Europe at a Crossroads: Governance, Business, and Policy Transformation in a Changing Global Economy. Proceedings of the International Conference EU-PAIR 2026

DOI:10.47743/eu-pair.2026-1-6

EXPERIMENTAL EVALUATION OF HYBRID QUANTUM-CLASSICAL LEARNING FOR DERMOSCOPIC SKIN LESION CLASSIFICATION

MARIA CIORÂȚANU
Alexandru Ioan Cuza University of Iași
Iași, Romania
mariacioratanu49@gmail.com

AIDA-GABRIELA PÂNCĂ
Alexandru Ioan Cuza University of Iași
Iași, Romania
aida.pnc9@gmail.com

ȘTEFAN-RĂZVAN BĂLĂUCĂ
Alexandru Ioan Cuza University of Iași
Iași, Romania
stefann.balauca@gmail.com

ADRIAN IFTENE
Alexandru Ioan Cuza University of Iași
Iași, Romania
adiftene@info.uaic.ro

Abstract
The growing use of artificial intelligence in medical image analysis raises important questions about reliability, transparency, and the responsible adoption of emerging computational technologies. This paper presents an experimental comparison between a classical convolutional neural network and a hybrid quantum-classical learning pipeline for binary dermoscopic skin lesion classification. The study uses the publicly available HAM10000 dataset and focuses on the distinction between benign and malignant skin lesions. The classical model is based on MobileNetV2, while the hybrid approach combines convolutional feature extraction, principal component analysis, and a variational quantum circuit used as a compact decision module. Both models are evaluated under the same train, validation, and test protocol, and the quantum pipeline is additionally validated for inference on IBM Quantum backends. The results show that the classical baseline clearly outperforms the hybrid model, reaching 95.46% test

accuracy, while the selected hybrid configuration obtains a balanced accuracy of 0.582 and an AUC of 0.610 on the held-out test set. These findings indicate that current hybrid quantum-classical models should not yet be viewed as competitive clinical decision-support tools, but rather as feasibility studies under noisy intermediate-scale quantum constraints. The paper contributes to the discussion on trustworthy and responsible AI by showing the importance of controlled evaluation, transparent limitations, and careful interpretation before adopting emerging AI technologies in sensitive domains such as healthcare.
Keywords: *responsible AI; medical image classification; hybrid quantum-classical learning; dermoscopy; trustworthy evaluation.*
JEL Classification: C63; I18; O33.

1. INTRODUCTION

Artificial intelligence is increasingly used in medical image analysis, where automated systems can support faster screening, provide an additional signal for clinicians, and assist educational or research workflows. In dermoscopy, the task remains difficult because benign and malignant skin lesions may share similar visual characteristics, while lesions belonging to the same class may also vary substantially in color, texture, shape, and border appearance. These characteristics make binary benign versus malignant classification a relevant and challenging case for studying the reliability of automated image analysis systems (Esteva et al., 2017; Tschandl, Rosendahl and Kittler, 2018).

At the same time, quantum machine learning has attracted attention as an emerging computational direction. However, current quantum hardware is still limited by a small number of noisy qubits, finite-shot measurement, restricted connectivity, and device-dependent execution constraints (Preskill, 2018; Cerezo et al., 2021). For realistic medical images, fully quantum image processing is therefore not yet a practical option. Most present studies rely on hybrid quantum-classical designs, in which a classical model extracts image features and a variational quantum circuit is used as a compact decision layer (Mari et al., 2020; Benedetti et al., 2019).

This paper studies such a hybrid design in the context of dermoscopic skin lesion classification. The goal is not to claim that the quantum model is superior to established deep learning approaches. Instead, the study asks how a hybrid quantum-classical decision module behaves when it is compared with a strong classical convolutional baseline under the same task definition, data splits, and evaluation protocol.

The broader relevance of the work lies in the governance of emerging AI systems. In high-stakes domains such as healthcare, novelty alone is not sufficient. A model must be evaluated transparently, compared against strong baselines, and presented with clear limitations before its potential role can be interpreted responsibly.

The main contributions of the paper are threefold. First, it presents a controlled comparison between a classical MobileNetV2 classifier and a hybrid quantum-classical pipeline on the HAM10000 dermoscopic dataset. Second, it evaluates the hybrid model using validation-based model selection and held-out test performance, including accuracy, balanced accuracy, F1 score, AUC, and confusion-matrix analysis. Third, it frames the results from a trustworthy AI perspective, showing that the hybrid pipeline is technically feasible and executable on IBM Quantum backends, but not competitive with the classical baseline for practical classification.

2. RESPONSIBLE FRAMING AND RELATED WORK

Deep convolutional neural networks are well-established tools for image classification and have been widely used in medical image analysis, including dermatologist-level skin cancer classification studies (Esteva et al., 2017). In skin lesion assessment, transfer learning is particularly useful because models trained on large-scale image datasets can be adapted to more specific medical tasks. MobileNetV2 is an efficient convolutional architecture that provides a strong trade-off between representational capacity and computational cost (Sandler et al., 2018). For this reason, it is suitable both as a classical baseline and as a feature extractor for hybrid experiments.

The HAM10000 dataset is a commonly used dermoscopic benchmark containing multiple diagnostic categories (Tschandl, Rosendahl and Kittler, 2018). In this study, these categories are grouped into two classes, benign and malignant, in order to create a binary classification task. This formulation is useful for experimental comparison, but it should not be confused with clinical diagnosis. A real clinical system would require broader validation, physician oversight, prospective testing, and careful assessment of safety, bias, and reliability.

Quantum machine learning explores whether quantum circuits can support learning tasks through parameterized transformations and measurement-based outputs (Biamonte et al., 2017; Benedetti et al., 2019). In the current noisy intermediate-scale quantum era, variational quantum algorithms are among the most practical approaches because they combine trainable quantum circuits with classical optimization loops (Cerezo et al., 2021; Preskill, 2018). Hybrid transfer-learning approaches are especially relevant for images because the high-dimensional visual representation can be handled by a classical model, while the quantum component operates on a reduced feature vector (Mari et al., 2020).

Recent studies have explored hybrid quantum-classical architectures for medical and skin lesion classification tasks, often combining classical feature extraction with variational quantum circuits or related quantum components (Hussein, Montaser and Elsayed, 2025; Frasca et al., 2025; Reka et al., 2024).

These works show that the area is active, but they also reinforce the need for careful baselines. A hybrid model should not be evaluated only by showing that it can run. It must also be compared with mature classical alternatives under a controlled protocol.

From a responsible AI perspective, this distinction is central. In medical imaging, a weak but novel model can create misleading expectations if its limitations are not clearly stated. The present paper therefore treats the hybrid quantum-classical pipeline as an experimental benchmark and feasibility study. The contribution is not a claim of clinical readiness, but an evidence-based analysis of what the current pipeline can and cannot do.

3. METHODOLOGY

3.1. Task and Dataset

The task is binary dermoscopic image classification. Given an input image of a skin lesion, the model must assign the image to one of two classes, benign or malignant. The experiments use the HAM10000 dataset, a public collection of dermoscopic images covering seven diagnostic categories (Tschandl, Rosendahl and Kittler, 2018). For the binary setup, the original labels are aggregated into two super-classes. The binary grouping used in the experiments is summarized in (Table 1).

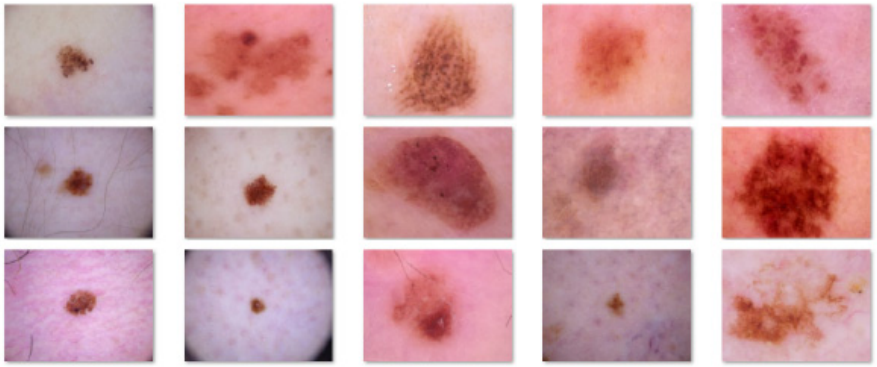
Table 1. Binary grouping of HAM10000 diagnostic categories

Binary class	HAM10000 categories used in the study
Malignant	Melanoma (mel), basal cell carcinoma (bcc), and actinic keratoses (akiec).
Benign	Melanocytic nevi (nv), benign keratosis-like lesions (bkl), dermatofibroma (df), and vascular lesions (vasc).

Source: authors' own processing based on Tschandl, Rosendahl and Kittler (2018)

Examples from the HAM10000 dataset are shown in (Figure 1).

All images are processed through the same input pipeline before classification. Images are resized to 224 x 224 pixels and normalized to the [0, 1] range. During training, lightweight augmentation through random rotations and flips is used to improve robustness to visual variability. Using the same preprocessing for both models is important because the comparison should reflect the behavior of the classifiers rather than differences in input preparation.



Source: authors' own processing of the HAM10000 dataset (Tschandl, Rosendahl and Kittler, 2018)

Figure 1. Representative dermoscopic examples from HAM10000

3.2. Classical MobileNetV2 Baseline

The classical baseline is a MobileNetV2-based convolutional neural network (Sandler et al., 2018). The model is initialized with ImageNet pretrained weights and adapted to the binary benign versus malignant classification task. For each input image, the model produces a probability score for the malignant class and a predicted label. This baseline has two roles in the study. First, it provides the main practical reference point because convolutional models remain strong solutions for dermoscopic image classification. Second, it provides the visual features used by the hybrid quantum-classical model.

3.3. Hybrid Quantum-Classical Pipeline

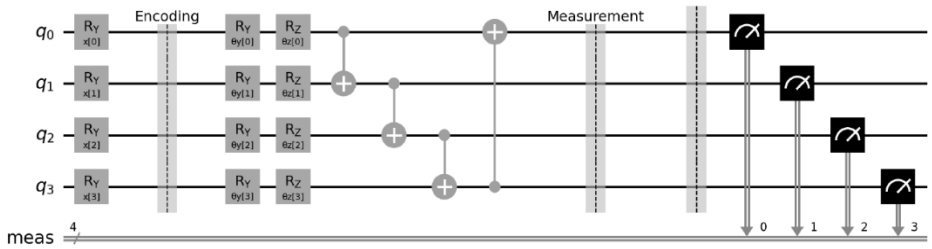
The hybrid model follows a late-fusion strategy. The image is first processed by the classical MobileNetV2 backbone, which extracts a high-dimensional feature vector. Because current quantum circuits cannot directly process such a large representation, the feature vector is standardized and reduced with principal component analysis. The reduced vector is then transformed so that it can be encoded into rotation angles.

$$z = \mathrm{PCA}(\mathrm{Scaler}(f(I))) \quad (1)$$

$$x = \tanh(z), \; x \in (-1, 1)^n \quad (2)$$

The quantum component is a variational quantum circuit implemented with Qiskit. The reduced features are encoded through angle-based rotations, trainable variational layers are applied, and entanglement is introduced through CNOT gates arranged according to the selected circuit configuration. Data re-uploading is used to increase *expressiveness by applying the input encoding*

more than once between variational layers (Perez-Salinas et al., 2020). Figure 2 provides a simplified four-qubit illustration of the variational quantum circuit, showing its main encoding, processing, and measurement stages. The diagram is intended as a conceptual representation rather than an exact depiction of the 12-qubit configuration used in the experiments.



Source: authors' own elaboration

Figure 2. Simplified four-qubit illustration of the variational quantum circuit

The final score is obtained from measurement statistics and mapped to a probability-like value. A threshold selected on the validation set is then used to assign the final binary label.

$$\text{pVQC} = (e + 1) / 2 \quad (3)$$

$$\hat{y} = \text{malignant if pVQC} \geq \tau, \text{otherwise benign} \quad (4)$$

This design is deliberately constrained. The quantum circuit does not replace the full image-processing pipeline. Instead, it acts as a compact decision module on top of classical CNN features. This distinction is important for interpreting the results because the experiment evaluates whether a quantum decision layer is useful after classical feature extraction, not whether a quantum computer can independently perform dermoscopic image recognition.

3.4. Real-Hardware Inference and Prototype Workflow

After simulator-based development, the selected hybrid pipeline is also validated for inference on IBM Quantum backends. For real-device execution, the circuit must be transpiled to hardware-supported operations and mapped to a valid backend layout. The decision flow remains unchanged because the classical backbone extracts features, PCA reduces them, the variational circuit processes the quantum-compatible representation, and the score is computed from shot-based measurements.

In addition to the model-level experiments, the project includes an end-to-end prototype. A Streamlit frontend allows the user to upload a dermoscopic image, while a FastAPI backend runs the classical CNN and the hybrid VQC model on the same input. The interface returns side-by-side predictions and scores. This

prototype supports demonstration, reproducible inference with saved artifacts, and inspection of cases in which the two models agree or diverge.

4. EXPERIMENTAL PROTOCOL

4.1. Evaluation Design

The evaluation uses disjoint train, validation, and test splits. Model development and selection are performed on training and validation data, while final performance is reported only on the held-out test set. This separation is especially important for the hybrid model because its threshold is selected on the validation set and then applied unchanged to the test set.

The comparison is controlled by using the same task definition, the same class grouping, the same dataset source, and the same evaluation logic. However, the comparison must still be interpreted carefully. The hybrid model reuses CNN-derived features and therefore should be understood as a quantum decision module built on top of a classical visual representation.

4.2. Metrics and Retained Configuration

The reported metrics are accuracy, balanced accuracy, F1 score, and area under the receiver operating characteristic curve. Accuracy summarizes overall correctness. Balanced accuracy is useful because it treats the two classes more evenly when class distribution matters. The F1 score captures the precision-recall trade-off, and AUC reflects ranking quality across thresholds. In addition, confusion matrices are used to inspect the distribution of false positives and false negatives.

The retained hybrid configuration uses 12 qubits, 4 variational repetitions, data re-uploading equal to 2, ring entanglement, and 192 trainable parameters. The circuit is evaluated with a shot-based estimator using 2048 shots. Parameters are optimized with SPSA for 500 iterations and 5 random restarts. The best retained restart achieved its best validation score at iteration 225. The chosen threshold is tau = 0.53, selected on the validation set using balanced accuracy as the target criterion.

5. RESULTS

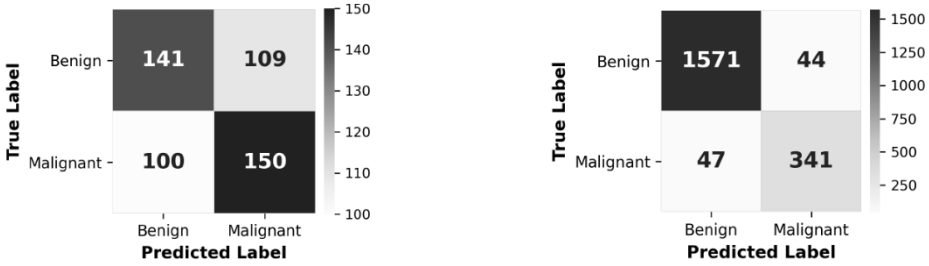
The results show a clear difference between the classical and hybrid approaches. The MobileNetV2 baseline reaches 95.46% test accuracy in the binary classification task. The hybrid VQC model performs above chance level, but remains substantially weaker. On the validation split, the retained hybrid configuration reaches accuracy = 0.64, balanced accuracy = 0.64, F1 = 0.65, and AUC = 0.62. On the held-out test split, the same configuration reaches accuracy = 0.582, balanced accuracy = 0.582, F1 = 0.589, and AUC = 0.610. The retained model results are summarized in (Table 2).

Table 2. Summary of retained model results

Model	Evaluation setting	Reported results	Interpretation
MobileNetV2 classical CNN	Held-out test split	95.46% test accuracy	Strong practical baseline for the binary dermoscopic task.
Hybrid VQC pipeline	Validation split	Acc = 0.64; BalAcc = 0.64; F1 = 0.65; AUC = 0.62	Best configuration selected before test evaluation.
Hybrid VQC pipeline	Held-out test split	Acc = 0.582; BalAcc = 0.582; F1 = 0.589; AUC = 0.610	Above chance, but far below the classical CNN.

Source: authors' own processing of experimental results

The confusion matrix comparison is shown in (Figure 3).



Source: authors' own processing of experimental results

Figure 3. Confusion matrix comparison on the HAM10000 test split

The test confusion matrix of the retained hybrid configuration is [[141, 109], [100, 150]], where rows correspond to true classes and columns correspond to predicted classes. This distribution shows that the hybrid model makes a substantial number of errors in both classes. The classical baseline has a much stronger diagonal concentration, reflecting better separation between benign and malignant samples.

The observed gap is consistent with the design constraints of the two models. The classical CNN operates in a representation space well suited to image classification and benefits from mature transfer learning. The hybrid model receives a compressed, quantum-compatible representation and optimizes a finite-shot stochastic objective. Compression is necessary because of the limited qubit budget, but it can remove information that is useful for lesion discrimination.

Repeated VQC training also shows variability. This behavior is influenced by finite-shot sampling noise, the stochastic nature of SPSA optimization, initialization, circuit depth, entanglement pattern, and data re-uploading. Some runs approach the best validation behavior, while others converge closer to chance level. This variability is not just an implementation detail. It is part of the practical behavior of current hybrid quantum-classical learning under noisy and low-dimensional constraints.

Validation on IBM Quantum hardware confirms that the trained circuit can be executed beyond simulation. This result should be interpreted as hardware compatibility and feasibility, not as evidence of a performance advantage. Queue-dependent latency, device noise, and backend constraints make real-hardware execution unsuitable as the main environment for iterative model development in this study.

6. TRUSTWORTHY AI, GOVERNANCE AND INTERPRETATION

The main idea of the study is not that the hybrid quantum-classical model should replace classical deep learning. The classical CNN remains clearly stronger. The interpretation is that the hybrid pipeline demonstrates feasibility, while also showing the limitations of current near-term quantum approaches in realistic medical image classification.

This conclusion is relevant for trustworthy AI because emerging technologies can easily be over-interpreted when evaluated without strong baselines. In a sensitive domain such as healthcare, a model should not be presented as useful just because it is novel or executable on advanced infrastructure. It should be assessed against practical alternatives and reported with limitations that are understandable to researchers, practitioners, and decision-makers.

The study supports three governance-oriented principles. First, controlled comparison is necessary. Both models must be evaluated on the same task, with the same data splits and clearly defined metrics. Second, limitations must be explicit. The hybrid model relies on CNN-derived features, dimensionality reduction, finite-shot measurement, and stochastic optimization. These constraints directly affect performance and stability. Third, deployment claims must be proportional to evidence. The current hybrid model is suitable as a research benchmark and educational prototype, not as a clinical decision-support tool. The responsible interpretation of the experimental findings is summarized in (Table 3).

Table 3. Responsible interpretation of the experimental findings

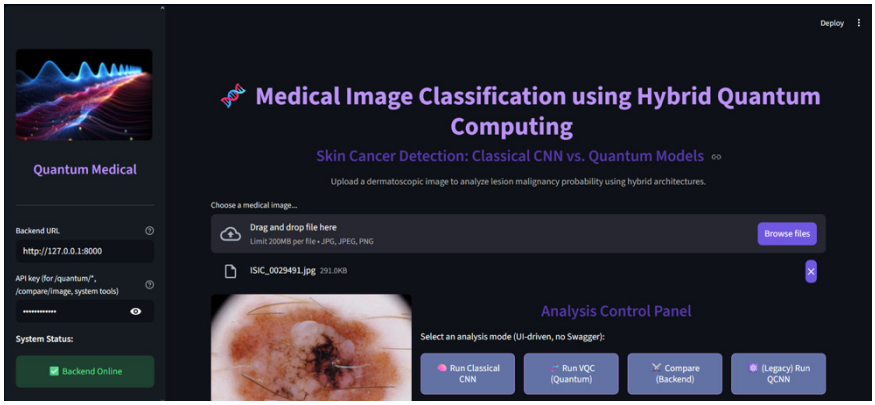
Aspect	Evidence from the study	Responsible interpretation
Predictive performance	The classical CNN reaches 95.46% test accuracy, while the hybrid model reaches 0.582 balanced accuracy and 0.610 AUC on the test set.	The hybrid model is not competitive with the classical baseline for practical classification.
Technical feasibility	The hybrid pipeline can be trained, selected on validation data, evaluated on test data, and executed on IBM Quantum backends.	The contribution is a feasibility study under current NISQ-era constraints.
Reliability	VQC training shows run-to-run variability due to finite-shot sampling and stochastic optimization.	Results must be reported with caution and should not be generalized beyond the experimental setting.
Healthcare relevance	The task concerns benign-versus-malignant dermoscopic classification, a high-stakes medical context.	The prototype is not a clinical tool and would require extensive validation before any medical use.
Governance relevance	The work compares a new AI direction with a mature baseline and documents limitations.	Transparent evaluation helps prevent exaggerated claims about emerging technologies.

Source: authors' own elaboration based on experimental findings

The prototype also has educational value. By allowing side-by-side comparison between the classical and hybrid predictions, it makes the evaluation process visible. Users can inspect cases of agreement and disagreement and understand why a weaker model should not be adopted simply because it belongs to a newer technological category.

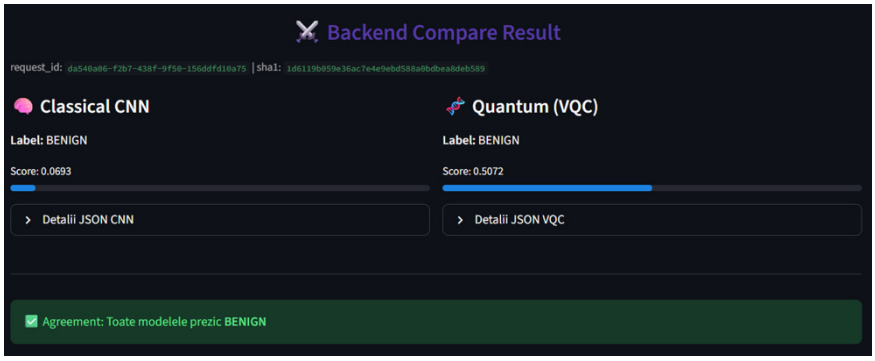
7. PROTOTYPE AND REPRODUCIBILITY SUPPORT

The software prototype was designed to support reproducible inference and transparent comparison. The backend exposes endpoints for single-model inference and comparative inference. The comparison workflow runs both models on the same uploaded image and returns structured outputs containing model labels, scores, agreement indicators, score differences, and selected comparison information. The upload interface is shown in Figure 4, while the comparison report is shown in Figure 5.



Source: authors' own screenshot of the developed prototype

Figure 4. Prototype interface for uploading a dermoscopic image



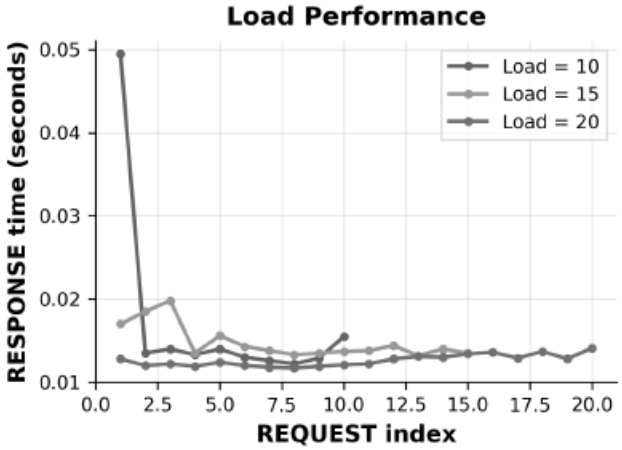
Source: authors' own screenshot of the developed prototype

Figure 5. Prototype interface showing the comparison report

The training process stores the preprocessing and model artifacts required for reproducible inference. For the hybrid model, these include the fitted scaler and PCA transformer, optimized circuit parameters, selected threshold, configuration metadata, and training reports. This artifact-based design allows the API to reload the same model state without retraining.

The system also includes operational support. Prediction caching is keyed by uploaded image hashes and can reduce response time for repeated submissions. Audit and monitoring utilities support inspection of API usage and model execution behavior. These elements improve the engineering transparency of the experimental platform.

The prototype was validated with 51 automated tests covering unit-level behavior, integration-level endpoints, repeated requests, and cache-related execution paths. The overall recorded code coverage was 73%, with 90% coverage for the VQC-related module and 100% coverage for the classical CNN module. Load testing indicated stable behavior for the tested repeated-request scenarios, with caching improving repeated submissions. The load-testing behavior of the deployed API is shown in Figure 6.



Source: authors' own processing of load-testing results

Figure 6. Load-testing performance of the deployed API

8. CONCLUSIONS

This paper presented an experimental evaluation of hybrid quantum-classical learning for binary dermoscopic skin lesion classification. The study compared a classical MobileNetV2 baseline with a hybrid pipeline that combines CNN feature extraction, PCA-based dimensionality reduction, and a variational quantum circuit. Both models were evaluated under the same train, validation, and test protocol on the HAM10000 dataset, and the selected hybrid pipeline was additionally validated for inference on IBM Quantum backends.

The results show that the classical CNN remains the stronger practical solution. It reaches 95.46% test accuracy, while the retained hybrid VQC configuration reaches 0.582 balanced accuracy and 0.610 AUC on the held-out test split. The hybrid model is therefore not competitive with the classical baseline for practical dermoscopic classification. Its value lies instead in demonstrating an executable research pipeline and in clarifying the present limitations of quantum-classical learning under NISQ-era constraints.

From a responsible AI perspective, the study reinforces the need for careful evaluation of emerging technologies. In healthcare-related tasks, claims should

be based on controlled comparison, transparent metrics, and explicit limitations. The hybrid quantum-classical model should be understood as a feasibility study and educational benchmark, not as a clinical decision-support system. Future work should investigate improved feature compression, more stable optimization, alternative circuit designs, noise-aware inference, and explainability mechanisms, while preserving the same discipline of comparison against strong classical baselines.

References

1) Benedetti, M., Lloyd, E., Sack, S. and Fiorentini, M. (2019). Parameterized quantum circuits as machine learning models. Quantum Science and Technology, 4(4), 043001.
2) Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. and Lloyd, S. (2017). Quantum machine learning. Nature, 549(7671), pp. 195-202.
3) Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L. and Coles, P.J. (2021). Variational quantum algorithms. Nature Reviews Physics, 3(9), pp. 625-644.
4) Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), pp. 115-118.
5) Frasca, M., Cutica, I., Pravettoni, G. and La Torre, D. (2025). Optimizing melanoma diagnosis: A hybrid deep learning and quantum computing approach for enhanced lesion classification. Intelligence-Based Medicine, 12, 100264.
6) Hussein, A.A., Montaser, A.M. and Elsayed, H.A. (2025). Skin cancer image classification using hybrid quantum deep learning model with BiLSTM and MobileNetV2. Quantum Machine Intelligence, 7(2), 66.
7) Mari, A., Bromley, T.R., Izaac, J., Schuld, M. and Killoran, N. (2020). Transfer learning in hybrid classical-quantum neural networks. Quantum, 4, 340.
8) Perez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E. and Latorre, J.I. (2020). Data re-uploading for a universal quantum classifier. Quantum, 4, 226.
9) Preskill, J. (2018). Quantum computing in the NISQ era and beyond. Quantum, 2, 79.
10) Reka, S.S., Karthikeyan, H.L., Shakil, A.J., Venugopal, P. and Muniraj, M. (2024). Exploring quantum machine learning for enhanced skin lesion classification: A comparative study of implementation methods. IEEE Access, 12, pp. 104568-104584.
11) Sandler, M., Howard, A., Zhu, M., Zhmoginov, A. and Chen, L.C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510-4520.
12) Tschandl, P., Rosendahl, C. and Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data, 5, 180161.