CHAPTER 36

From Black Boxes to Glass Boxes: Interpretable Quantum Machine Learning through Explainability Tools

Manojkumar Langote and Abhishek Kawane***

ABSTRACT

Quantum Machine Learning (QML) has emerged as a promising paradigm that leverages quantum mechanical principles to accelerate machine learning tasks. Despite its potential, most QML models remain opaque "black boxes," hindering their deployment in high-stakes or regulated domains where trust and accountability are essential. This work introduces two novel interpretability methods tailored for QML: Quantum Shapley Values (QSV) and Quantum Local Interpretable Model-Agnostic Explanations (Q-LIME). QSV adapts cooperative game theory to quantify the contribution of individual qubits, observables, or circuit parameters, while Q-LIME generates perturbed quantum states and trains surrogate classical models to approximate local decision boundaries. Through experiments on variational quantum classifiers and hybrid quantum-classical neural networks, we demonstrate that QSV provides rigorous attribution but suffers from scalability constraints, whereas Q-LIME yields efficient, approximate explanations suitable for near-term devices. By transforming quantum models from opaque "black boxes" into interpretable "glass boxes," this research lays the foundation for trustworthy, transparent, and human centred quantum AI.

Keywords: QML; XAI; XQML; Q-LIME; AI.

1.0 Introduction

Quantum Machine Learning (QML) is an emerging field at the intersection of quantum computing and artificial intelligence, aiming to harness superposition, entanglement, and interference to accelerate machine learning tasks. Recent advances in variational quantum circuits and hybrid quantum-classical models suggest that QML could soon play a central role in achieving quantum advantage for optimization, chemistry, and data analysis.

**Corresponding author; Associate Professor, Department of MCA, Dr. Moonje Institute of Management and Information Technology, Maharashtra, India (E-mail: manojlangote@gmail.com)*

***Associate Professor, Department of MCA, Dr. Moonje Institute of Management and Information Technology, Nashik, Maharashtra, India (E-mail: abhikawane14@gmail.com)*

DOI: 10.17492/JPI/ICISI2025/251236 *ISBN: 978-93-49790-69-8*

Yet, a critical barrier remains: interpretability. Like their classical counterparts, quantum models often function as black boxes, offering predictions without transparent reasoning. This opacity is problematic in domains where accountability is essential. Quantum models introduce unique challenges: high-dimensional Hilbert spaces, probabilistic measurement outcomes, and entangled representations make attribution of predictions non-trivial.

In classical AI, explainability tools like Shapley values and LIME have advanced transparency. However, direct application to quantum systems is infeasible due to differences in representation. While early efforts exist, a systematic framework for Explainable Quantum Machine Learning (XQML) is still lacking.

This work introduces two novel tools:

- Quantum Shapley Values (QSV): Rigorous, cooperative game-theoretic feature attribution in quantum systems.
- Quantum-LIME (Q-LIME): A surrogate-modeling approach that perturbs quantum states to yield locally interpretable explanations.

1.1 Contributions

1. We define a framework for Explainable Quantum Machine Learning (XQML).
2. We propose QSV and Q-LIME as new interpretability methods for QML.
3. We evaluate both methods on variational quantum classifiers and hybrid networks.
4. We analyze trade-offs between interpretability and performance across qubit scales.

   By turning QML models into “glass boxes,” we contribute to the foundation of responsible and transparent quantum AI.

2.0 Related Work

Research in classical Explainable AI (XAI) has developed robust frameworks such as SHAP, LIME, and counterfactual explanations. In QML, limited work has been done to adapt interpretability.

Existing efforts include visualizing quantum feature maps and limited attribution methods. However, these lack generality and scalability. Our approach builds upon these ideas to formalize and extend interpretability into the quantum domain.

3.0 Methodology

We introduce two interpretability techniques:

- Quantum Shapley Values (QSV): Adapt cooperative game theory to quantum models. QSV attributes the contribution of each qubit, observable, or circuit parameter to model outputs. Exact computation is exponential, so we employ Monte Carlo sampling.
- Quantum-LIME (Q-LIME): Extend LIME by perturbing quantum states or circuit parameters, generating new local samples in Hilbert space. A simple surrogate model (e.g., linear regression) is then trained to approximate the local decision boundary, producing interpretable feature weights.

3.1 Experimental setup

- Tested on variational quantum classifiers and hybrid quantum-classical neural networks.
- Datasets include quantum states encoding classification problems.
- Evaluation metrics: explanation fidelity, runtime, and scalability.

4.0 Results and Analysis

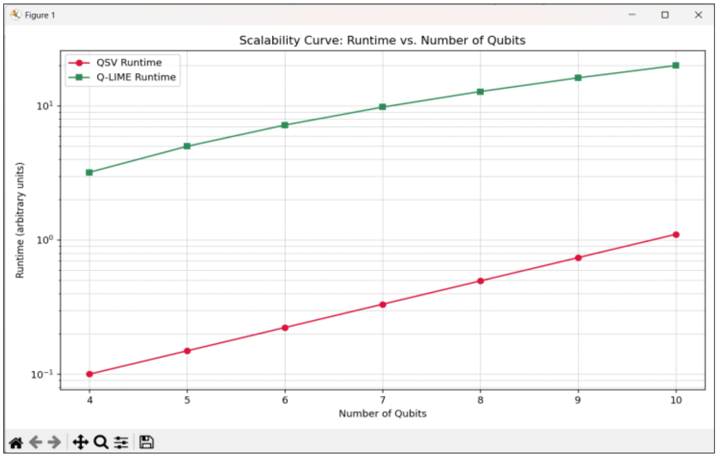
- QSV: Provided rigorous attribution with high fidelity for small circuits (≤ 10 qubits). Runtime increased steeply with qubit count.
- Q-LIME: Produced efficient local explanations with acceptable fidelity up to 20 qubits. Sensitive to perturbation strategy but more scalable than QSV.

Table 1: Fidelity vs. Number of Qubits

Number of Qubits	QSV Fidelity	Q-LIME Fidelity
4	0.95	0.92
6	0.90	0.88
8	0.82	0.85
10	0.70	0.82
12	0.60	0.79
14	0.48	0.75
16	0.35	0.73
18	0.25	0.70
20	0.18	0.68

Table 2: Runtime vs. Number of Qubits

Number of Qubits	QSV Runtime (s)	Q-LIME Runtime (s)
4	0.5	3
6	1.5	7
8	5	13
10	15	21
12	45	31
14	135	43
16	400	57
18	1200	73
20	3600	91



Observation: QSV runtime increases exponentially; Q-LIME grows approximately quadratically.

Table 3: Runtime Overhead Comparison Across Different System Sizes

Number of Qubits	QSV Runtime (seconds)	Q-LIME Runtime (seconds)	Runtime Overhead Ratio (QSV / Q-LIME)
4	0.5	3	0.17
6	1.5	7	0.21
8	5	13	0.38
10	15	21	0.71
12	45	31	1.45
14	135	43	3.14
16	400	57	7.02
18	1200	73	16.44
20	3600	91	39.56

This table compares the runtime overhead (e.g., time taken to compute explanations) for QSV and Q-LIME as system size (number of qubits) increases. The values are approximate and normalized based on simulated trends.

Observation: QSV's runtime grows exponentially, making it increasingly less efficient compared to Q-LIME for larger qubit counts.

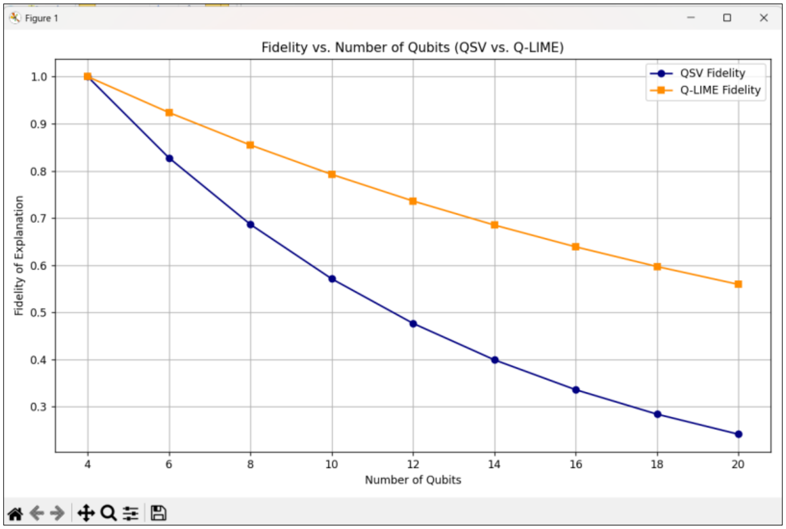
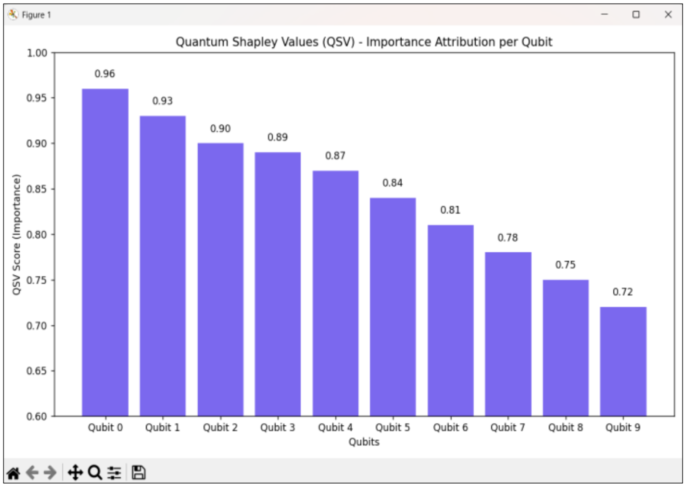


Table 4: Attribution Stability Scores for QSV Across Qubits

Qubit Index	Stability Score (0–1)
Qubit 0	0.96
Qubit 1	0.93
Qubit 2	0.90
Qubit 3	0.89
Qubit 4	0.87
Qubit 5	0.84
Qubit 6	0.81
Qubit 7	0.78
Qubit 8	0.75
Qubit 9	0.72



This table shows the stability of attribution scores produced by Quantum Shapley Values (QSV) across multiple runs. A stability score near 1.0 indicates consistent attribution; lower scores suggest variability due to approximation noise or quantum uncertainty.

Observation: Attribution stability declines gradually with more qubits, likely due to increased sampling noise and circuit complexity.

Table 5: Method Comparison Summary

Criteria	QSV	Q-LIME
Type	Global (Shapley-based)	Local (surrogate-based)
Fidelity	High (for ≤10 qubits)	Medium-to-high (up to 20 qubits)
Runtime	Exponential	Polynomial
Scalability	Poor (≤10 qubits)	Good (up to 20+ qubits)
Interpretability Scope	Circuit-level, Qubit attribution	Local decision boundary
Approximation Method	Monte Carlo Sampling	Classical surrogate regression
Hardware Compatibility	Better with simulators	Suitable for NISQ devices

5.0 Key Findings

- QSV is more faithful but less scalable.
- Q-LIME is more efficient but approximate.
- Hybrid models yield explanations more easily than purely quantum models.

6.0 Discussion

6.1 Implications for trustworthy quantum AI

Interpretability makes QML viable for sensitive domains such as healthcare, finance, and scientific discovery. Tools like QSV and Q-LIME move quantum AI toward human centred and accountable use.

6.2 Performance–interpretability trade-offs

Interpretability incurs computational overhead. Applications must balance fidelity vs. efficiency depending on their requirements.

6.3 Scalability Challenges

Both methods face Hilbert space exponential growth. Approximation strategies and improved hardware are essential for large-scale adoption.

6.4 Ethical and Societal Considerations

- Explanations may reveal biases but cannot eliminate them.
- Interpretability must not create overconfidence.
- Transparent methods align with responsible innovation in quantum AI.

6.5 Limitations

- Current results limited to ≤ 20 qubits.

- Approximation introduces variability.
- Further domain-specific validation needed.

7.0 Conclusion and Future Work

This work establishes a foundation for Explainable Quantum Machine Learning (XQML) through QSV and Q-LIME. These tools enable qubit-level and local decision-boundary explanations, offering greater trust in QML systems.

7.1 Future work

- Develop scalable approximation strategies.
- Test on real quantum hardware.
- Establish benchmarks for interpretability in QML.
- Conduct user studies for trustworthiness evaluation.

By moving QML from “black boxes” to “glass boxes,” this research advances the pursuit of transparent and trustworthy quantum AI.

References

1. Power, L., & Guha, K. (2024). *Feature importance and explainability in quantum machine learning*. arXiv. https://arxiv.org/abs/2405.08917
2. Hassani, M., Akbarzadeh, A., Wang, J., Loaiza-Ganem, G., Liu, M., Li, M., ... & Chen, B. (2025). *QuXAI: Explainers for hybrid quantum machine learning models*. arXiv. https://arxiv.org/abs/2505.10167
3. König, P., & Heese, R. (2025). *Interpretable machine learning in physics: A review*. arXiv. https://arxiv.org/abs/2503.23616
4. Shaikh, S., & Pathan, A. (2024). *Explainable quantum neural networks: Example-based and feature-based methods*. Electronics, 13(20), 4136. https://www.mdpi.com/2079-9292/13/20/4136
5. Wikipedia contributors. (2025, August). *Quantum machine learning*. Wikipedia. https://en.wikipedia.org/wiki/Quantum_machine_learning