SAFE Quantum Machine Learning with Variational Quantum Classifiers

Anonymous Author(s)

Abstract

We propose a variational quantum classifier operating on high-dimensional deep representations via amplitude encoding, stabilized by a learnable classical pre-encoding layer. By combining normalized amplitude embeddings with bounded quantum observables, the resulting model induces a structured and smooth hypothesis class with controlled sensitivity to input variations. Model reliability is assessed using SAFE-AI metrics derived from the Cramér–von Mises divergence, enabling consistent evaluation across accuracy, robustness, and explainability dimensions. Empirical results show that the proposed quantum model provides comparable accuracy to strong classical baselines while exhibiting a more balanced SAFE reliability profile, with improved robustness to noise and stability under structured feature removal. These findings suggest that variational quantum circuits offer a principled mechanism for stability-oriented SAFE learning in safety-critical settings.

Keywords

Security, Accuracy, Explainability, Quantum AI

1 Introduction

Machine learning models are increasingly used in high-stakes domains such as medical imaging, finance, and public decision-making. In these settings, predictive accuracy alone is insufficient: models must also remain stable under perturbations, degrade gracefully when information is removed, and provide behavior that can be meaningfully inspected by human stakeholders. This motivates the Secure, Accurate, Fair, and Explainable AI (SAFE-AI) perspective [1], where multiple reliability dimensions are evaluated jointly rather than as independent post hoc criteria.

Deep learning models have achieved strong performance in high-dimensional image analysis, but their high flexibility can lead to sensitivity under noise, distributional shift, or structured input degradation. Quantum machine learning provides a complementary modelling paradigm. Variational quantum circuits operate through normalized state embeddings, unitary transformations, and bounded measurements, potentially inducing structured hypothesis classes with controlled sensitivity. However, most empirical quantum machine learning studies focus mainly on predictive accuracy, while robustness and explainability remain less systematically evaluated.

This work investigates whether a hybrid variational quantum classifier can support SAFE learning in high-stakes image classification. We use brain tumor MRI classification as a representative medical task and compare the proposed quantum model with classical baselines under a unified SAFE-AI evaluation protocol.

2 Method

The proposed framework combines classical deep feature extraction, a hybrid variational quantum classifier, and a unified SAFE-AI evaluation protocol. The images are first cropped to remove the non-informative background, resized to 224×224, normalized, and processed by a pretrained ResNet-18 whose final layer is replaced by an identity mapping. This produces a 512-dimensional feature vector for each image. Before quantum encoding, the feature vector is transformed by a learnable classical pre-VQC layer consisting of a linear mapping followed by a GELU activation. The transformed vector is then normalized and embedded into a 9-qubit quantum state through amplitude encoding [6],

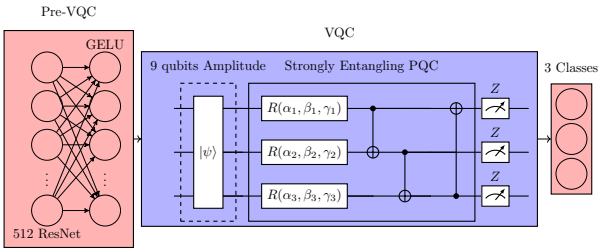

Figure 1: Hybrid quantum–classical architecture combining ResNet feature extraction, a classical pre-VQC projection layer, amplitude encoding into a variational quantum circuit, and a final linear classifier.

$$|\psi(x)\rangle = \sum_{i=0}^{511} x_i |i\rangle. \tag{1}$$

Figure 1 illustrates the overall hybrid quantum–classical architecture. The encoded quantum state is processed by a variational quantum circuit with strongly entangling layers. Expectation values of Pauli-Z operators are measured on each qubit to produce bounded quantum features, which are mapped to class logits through a final linear classifier. The full hybrid model is trained end-to-end using cross-entropy loss.

Reliability is evaluated using SAFE-AI metrics derived from the Rank Graduation (RG) framework, which is based on the Cramér–von Mises divergence between empirical distributions. Given two random variables Y and Y' with cumulative distribution functions F_Y and $F_{Y'}$, the p-th order Cramér–von Mises divergence is defined as

$$\mathrm{CvM}_p(F_Y, F_{Y'}) = \int_{-\infty}^{\infty} |F_Y(u) - F_{Y'}(u)|^p \, dF_Y(u). \tag{2}$$

Based on the first-order divergence, the Rank Graduation metric is defined as

$$\mathrm{RG}_1(Y, Y') = 1 - \frac{\mathrm{CvM}_1(F_Y, F_{Y'})}{G(Y)}, \tag{3}$$

where $G(Y)$ denotes the Gini index computed on the empirical distribution of Y. This normalization ensures that RG takes values in $[0, 1]$ and provides a ranking-based measure of predictive agreement. In the binary classification case, when Y represents

the ground truth and $Y' = \hat{Y}$ the predicted probabilities, the Rank Graduation metric coincides with the Area Under the ROC Curve [2].

Different SAFE-AI dimensions are obtained by selecting different pairs of variables (Y, Y') within Equation 3. Rank Graduation Accuracy (RGA) evaluates agreement between true labels and predicted class probabilities. Rank Graduation Robustness (RGR) evaluates agreement between predictions before and after controlled perturbations of the input features. Rank Graduation Explainability (RGE) evaluates agreement between predictions before and after structured removal of relevant image regions identified through Grad-CAM.

For multiclass classification, RG is computed using a one-vs-rest strategy. A class-specific RG score is evaluated independently for each class by comparing the corresponding class indicator with the predicted class probability. The final multiclass score is obtained as a convex combination of class-specific RG values weighted by empirical class frequencies. The corresponding areas under the degradation curves, AURGA, AURGR, and AURGE, summarize how gracefully model predictions degrade under data removal, perturbation, and occlusion.

3 Experimental Setting

Experiments are conducted on the Brain Cancer MRI dataset [5], containing 6,056 clinically verified MRI images from three classes: glioma, meningioma, and brain tumor. All models are evaluated on the same standardized 512-dimensional ResNet-18 [4] feature space using 5-fold cross-validation. The proposed quantum machine learning model is compared with logistic regression, random forest, support vector machine, and multilayer perceptron. The MLP baseline is deliberately chosen to mirror the classical component of the hybrid quantum model, enabling a controlled comparison between purely classical and quantum-enhanced nonlinear mappings.

Robustness is evaluated at three levels. First, Gaussian noise is added to the standardized ResNet feature vectors. Second, for differentiable models, adversarial feature-space robustness is evaluated using the Fast Gradient Sign Method [3]. Third, image-level spatial robustness is assessed by applying controlled rotations and translations before feature extraction. Explainability stability is evaluated through Grad-CAM-guided patch occlusion [7], where the most relevant image regions are progressively blurred and the resulting predictions are compared with predictions on the original images.

4 Results

Table 1 reports the main predictive and SAFE-AI results. The SVM achieves the highest predictive accuracy and macro-F1 score, while the proposed quantum model closely follows it and outperforms the remaining baselines. In terms of SAFE-AI metrics, the SVM obtains the strongest ranking accuracy and AURGA, the random forest achieves the highest explainability stability, and the quantum model achieves the strongest robustness to Gaussian feature perturbations. These differences indicate that predictive accuracy alone does not fully characterize model reliability, and SAFE-AI metrics reveal complementary stability properties across model classes.

Results in Table 2 show that adversarial perturbations produce substantially larger degradation than spatial transformations, while spatial robustness remains high across all models. Among differentiable models, QML achieves the highest FGSM robustness and remains strongly stable under spatial perturbations, indicating improved resistance to gradient-based feature distortions.

Table 1: Mean 5-fold predictive performance and SAFE-AI scores on the MRI dataset.

Model	Acc.	F1	RGA	AURGA	AURGR	AURGE
Linear	0.938	0.938	0.9976	0.9289	0.8326	0.6477
RF	0.933	0.933	0.9994	0.9404	0.8891	**0.7132**
SVM	**0.983**	**0.983**	**0.9999**	**0.9488**	0.9393	0.7090
MLP	0.959	0.959	0.9979	0.9283	0.9360	0.7004
QML	0.977	0.977	0.9996	0.9469	**0.9409**	0.6880

Table 2: Supplementary robustness analysis using FGSM feature-space perturbations and image-level spatial perturbations

Model	AURGR-FGSM	AURGR-Spatial
Linear	0.4508	0.8938
RF	n/a	0.9240
SVM	n/a	**0.9389**
MLP	0.6683	0.9248
QML	**0.6961**	0.9163

These results show that the proposed quantum model achieves predictive performance close to the strongest classical baseline while exhibiting a competitive and complementary reliability profile, particularly under stochastic feature perturbations and adversarial stress among differentiable models. This behavior suggests that hybrid quantum models can provide useful stability-oriented properties in medical image analysis, where graceful degradation under stress may complement high predictive accuracy.

5 Discussion & Conclusion

This work evaluated hybrid variational quantum classifiers under a unified SAFE-AI reliability framework on medical MRI classification. The proposed model achieves predictive performance comparable to strong classical baselines while exhibiting improved robustness under stochastic and adversarial perturbations. These results suggest that normalized amplitude embeddings combined with variational quantum transformations can induce stable hypothesis classes suitable for reliability-sensitive applications.

Future work will investigate scalability to higher-resolution quantum encodings and extension of SAFE-AI evaluation to fairness-aware and distribution-shift settings.

References

[1] Golnoosh Babaei, Paolo Giudici, and Emanuela Raffinetti. 2025. A SAFE Rank Graduation Box. *Expert Systems with Applications* 259 (2025), 125047. doi:10.1016/j.eswa.2024.125047

[2] Paolo Giudici and Emanuela Raffinetti. 2025. RGA: a unified measure of predictive accuracy. *Advances in Data Analysis and Classification* 19 (2025), 67–93. doi:10.1007/s11634-023-00574-2

[3] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *arXiv 1412.6572* (12 2014).

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Las Vegas, NV, 770–778.

[5] Md Mizanur Rahman. 2024. *Brain Cancer – MRI dataset.* doi:10.17632/mk56jw9rns.1

[6] M. Schuld and N. Killoran. 2019. Quantum Machine Learning in Feature Hilbert Spaces. *Physical Review Letters* 122, 4 (2019), 040504. doi:10.1103/PhysRevLett.122.040504

[7] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 618–626. doi:10.1109/ICCV.2017.74