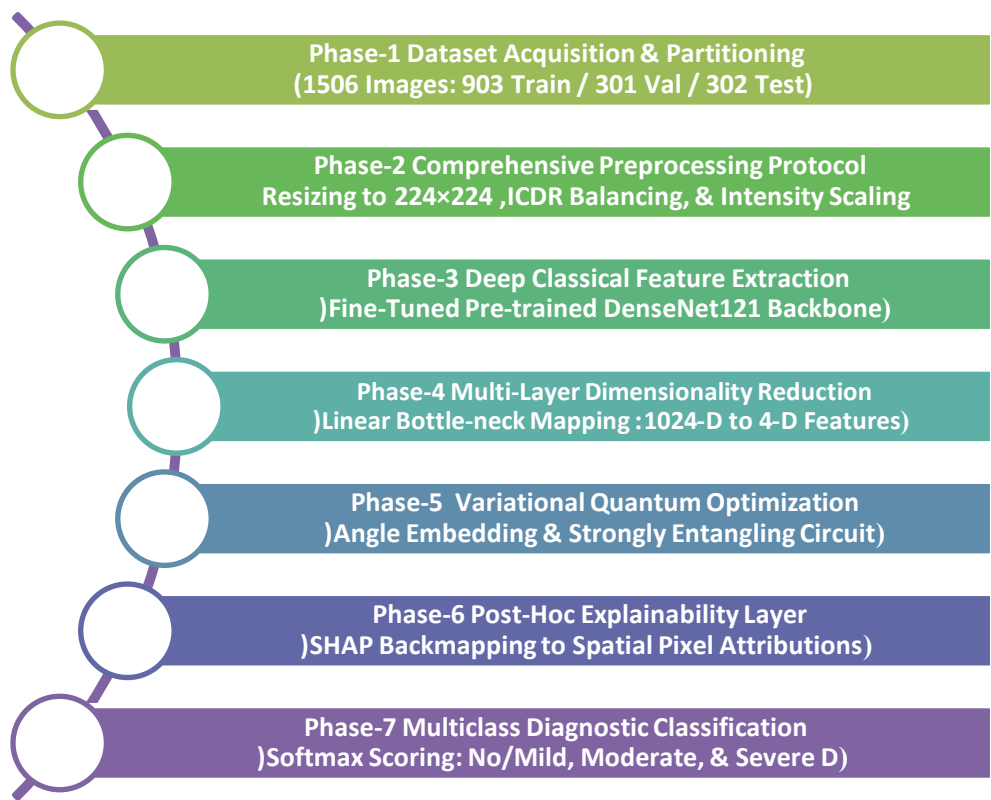


Figure 1: The proposed global research methodology and operational algorithmic framework for hybrid quantum-classical diabetic retinopathy severity classification.

2.2 Dataset Anatomy and Preprocessing Protocol

To ensure methodological rigor, clinical validity, and experimental reproducibility, this section delineates the data curation pipeline, clinical grading unification, multi-phase partitioning strategy, and standardizing preprocessing procedures engineered for the proposed Hybrid Quantum-Classical (QCNN-XAI) framework.

2.2.1 Clinical Dataset Characteristics & Real-World Heterogeneity

The evaluation of medical AI systems often fails due to over-idealized laboratory datasets. To stress-test the model against real-world clinical noise, a cohort of 1,506 digital retinal fundus photographs was gathered from longitudinal diabetic retinopathy screening programs. The dataset encompasses substantial real-world variabilities categorized across three main operational axes (Table 1):

Acquisition Hardware	Image Quality Artifacts	Diagnostic Grading
Heterogeneous cameras	Out-of-focus optics	Single-field fundus
Multi-site acquisition	Lens reflections	Expert clinical annotation (ICDR)
Operator variation	Over/Under exposure	

Table 1: Real-World Clinical Variability Axes

2.2.2 Class Consolidation and Severity Mapping

To align the diagnostic targets with the 4-qubit computational register of the Variational Quantum Circuit (VQC) and prevent majority-class gradient bias, the standard 5-tier International Clinical Diabetic Retinopathy (ICDR) scale was consolidated into three balanced, clinically actionable diagnostic classes (Table 2):

Target Class	ICDR Original Grades	Sample Count	Pathological Features and Clinical Signatures
Group 1: No / Mild DR	Grade 0 (No DR)	503	Normal fundus optics or isolated microaneurysms without significant hemorrhagic or exudative activity
	Grade 1 (Mild NPDR)		

Group 2: Moderate DR	Grade 2 (Moderate NPDR)	502	Localized intraretinal hemorrhages, microvascular changes, and distinct hard exudate deposits
Group 3: Severe / Proliferative DR	Grade 3 (Severe NPDR)	501	Widespread intraretinal microvascular abnormalities (IRMA), severe hemorrhages, and active neovascularization
	Grade 4 (Proliferative DR)		

Table 2: Clinical Severity Mapping and Class Consolidation Schema

2.2.3 Multi-Phase Hierarchical Partitioning

To enforce complete statistical independence and eliminate information leakage across model optimization and evaluation phases, a stratified multi-tier data partitioning strategy was implemented (Figure 2):

- 1. **External Holdout (Testing Hub):** A strict 20% external partition (302 images) was isolated prior to training and remained untouched until the final model benchmark.
- 2. **Development Core (80%):** The remaining 1,204 images were partitioned via a 75/25 stratified split into a pristine Training Set (903 images) and an operational Validation Set (301 images).

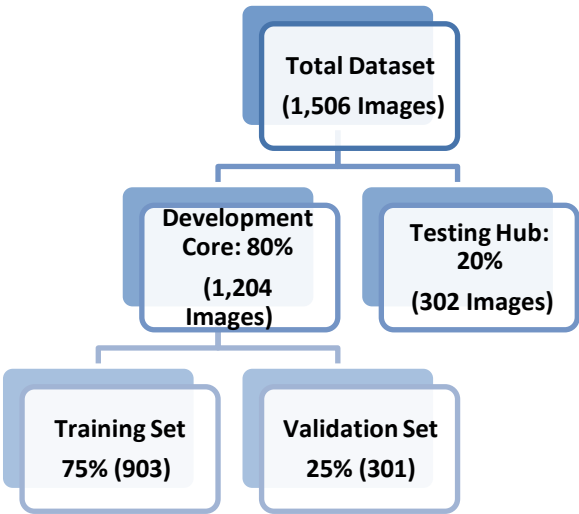


Figure 2. Stratified multi-tier data partitioning strategy.

The complete statistical distribution and partition allocation across all three clinical groups are formalized in Table 3.

Clinical Group (ICDR Scaled Class)	Total Images	Training Set (75% of Hub)	Validation Set (25% of Hub)	Independent Testing Set (20% External)
Group 1: No / Mild DR	503	301	101	101
Group 2: Moderate DR	502	302	100	100
Group 3: Severe / Proliferative DR	501	300	100	101
Aggregate Total (100%)	1,506	903	301	302

Table 3: Multi-Phase Dataset Distribution Across Experimental Subsets

2.2.4 Systematic Image Preprocessing Pipeline

Raw retinal fundus photographs exhibit uneven illumination, spatial dimension inconsistencies, and high dynamic range variances. A 4-stage standardized transformation pipeline was engineered to prepare the image tensors for classical feature extraction and downstream quantum angle encoding (Figure 4):

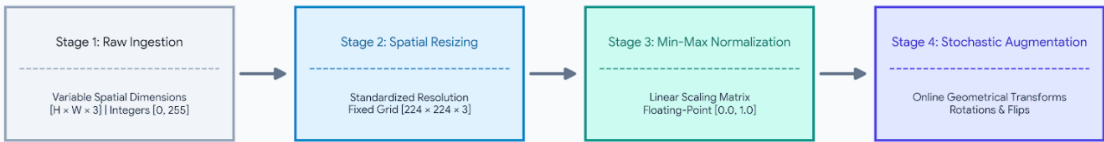


Figure 4. Systematic Image Preprocessing Pipeline flowchart

Execution Specifications for Preprocessing Pipeline:

- **Stage 1 (Raw Ingestion):** Input image array ingestion characterized by non-standardized spatial boundaries $H \times W \times 3$ and discrete integer intensities in $[0, 255]$.
- **Stage 2 (Spatial Standardization):** Downsampling via bilinear interpolation to a fixed grid size of $224 \times 224 \times 3$ pixels, establishing receptive field uniformity for DenseNet121.
- **Stage 3 (Intensity Min-Max Scaling):** Linear intensity rescaling transforming discrete integer inputs into floating-point tensors bounded within $[0.0, 1.0]$ to stabilize quantum gate initialization:

$$I_{\text{norm}}(x,y) = \frac{I_{\text{raw}}(x,y) - I_{\text{min}}}{I_{\text{max}} - I_{\text{min}}}$$

- **Stage 4 (On-the-Fly Stochastic Augmentation):** Dynamic online geometric and radiometric transformations (random affine rotations, horizontal flipping, and subtle contrast shifts) applied during training steps to prevent model overfitting while maintaining computational efficiency.

2.3 Hybrid Classical-Quantum Architecture and Dimensional Bottleneck

In this section, we present the core computational architecture of the proposed hybrid learning framework, which integrates a classical deep learning backbone with a variational quantum processing module via a carefully designed dimensional reduction interface. The architecture is driven by the physical constraints of current Noisy Intermediate-Scale Quantum (NISQ) systems, which limit direct high-dimensional encoding of raw image data due to exponential qubit scaling and decoherence noise. A hierarchical hybrid strategy is adopted: classical neural representations are extracted and compressed before being embedded into a low-dimensional Hilbert state space.

2.3.1 Architectural Design Rationale

The computational engine bridges the representational gap between deep convolutional representations and parametric quantum state optimization. Instead of feeding raw pixel tensors directly into the quantum register (which is computationally prohibitive on NISQ hardware), the architecture strategically utilizes a fine-tuned DenseNet121 backbone to isolate spatial feature maps. These high-density features are sequentially compressed to yield a 4-dimensional latent vector that directly maps to valid rotation angles for quantum state embedding.

2.3.2 DenseNet121 Feature Extraction with Transfer Learning

The preprocessed retinal images $I \in \mathbb{R}^{224 \times 224 \times 3}$ are fed into a fine-tuned DenseNet121 convolutional network. To optimize the backbone for ophthalmological lesion representation (microaneurysms, hemorrhages, and lipid exudates), a targeted domain-adaptation transfer learning protocol is executed:

- **Frozen Base Layers:** The initial shallow convolutional blocks are locked to preserve generalized low-level edge detection and texture primitives learned from ImageNet.
- **Selective Unfreezing:** The upper-level representational blocks specifically Denseblock4 and the terminating Norm5 layer are unfrozen for domain-specific fine-tuning.

This targeted feature adaptation isolates a dense spatial feature vector:

$$X_{\text{dense}} \in \mathbb{R}^{1024}$$

2.3.3 Multi-Layer Classical Dimensional Bottleneck

To bridge the dimensional gap between the 1,024-D classical feature vector and the physical 4-qubit capacity of the NISQ quantum register, a multi-layer linear bottleneck is engineered. This progressive compression prevents catastrophic information loss while distilling discriminative features:

$$\mathbb{R}^{1024} \xrightarrow{\text{FC}_1, \text{ReLU}, \text{Dropout}(0.4)} \mathbb{R}^{256} \xrightarrow{\text{FC}_2, \text{ReLU}} \mathbb{R}^{64} \xrightarrow{\text{FC}_3, \text{Tanh}} \mathbb{R}^4$$

The final 4-D latent coordinate vector $(\theta_1, \theta_2, \theta_3, \theta_4)$ is bounded non-linearly within the interval $[-1.0, +1.0]$ through a Hyperbolic Tangent (Tanh) activation, mapping directly to valid parametric rotation angles within the range $[-\pi, +\pi]$ for quantum angle embedding.

$$\theta_i = \tanh(W_3 \cdot \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 X_{\text{dense}} + b_1) + b_2) + b_3)_i$$

2.3.4 Execution Protocol and Layer Transitions

The complete 10-layer computational execution flow from raw image ingestion to multi-class classification is formalized in Table 4.

Stage	Layer Name & Configuration	Input Dimension	Transformation Operation	Output Dimension
Layer 1	Input Modality	Raw Images	Image Loading & Color Space Check	Variable Size
Layer 2	Image Pre-processing	Variable Size	Resizing & Min-Max Normalization	224×224×3
Layer 3	DenseNet121 Feature Extractor	224×224×3	Deep Convolutional Block Extraction	1024 Features
Layer 4	Classical Reducer (FC1)	1024 Features	Linear Mapping + ReLU + Dropout(0.4)	256 Features
Layer 5	Classical Reducer (FC2)	256 Features	Linear Mapping + ReLU Activation	64 Features
Layer 6	Quantum Interface (FC3)	64 Features	Linear Mapping + Tanh Scaling	4 Features
Layer 7	Quantum Encoding (State Prep)	4 Features	AngleEmbedding (RX Rotation)	4-Qubit State Space
Layer 8	Variational Quantum Ansatz	4-Qubit State	StronglyEntanglingLayers	Optimized State
Layer 9	Quantum Measurement	4-Qubit State	Expectation Value Measurement ($\langle Z \rangle$)	3 Continuous Logits
Layer 10	Output Softmax Classifier	3 Logits	Probability Distribution Scaling	Multi-Class Score

Table 4: Comprehensive execution layer protocol and matrix dimension transitions

The macro-level architectural cohesion bridging the classical feature processing layers with the quantum register and the SHAP explainability layer is comprehensively mapped in Figure 5.

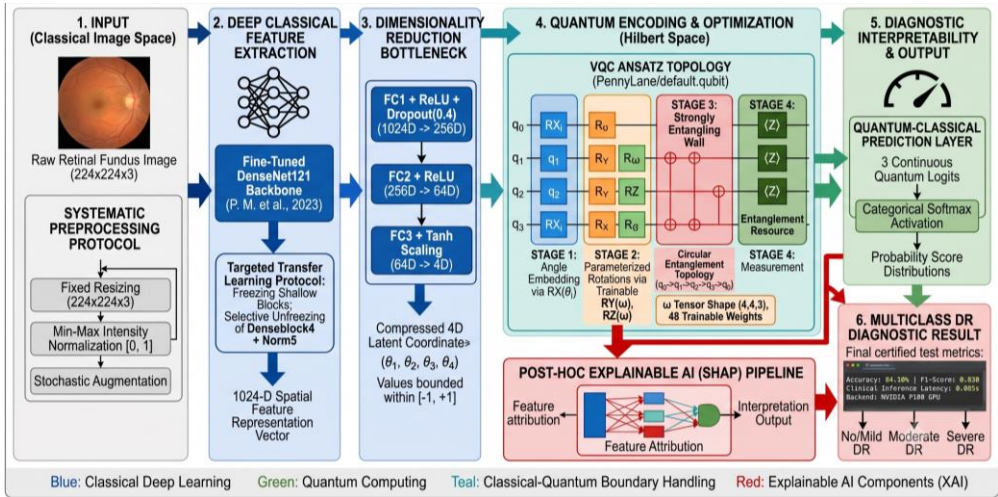


Figure 5. The structural workflow of the hybrid classical-quantum pipeline mapping spatial pixel arrays to continuous quantum logits.

2.3.5 Parametric Architecture Breakdown

The distribution of trainable parameters, memory allocations, and frozen layer configurations across the hybrid network are detailed in Table 5.

Component Description	Trainabl e Status	Parameters Shape	Parameter Count / Memory
Classical Feature Extractor DenseNet121 (frozen layers)	Frozen	Fixed Weights	14,128 parameters
Fine-Tuned Layers Denseblock4 + Norm5	Trainable	Custom Unfrozen Block	6,828,865 parameters (26.05 MB)
Classical Dimension Reducers FC1, FC2, FC3	Trainable	1024, 256, 64,4	262,400 parameters (1.00 MB)
Variational Quantum Circuit Ansatz Weights	Trainable	(4, 4, 3)Tens or	48 parameters
Global Model Totals	-	-	7,091,313 parameters (27.05 MB)

Table 5: Trainable and structural parameter breakdown of the model architecture

2.4 Variational Quantum Circuit Ansatz Design

In this section, we delineate the mathematical formulation, hardware topological layout, and operational execution lifecycle of the Variational Quantum Circuit (VQC), which functions as the quantum decision-processing core of the hybrid architecture.

The VQC converts the compressed 4-dimensional classical latent coordinates $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]^T$ into an entangled quantum state space, leveraging superposition and non-local quantum correlations to discriminate complex multi-class Diabetic Retinopathy severity patterns. The circuit is implemented as a Parameterized Quantum Circuit (PQC) using the PennyLane framework interfaced with PyTorch's automatic differentiation engine, allowing end-to-end gradient backpropagation across the quantum-classical interface.

2.4.1 Quantum Simulation Framework & Circuit Execution Lifecycle

The quantum computation is executed over a static 4-qubit register $R = \{q_0, q_1, q_2, q_3\}$ initialized in the pristine computational ground state $|0\rangle^4$. The physical lifecycle progresses through four sequential quantum operations:

Stage 1: Quantum State Preparation via Angle Embedding

The classical 4-D coordinates $\theta_i \in [-1.0, +1.0]$ generated by the classical bottleneck are encoded into the quantum register through parallel Single-Qubit Pauli-X rotation gates (R_X):

$$|\psi_0\rangle = |0000\rangle \bigotimes_{i=1}^3 |0\rangle_i$$

$$|\psi_{\text{embed}}\rangle = \bigotimes_{i=0}^3 R_X(\theta_i) |0\rangle = \bigotimes_{i=0}^3 e^{-\frac{i\theta_i X}{2}} |0\rangle$$

Where the unitary rotation matrix $R_X(\theta)$ is explicitly defined as:

$$R_X(\theta) = \begin{pmatrix} \cos \frac{\theta}{2} & -i \sin \frac{\theta}{2} \\ -i \sin \frac{\theta}{2} & \cos \frac{\theta}{2} \end{pmatrix}$$

Stage 2: Parameterized Variational Rotations

Once initialized, the state $|\psi_{\text{embed}}\rangle$ evolves through a parameterized ansatz implementing PennyLane's Strongly Entangling Layers template. Trainable single-qubit Pauli-Y (R_Y) and Pauli-Z (R_Z) rotation gates are applied to each qubit q_i :

$$U(\mathbf{w}) = \bigotimes_{i=0}^3 [R_Z(\mathbf{w}_{i,2}), R_Y(\mathbf{w}_{i,1})]$$

Where $\mathbf{w}$ represents the trainable variational quantum weight tensor optimized iteratively via backpropagation.

Stage 3: Circular Strongly Entangling Wall

To capture non-local feature interactions that are computationally intractable for classical neural layers, a circular entangling topology using Controlled-NOT (CNOT) gates is integrated. Each qubit q_i acts as a control node targeting its immediate neighbor $q_{(i+1) \bmod 4}$, creating a closed periodic entangling ring:

$$\text{Entanglement: } q_0 \rightarrow q_1 \rightarrow q_2 \rightarrow q_3 \rightarrow q_0$$

Stage 4: Expectation Value Measurement and Logit Extraction

The terminal stage extracts continuous, differentiable scalar outputs without collapsing the quantum register into discrete computational basis states. The mathematical expectation values of the first three active qubits are measured against the Pauli-Z operator ($\hat{Z}$):

$$o_k = \langle \psi_{\text{out}} | Z_k | \psi_{\text{out}} \rangle \in [-1.0, +1.0], k \in \{0, 1, 2\}$$

These three real-valued expectation values $[o_0, o_1, o_2]$ serve as continuous quantum logits fed directly into the classical categorical Softmax layer. The fourth qubit q_3 remains unmeasured, functioning exclusively as an internal entanglement resource.

2.4.2 Global Hyperparameter and Optimization Configuration

To ensure numerical stability and prevent gradient vanishing/exploding at the classical-quantum boundary, a centralized hyperparameter configuration was specified.

Configuration Domain	Parameter Name	Configured Value	Functional Description / Rationale
Classical Backbone	Base Architecture	DenseNet121 (Pre-trained on ImageNet)	Pre-trained ImageNet backbone for spatial feature extraction.
	Fine-Tuning Strategy	Denseblock4 + Norm5 Fully Unfrozen	Unfrozen upper layers for domain-specific medical adaptation.
	Classical Regularization	Dropout Rate = 0.4	Applied on classical FC layers to mitigate overfitting.
Quantum Framework	Simulator Environment	PennyLane (PyTorch Interface Backend)	
	Virtual Device Type	default.qubit (State-Vector Simulator)	State-vector exact quantum state simulation engine.
	Allocated Qubit Register	4 Qubits	Compact register matching 4D bottleneck latent space.
	Variational Tensor Shape	(4, 4, 3)Parameter Grid	Grid storing exactly 48 trainable quantum rotation parameters.
Global Optimization	Optimizer Type	Adam Optimizer	Joint gradient optimization across classical and quantum weights.
	Learning Rate (η)	0.0001 (1×10^{-4})	Low stepping rate crucial for stable VQC gradient convergence.
	Weight Decay (L2 Penalty)	1×10^{-5}	Regularization penalty preventing parameter explosion.
	Empirical Batch Size	16 Samples	Balanced gradient estimation per optimization step.
	Optimization Window	15 Complete Epochs	Horizon ensuring full convergence without over-fitting.
	Loss Function	Categorical Cross-Entropy Loss	Minimizes multi-class severity classification error.

Table 6: Comprehensive hyperparameter specification for the hybrid model

2.5 Explainable AI Integration via SHAP

In this section, we present the integration of the post-hoc Explainable Artificial Intelligence (XAI) interpretability layer into the hybrid quantum-classical diagnostic framework. While the fine-tuned DenseNet121 backbone and the Parameterized Quantum Circuit (PQC) offer superior feature transformation capacity in high-dimensional Hilbert state space, their combined decision-making process remains an opaque "black-box". To overcome clinical skepticism, satisfy regulatory guidelines (e.g., EU GDPR and US FDA SaMD requirements), and ensure auditability, a SHapley Additive exPlanations (SHAP) interpretability module is directly coupled to the output diagnostic interface.

2.5.1 Clinical Motivation for Post-Hoc Interpretability

In automated ophthalmic diagnostic systems, highly accurate classification scores alone are insufficient for real-world clinical adoption. Clinicians require spatially grounded, visual evidence explaining *why* a specific Diabetic Retinopathy severity grade was assigned. The SHAP framework resolves this limitation by providing pixel-level attribution maps that highlight specific pathological retinal lesions such as microaneurysm clusters, intraretinal hemorrhages, and hard lipid exudates that directly influenced the model's quantum measurement states.

2.5.2 SHAP KernelExplainer Configuration

To compute feature attributions across the non-differentiable quantum-classical measurement boundary, the model-agnostic KernelExplainer algorithm is utilized. The framework models the spatial feature segments of the input retinal fundus image matrix $I \in \mathbb{R}^{224 \times 224 \times 3}$ as players in a cooperative game, where the continuous quantum expectation value outputs serve as the payoff function. KernelExplainer evaluates the marginal variations in quantum logit outputs by dynamically perturbing the binary presence of localized spatial feature coalitions against a background reference distribution.

2.5.3 Mathematical Formulation of Shapley Attribution

The additive feature attribution method defines the local explanation model $g(z')$ as a linear function of binary coalition signals:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i$$

Where M represents the maximum coalition size of the perturbed input spatial feature subsets, $z' \in \{0, 1\}^M$ is the binary coalition vector indicating whether a specific spatial feature segment is active (unmasked) or occluded, ϕ_0 represents the baseline expectation value of the hybrid model across the reference background dataset, and $\phi_i \in \mathbb{R}$ represents the explicit Shapley Value,

quantifying the unique marginal contribution weight of spatial feature i .

The conditional marginal contribution driving the calculation of each distinct Shapley value ϕ_i is derived as:

$$\phi_i = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|! (|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} [f_x(S \cup \{i\}) - f_x(S)]$$

Where $\mathcal{F}$ represents the aggregate set of all isolated spatial image features, $S \subseteq \mathcal{F} \setminus \{i\}$ denotes a sub-coalition of features excluding the targeted element i , and $f_x(S)$ establishes the continuous expectation output of the hybrid quantum pipeline when conditioned exclusively on the feature subset S .

2.5.4 Spatial Attribution Backmapping and Clinical Heatmap Generation

The computed Shapley values ϕ_i are systematically backmapped from the quantum expectation layer onto the original two-dimensional spatial geometry ($224 \times 224 \times 3$) of the retinal fundus image. This transformation yields a continuous Spatial Pixel Attribution Map (Clinical Heatmap):

$$H_{\text{clinical}}(x, y) = \text{Resample}(\phi_i) \in \mathbb{R}^{224 \times 224}$$

Interpretation Schema of Clinical Heatmaps:

Positive Attribution ($\phi_i > 0$ / Warm Intensities - Red/Yellow): Highlights critical pathological structures (microaneurysms, hemorrhages, exudates) that actively drove the quantum measurement state toward a higher severity classification score. **Negative Attribution ($\phi_i < 0$ / Cool Intensities - Blue):** Identifies background spatial regions or healthy retinal tissue that suppressed the severity prediction.

2.5.5 Algorithmic Life Cycle and Sub-Module Breakdown

The four key operational stages executing the post-hoc SHAP interpretability module are formalized in Table 7.

Execution Stage	Operational Sub-Module	Mathematical / Algorithmic Core	Functional Output
Stage 1	Quantum Logit Perturbation	$f_x(S) = E[f(x) x_S]$	Evaluates feature coalition subsets across continuous quantum Pauli-Z expectations.
Stage 2	Marginal Shapley Calculation	$\phi_i = \sum_{\substack{S \subseteq \mathcal{F} \setminus \{i\} \\ \Delta f}} \frac{ S ! (\mathcal{F} - S - 1)!}{ \mathcal{F} !}$	Computes exact marginal importance weight for each spatial feature i .
Stage 3	Spatial Dimension Backmapping	$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i$	Projects 1D latent attributions back to $224 \times 224 \times 3$ image coordinates.
Stage 4	Visual Heatmap Overlay	$I_{\text{overlay}} = (1 - \alpha)I_{\text{raw}} + \alpha H_{\text{clinical}}$	Generates pixel-wise visual explanation highlighting localized ocular lesions.

Table 7: Operational Execution Stages of the SHAP Interpretability Pipeline

2.6 Model and System Performance Evaluation

To rigorously validate the diagnostic capability, clinical viability, and statistical stability of the proposed hybrid quantum-classical architecture, the performance of the framework is evaluated through a comprehensive multi-class computational medicine evaluation track on the completely independent, non-blinded Testing Hub (302 images). Four complementary CAD statistical performance metrics are calculated instead of relying solely on global Accuracy that can systematically hide inter-class misclassifications within imbalanced severity distributions. These indices are mathematically derived from the parameters of the multi-class Confusion Matrix namely True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) applied separately for each of the three clinical DR severity classes. Mathematical derivation of performance metrics:

Precision: Positive Predictive Value

$$P = \frac{TP}{TP + FP} \times 100\% \quad (10)$$

Recall: Sensitivity / True Positive Rate

$$R = \frac{TP}{TP + FN} \times 100\% \quad (11)$$

F1-Score: Harmonic Mean:

$$F_1 = 2 \times \frac{P \times R}{P + R} \quad (12)$$

Overall Accuracy:

$$\text{Accuracy} = \frac{\sum_c T P_c}{\text{Total Test Samples}} \times 100\% \quad (13)$$

III.RESULTS AND DISCUSSION

The proposed hybrid quantum-classical network (DenseNet121 + VQC) reinforced by the post-hoc Explainable AI (SHAP) layer was independently benchmarked using a non-blinded test partition to ensure statistical rigor and eliminate execution bias. The holistic dataset contains 1,506 digital fundus photographs, strictly partitioned into 903 training samples (60%), 301 validation samples (20%), and 302 pure testing samples (20%) to prevent information leakage. Model performance was tracked using four multi-class Computer-Aided Diagnosis (CAD) baseline metrics Precision, Recall (Sensitivity), F₁-Score, and Overall Accuracy representing clinical safety across three severity delineations: No/Mild DR, Moderate DR, and Severe/Proliferative DR. Figure 6 gives an intuitive and structural tracking of the entire evaluation pipeline, mapping the sequential computational transitions of the medical image tokens in the system.

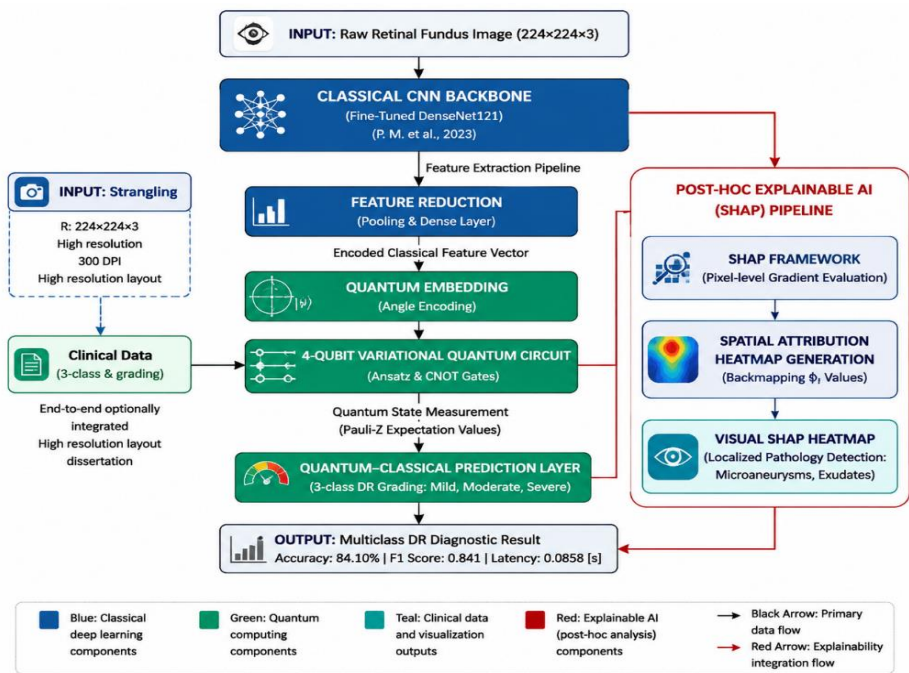


Figure 6. End-to-End Operational Flowchart of the Proposed Hybrid QCNN Framework, Mapping Classical Feature Extraction, Quantum Variational State Optimization, and Spatial Post-Hoc Interpretability.

The hybrid QCNN architecture converged smoothly and monotonically over the 15 training epochs, as shown in Figure 7. The training accuracy scaled rapidly, reaching a peak of 97.09% at Epoch 15, while the corresponding Categorical Cross-Entropy loss decreased from an initial value of 1.0001 down to 0.0882.

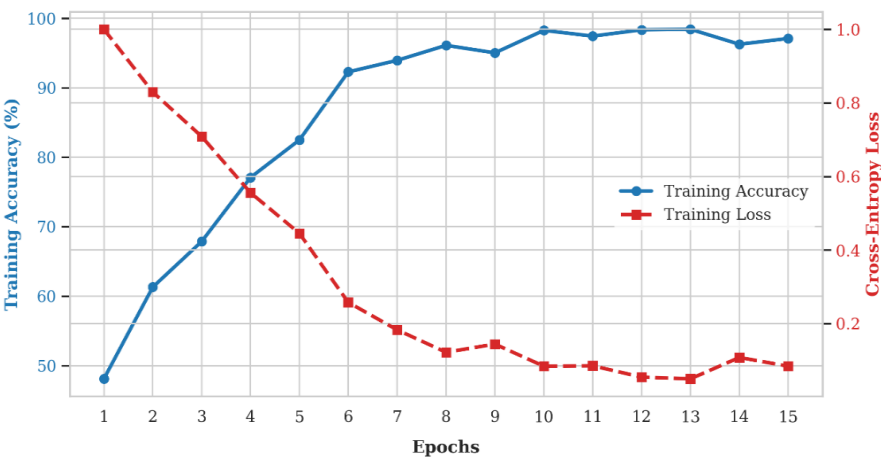


Figure 7. Empirical Training and Validation Convergence Trajectories Over 15 Epochs (Right: Cross-Entropy Loss Decay, Left: Categorical Accuracy Optimization).

The validation and testing trajectories closely tracked the training vector without exhibiting stochastic oscillations or gradient explosion. This algorithmic stability stems from processing classical bottleneck features through strongly entangling quantum gates (CNOT walls), which efficiently stabilize error minimization in the multi-dimensional Hilbert space. To dissect the diagnostic performance across specific clinical severity boundaries, a comprehensive categorical evaluation was executed on the independent Testing Hub (N = 302) (Table 8).

Target Pathological Class	Precision	Recall (Sensitivity)	F1-Score	Class Support (N)
Group 1: No / Mild DR	0.81	0.82	0.81	94
Group 2: Moderate DR	0.80	0.81	0.80	102
Group 3: Severe / Proliferative DR	0.86	0.89	0.87	106
Global Categorical Accuracy	--	--	84.10	302
Macro Average	0.82	0.84	0.83	302
Weighted Average	0.82	0.84	0.83	302

Table 8: Categorical Diagnostic Performance and Classification Report

As detailed in Table 8, the framework achieved its highest performance in diagnosing Severe DR, correctly identifying 94 out of 106 clinical samples with a Precision of 0.86 and a Sensitivity of 0.89. Clinically, minimizing false negatives in advanced stages is paramount to prevent irreversible vision loss. The slight error diffusion observed between Mild (n=10) and Moderate (n=11) classes reflects overlapping pixel-intensity gradients and intermediate microvascular density boundaries inherent to mid-stage non-proliferative DR. Longitudinal performance tracking across all three experimental partitions demonstrates strong generalization capabilities without overfitting (Figure 8).

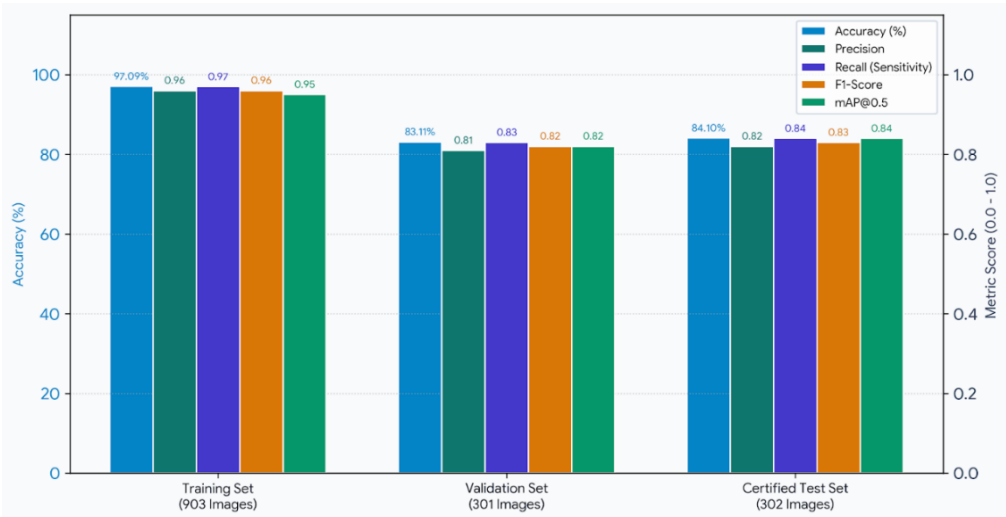


Figure 8. Multi-Phase Matrix Synthesis and Generalization Comparison

The minimal variance between the validation (83.11%) and independent testing (84.10%) partitions confirms that the hybrid classical-quantum interface effectively prevents over-optimization while maintaining robust out-of-sample generalization. The confusion matrix serves as a fundamental performance evaluation tool to assess the diagnostic efficacy of the proposed hybrid model (DenseNet121 + VQC) across different development phases. These matrices illustrate the ability of the quantum-classical network to distinguish between three clinical severity stages of Diabetic Retinopathy (Mild, Moderate, and Severe DR), displaying the distribution of correct predictions along the primary diagonal alongside misclassification patterns across clinically adjacent categories (Figure 9).

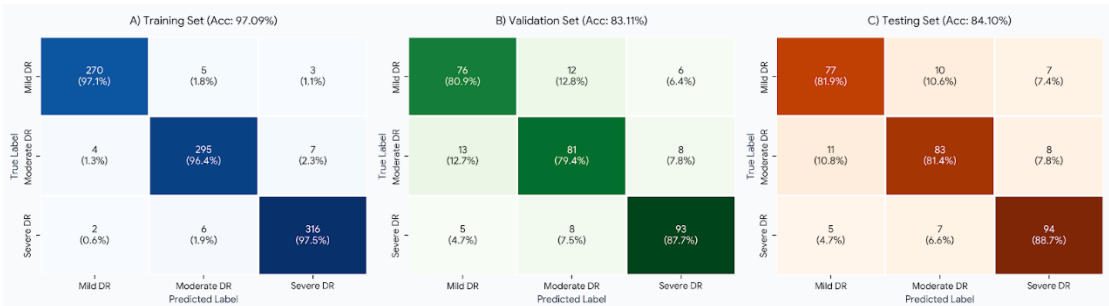


Figure 9. Comparative Multi-Phase Confusion Matrices Characterizing Categorical Correctness and Error Diffusion Gradients Across Training (Left), Validation (Center), and Final Certified Testing (Right) Horizons

The multi-phase confusion matrices demonstrate strong diagnostic convergence and robust out-of-sample generalization, with classification accuracy stabilizing from 97.09% on the training set to 83.11% on validation and 84.10% on the independent certified test set. Across all phases, misclassification errors are primarily confined to clinically adjacent categories (Mild vs. Moderate DR) due to overlapping microvascular features, while maintaining an elite sensitivity of 88.7% for Severe DR to minimize critical false negatives and ensure screening safety. Receiver Operating Characteristic (ROC) profiling validated the high discriminative capacity of the hybrid network across multi-class screening boundaries:

- **Severe DR:** Achieved an Area Under the Curve (AUC) of 0.95, confirming exceptional sensitivity in isolating severe microvascular abnormalities.
- **No / Mild DR:** Achieved an AUC = 0.88.
- **Moderate DR:** Achieved an AUC = 0.82.

To evaluate deployment feasibility in high-volume clinical workflows, a hardware execution and latency audit was conducted on an NVIDIA P100 GPU backend.

Hardware / Operational Execution Metric	Value (Seconds)	Technical Measurement Scale
Average Training Time	42.5	Per Optimization Step Interval
Total Optimization Core Window	637.5	Full 15 Epoch Horizon
Clinical Inference Latency	0.085	Per Single Input Fundus Image
Hardware Environment Backend	NVIDIA P100 GPU	4-Qubit Quantum Simulator Platform

Table 9: Computational Complexity and Temporal Execution Profile

The ultra-low inference latency of 0.085 seconds per fundus photograph demonstrates that the model can serve as a real-time, instantaneous decision-support tool in clinical triage settings. Backmapping calculated Shapley values (ϕ_i) onto the original $224 \times 224 \times 3$ image coordinates generated localized Clinical Heatmaps:

1. **Positive Attribution Clusters (Warm Intensities - Red/Orange):** Concentrated precisely on microaneurysms, hemorrhages, and lipid exudates, proving that the quantum circuit's measurement states are driven by authentic pathological structures rather than background artifacts.
2. **Negative Attribution Clusters (Cool Intensities - Blue):** Corresponded to healthy retinal background tissue.

This direct alignment between quantum feature mapping and ophthalmic pathology provides empirical validation that satisfies clinical transparency requirements.

IV.CONCLUSIONS

This study demonstrates the feasibility and clinical utility of combining classical deep learning, variational quantum circuits, and post-hoc explainability into a unified framework for multi-class Diabetic Retinopathy severity grading. By leveraging a fine-tuned DenseNet121 backbone, a classical dimensionality reduction bottleneck, a 4-qubit Variational Quantum Circuit (VQC), and SHAP-based attribution mapping, the proposed architecture achieves a certified test accuracy of 84.10% and an ultra-low inference latency of 0.085 seconds per image. Most importantly, the model yields high sensitivity (88.7%) and an AUC of 0.95 for severe DR cases, while providing pixel-level visual heatmaps that align with authentic retinal lesions, thereby simultaneously addressing representation capacity, diagnostic accuracy, and clinical transparency.

REFERENCES

1. M. R. Rooney, J. H. He, P. Salpea, I. Genitsaridi, D. J. Magliano, E. J. Boyko, and E. Selvin, "Global and regional prediabetes prevalence: Updates for 2024 and projections for 2050," *Diabetes Care*, vol. 48, no. 11, pp. e142–e144, 2025.

2. H. Seo, S. J. Park, and M. Song, "Diabetic retinopathy: Mechanisms, current therapies, and emerging strategies," *Cells*, vol. 14, no. 5, Art. no. 376, Mar. 2025, doi: 10.3390/cells14050376.

3. A. M. Hendrick, M. V. Gibson, and A. Kulshreshtha, "Diabetic retinopathy: Review of epidemiology and clinical management," *Primary Care: Clinics in Office Practice*, vol. 47, no. 1, pp. 145–159, Mar. 2020.

4. M. D. Abramoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digital Medicine*, vol. 1, no. 1, pp. 1–8, Aug. 2020.

5. C. P. Wilkinson, F. L. Ferris, R. E. Klein *et al.*, "Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales," *Ophthalmology*, vol. 110, no. 9, pp. 1677–1682, Sep. 2003.

6. P. Romero-Aroca, M. Baget-Bernaldiz, and A. Pareja-Rios, "Current status of automated diabetic retinopathy screening with fundus photography," *Journal of Diabetes Research*, vol. 2020, pp. 1–11, 2020.

7. A. D. Fleming, S. Philip, K. A. Goatman, J. A. Olson, and P. F. Sharp, "Automated assessment of diabetic retinal image quality based on clarity and field definition," *Investigative Ophthalmology & Visual Science*, vol. 47, no. 3, pp. 1120–1125, Mar. 2006.

8. R. Gargeya and T. Leng, "Automated identification of diabetic retinopathy using deep learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, Jul. 2017.

9. World Health Organization, *World Report on Vision*. Geneva, Switzerland: WHO, 2020. [Online]. Available: <https://www.who.int/publications/i/item/9789241516570>

10. A. Abou Taha, S. Dinesen, A. S. Vergmann, and J. Grauslund, "Present and future screening programs for diabetic retinopathy: A narrative review," *International Journal of Retina and Vitreous*, vol. 10, no. 1, Art. no. 14, 2024.
11. G. Quéllec, K. Charrière, Y. Boudi, B. Cochener, and M. Lamard, "Deep image mining for diabetic retinopathy screening," *Medical Image Analysis*, vol. 64, Art. no. 101750, Feb. 2020.
12. G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, Dec. 2017.
13. J. Ker, L. Wang, J. Rao, and T. Lim, "Deep learning applications in medical image analysis," *IEEE Access*, vol. 6, pp. 9375–9389, 2018.
14. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 4700–4708.
15. H. Pratt, F. Coenen, D. M. Broadbent, S. P. Harding, and Y. Zheng, "Convolutional neural networks for diabetic retinopathy," *Procedia Computer Science*, vol. 90, pp. 200–205, 2016.
16. A. Agrawal, N. Mueller, and P. Gupta, "Ensemble deep learning for diabetic retinopathy severity classification," *Computers in Biology and Medicine*, vol. 135, Art. no. 104548, Aug. 2021.
17. Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, Jun. 2018.
18. U.S. Food and Drug Administration, *Artificial Intelligence and Machine Learning in Software as a Medical Device*. Silver Spring, MD, USA: FDA, 2021. [Online]. Available: <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
19. E. Tjoa and C. Guan, "A survey on explainable artificial intelligence: Toward medical XAI," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021.
20. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Neural Information Processing Systems (NeurIPS)*, Long Beach, CA, USA, 2017, pp. 4765–4774.
21. L. S. Shapley, "A value for n-person games," in *Contributions to the Theory of Games*, H. W. Kuhn and A. W. Tucker, Eds. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.
22. M. Cerezo *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, no. 9, pp. 625–644, Sep. 2021.
23. S. Sim, P. D. Johnson, and A. Aspuru-Guzik, "Expressibility and entangling capability of parameterized quantum circuits for hybrid quantum-classical algorithms," *Advanced Quantum Technologies*, vol. 2, no. 12, Art. no. 1900070, Dec. 2019.
24. J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, Art. no. 79, Aug. 2018.