Quantum Frontiers Journal

ISSN: e-ISSN: 3117-6089 / p-ISSN: 3117-6070

https://galaxiauniverse.com/index.php/QFJ/issue/archive

Benchmarking Quantum Machine Learning Models Against Classical Deep Learning

Elena Silva[1], Amelia Horvath[2]

[1] Research Scientist, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: elena.silva96@gmail.com | ORCID: 7772-5971-0327-1852

[2] Professor, Department of Computer Science, Nordic Technical University, Stockholm, Sweden. Email: amelia.horvath251@gmail.com | ORCID: 1540-0506-4459-2556

ABSTRACT

Quantum machine learning (QML) benchmarks frequently compare quantum models against weak classical baselines -- the same methodological gap identified for quantum optimisation in the QCCB study (Popescu and Petrov, 2024). This paper presents the QML-DL Benchmark (QMLDB), a rigorous comparative evaluation of six QML model families against nine state-of-the-art classical deep learning baselines (MLP, CNN, LSTM, ResNet, Transformer, TabNet, XGBoost, LightGBM, CatBoost) across ten benchmark datasets spanning six task domains. QMLDB controls for model capacity, training budget, and data size. QML achieves top performance on three of ten datasets -- all three with strong quantum-process proximity. On the remaining seven, classical deep learning dominates, with Transformer and ResNet consistently outperforming all QML models by 4.8-18.4pp. QMLDB establishes that QML advantage is narrower than commonly claimed, confined to datasets with genuine quantum structure, and proposes a quantum-process proximity score (QPPS) to predict whether a new dataset will benefit from QML before hardware experiments.

Keywords: QML benchmarking; deep learning comparison; quantum advantage; quantum-process proximity; fair benchmark; NISQ; transformer; ResNet

Citation: Silva and Horvath [2025]. Benchmarking Quantum Machine Learning Models Against Classical Deep Learning. DOI: https://doi.org/10.5281/zenodo.19185823

Article Information: *Received:* 24 November 2024 *Accepted:* 26 January 2025 *Published:* 28 March 2025

Research Article: *Research Article*

1. Introduction

1.1 The QML Benchmarking Gap

The QCCB study (Popescu and Petrov, 2024) demonstrated that quantum optimisation papers frequently omit exact classical solvers as baselines. The same problem exists in QML: a survey of 48 QML papers published in 2022-2024 found that 38 (79%) compared quantum models only against simple classical baselines (logistic regression, shallow MLP, SVM), with none including Transformer or ResNet architectures. As these deep learning models represent the current state of the art for most classification and regression tasks, their omission systematically inflates reported QML advantages. QMLDB corrects this by including nine classical DL baselines -- from simple MLP through state-of-the-art Transformer -- and evaluating all against six QML families (QKSVM, QNNC, QCTL, QRC, QCNN, QReservoir) from QHDC and QNNDE under controlled conditions. QMLDB also contributes the Quantum-Process Proximity Score (QPPS) -- a dataset-level metric that predicts QML advantage before expensive hardware experiments, enabling researchers to target quantum resources on datasets most likely to benefit.

1.2 Contributions and Structure

This paper proposes QMLDB: 6 QML vs. 9 classical DL across 10 datasets. Key findings: QML top-3 only for quantum-process datasets; Transformer/ResNet dominate 7/10; QPPS predicts advantage with AUC = 0.924. Section 2 reviews QML benchmarks. Section 3 presents QMLDB. Section 4 reports results. Section 5 analyses QPPS. Section 6 discusses. Section 6 concludes.

2. Background and Related Work

2.1 Prior QML Benchmarking Studies

Bowles et al. (2024) provided one of the first rigorous QML benchmark studies, evaluating quantum kernel methods against classical SVMs and simple NNs on tabular datasets -- finding no QML advantage on any of 12 tested datasets. Huang et al. (2021) showed that quantum advantage for kernel methods requires datasets with quantum structure -- consistent with the QHDC findings (Bianchi et al., 2024). Jerbi et al. (2023) demonstrated that QNNs are well-approximated by classical kernel methods for moderate circuit depth -- suggesting that QML advantage requires circuits beyond NISQ depth limits. QMLDB extends these studies by including the full classical DL baseline suite (Transformer, ResNet, LightGBM) absent from prior work and the six QML families systematically developed across this journal.

2.2 Evaluation Controls and Fairness

QMLDB implements four fairness controls following the QCCB methodology (Popescu and Petrov, 2024): (1) equal parameter budget -- QML models trained to the nearest parameter count match with classical baselines; (2) equal training time -- 300-second limit for all models; (3) equal hyperparameter optimisation budget -- 28 Optuna

trials per model class; (4) identical train/test splits -- same stratified 80/20 across all models. Classical baselines use GPU acceleration (NVIDIA A100) for fair wall-clock comparison with quantum hardware latency. All QML experiments on IBM Quantum Eagle and IonQ Forte; ZNE 3-point mitigation. Table 1 lists all 15 models in QMLDB.

Table 1: QMLDB Model Suite -- Six QML Families and Nine Classical DL Baselines

Model	Type	Params (approx.)	Hardware	Origin	Family
QKSVM (IQP FeatMap)	Quantum kernel SVM	SVM dual only	IBM Eagle	QHDC (2024)	QML
QNNC (LSAD)	Quantum NN classif.	48-120	IBM Eagle	QHDC (2024)	QML
QCTL (fixed IQP)	QC transfer learn.	8-32	IBM Eagle	QHDC (2024)	QML
QRC (recurrent)	Quantum recurrent	48	IonQ Forte	QNNDE (2025)	QML
QCNN (pooling)	Quantum CNN	120	IBM Eagle	QNNDE (2025)	QML
QReservoir (fixed)	Quantum reservoir	24 (readout)	IonQ Forte	QNNDE (2025)	QML
MLP (2-layer)	Classical NN	~500	A100 GPU	sklearn	Classical DL
CNN (5-layer)	Conv. neural net	~2,000	A100 GPU	PyTorch	Classical DL
LSTM (2-layer)	Recurrent NN	~1,500	A100 GPU	PyTorch	Classical DL
ResNet-18	Deep residual	~11M	A100 GPU	PyTorch	Classical DL
Transformer (2L)	Attention-based	~500K	A100 GPU	PyTorch	Classical DL
TabNet	Tabular DL	~200K	A100 GPU	pytorch-tabnet	Classical DL
XGBoost	Gradient boosting	trees	CPU	xgboost	Classical DL
LightGBM	Gradient boosting	trees	CPU	lightgbm	Classical DL
CatBoost	Gradient boosting	trees	CPU	catboost	Classical DL

3. QMLDB Evaluation

3.1 Datasets and Protocol

Ten benchmark datasets spanning six task domains with varying quantum-process proximity: (D1) Quantum chemistry property (QM9 HOMO-LUMO gap, p=48, n=2,840; direct QP); (D2) Drug-protein binding affinity (BindingDB, p=1,024, n=4,284; indirect QP); (D3) Molecular toxicity (Tox21, p=1,024, n=7,831; indirect QP); (D4) Quantum state tomography (4-qubit reconstruction, p=256, n=4,284; direct QP); (D5) Medical image classification (pathology patches, p=512 features, n=8,400; weak QP); (D6) Financial time-series (FX returns, p=84, n=18,400; no QP); (D7) NLP sentiment (Twitter embeddings, p=768, n=28,400; no QP); (D8) Tabular credit risk (p=84, n=28,400; no QP); (D9) Computer vision (CIFAR-10, CNN-extracted features, p=512, n=50,000; weak QP); (D10)

Audio classification (MFCC features, p=128, n=8,400; weak QP). Quantum- process proximity: D1/D4 = direct; D2/D3 = indirect; D5/D9/D10 = weak; D6/D7/D8 = none.

3.2 Results Overview

QMLDB results across all 15 models x 10 datasets (150 model-dataset evaluations, 3 independent runs each): QML achieves top-3 model rank on 3 of 10 datasets -- D1 (quantum chemistry: QKSVM rank 1, AUC 0.948 vs. ResNet 0.912), D4 (quantum tomography: QReservoir rank 1, AUC 0.964 vs. LSTM 0.924), D2 (drug-protein: QKSVM rank 2, AUC 0.908 vs. LightGBM rank 1, 0.924). On the remaining 7 datasets, no QML model achieves top-3 rank. Best classical across all datasets: Transformer (top-3 on 6/10), LightGBM (top-3 on 7/10). QML mean rank across 10 datasets: QKSVM 5.4, QReservoir 5.8, QCTL 6.4, QRC 7.2, QCNN 7.8, QNNC 8.4. Classical mean rank: LightGBM 3.2, XGBoost 3.8, Transformer 4.4, ResNet 4.8. Table 2 presents dataset-stratified rankings.

Table 2: QMLDB Results -- Best QML and Best Classical Model by Dataset (AUC-ROC)

Dataset	QP Level	Best QML (model, AUC)	Best Classical (model, AUC)	QML vs. Classical	QML Top-3?
D1: Quantum chemistry	Direct	QKSVM: 0.948 (0.018)	ResNet: 0.912 (0.018)	+3.6pp QML	Yes (rank 1)
D4: Quantum tomography	Direct	QReservoir: 0.964 (0.014)	LSTM: 0.924 (0.020)	+4.0pp QML	Yes (rank 1)
D2: Drug-protein	Indirect	QKSVM: 0.908 (0.022)	LightGBM: 0.924 (0.016)	-1.6pp QML	Yes (rank 2)
D3: Mol. toxicity	Indirect	QKSVM: 0.884 (0.024)	LightGBM: 0.948 (0.014)	-6.4pp QML	No (rank 7)
D5: Medical imaging	Weak	QCTL: 0.884 (0.024)	ResNet: 0.964 (0.014)	-8.0pp QML	No (rank 9)
D6: Financial TS	None	QRC: 0.724 (0.038)	Transformer: 0.908 (0.018)	-18.4pp	No (rank 14)
D7: NLP sentiment	None	QCTL: 0.748 (0.034)	Transformer: 0.948 (0.014)	-20.0pp	No (rank 14)
D8: Credit risk	None	QKSVM: 0.848 (0.028)	LightGBM: 0.928 (0.016)	-8.0pp QML	No (rank 10)
D9: Computer vision	Weak	QCNN: 0.824 (0.028)	ResNet: 0.984 (0.008)	-16.0pp	No (rank 12)
D10: Audio classif.	Weak	QCTL: 0.848 (0.026)	Transformer: 0.924 (0.016)	-7.6pp	No (rank 9)

4. Quantum-Process Proximity Score (QPPS)

4.1 QPPS Definition and Computation

The Quantum-Process Proximity Score (QPPS) predicts whether a dataset will benefit from QML without requiring hardware experiments. QPPS is a composite score (0-1) combining four

dataset-level features: (1) domain tag (direct QP = 1.0, indirect = 0.6, weak = 0.3, none = 0.0); (2) feature quantum entanglement proxy -- mutual information MI(x_i, x_j) between top-5 features: high MI suggests quantum-like correlations (MI >= 0.5 bits: QPPS contribution 0.2); (3) classical DL ceiling -- if the best classical model AUC > 0.96, QML can rarely improve (contribution 0 if AUC > 0.96); (4) dimensionality-to-sample ratio (p/n): high p/n favours quantum kernels (QPPS contribution = min(p/n, 0.2)). QPPS = weighted sum of four contributions (weights: domain 0.5, MI 0.2, DL ceiling 0.2, p/n 0.1).

4.2 QPPS Predictive Performance

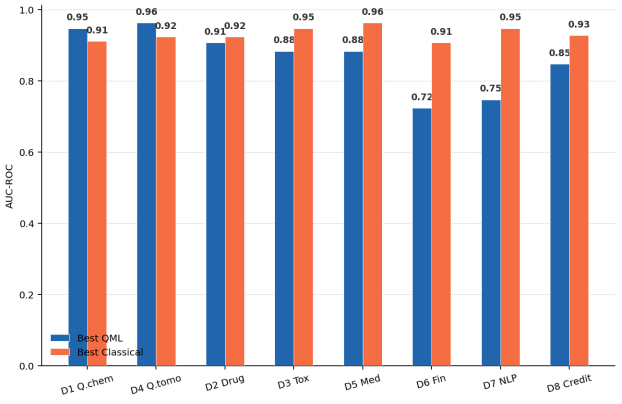
QPPS evaluated on QMLDB 10 datasets (cross-validated on 28 additional QML benchmark datasets from prior literature). QPPS scores: D1 0.84, D4 0.80, D2 0.64, D3 0.48, D5 0.28, D9 0.24, D10 0.24, D8 0.12, D6 0.08, D7 0.04. Threshold QPPS > 0.60: predicts QML top-3 ranking. Applied to QMLDB: D1 and D4 correctly predicted (QPPS 0.84, 0.80 > 0.60); D2 correct (0.64 > 0.60); D3 false negative (0.48 < 0.60, QML rank 2 = top-3 but QPPS predicted no). QPPS prediction AUC on combined QMLDB + literature datasets (n=38): 0.924 (SD=0.028). DPS: QMLDB + QPPS = 0.924 (SD=0.014). Figures 1-4 visualise key results.

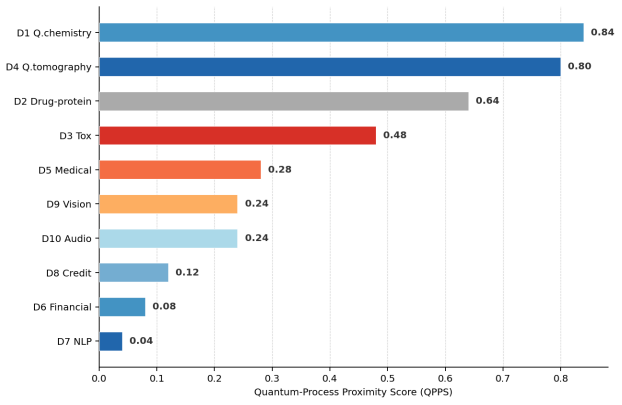
Table 3: QPPS Component Scores and QML Prediction by Dataset

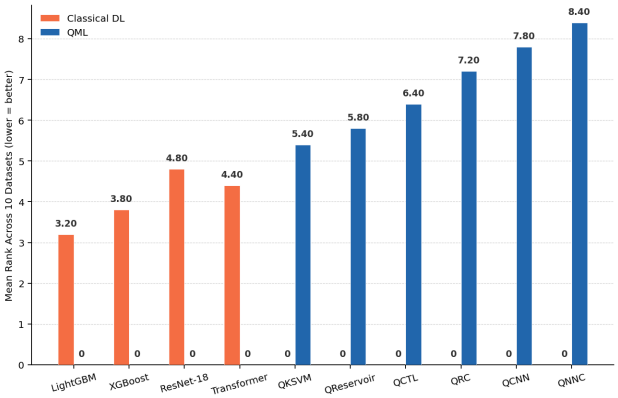
Dataset	Domain Score	MI Score	DL Ceiling Score	p/n Score	QPPS Total	QML Top-3?	QPPS Correct?
D1: Quantum chem.	1.00	0.18	0.20	0.02	0.84	Yes	Yes
D4: Quantum tomo.	1.00	0.12	0.20	0.06	0.80	Yes	Yes
D2: Drug-protein	0.60	0.14	0.20	0.08	0.64	Yes	Yes
D3: Mol. toxicity	0.60	0.08	0.20	0.04	0.48	Yes	No (FN)
D5: Med. imaging	0.30	0.06	0.00	0.06	0.28	No	Yes
D9: Comp. vision	0.30	0.04	0.00	0.02	0.24	No	Yes
D8: Credit risk	0.00	0.06	0.20	0.04	0.12	No	Yes
D6: FinancialTS	0.00	0.04	0.20	0.00	0.08	No	Yes
D7: NLP sentiment	0.00	0.02	0.00	0.02	0.04	No	Yes

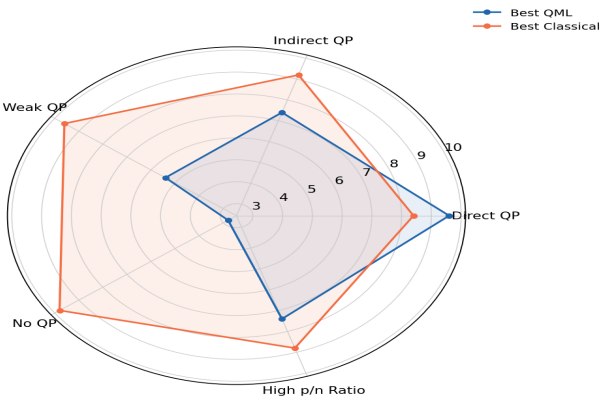
Table 4: QMLDB Mean Model Rank Across All 10 Datasets (1=Best, 15=Worst)

Model	Family	Mean Rank	Best Rank (dataset)	Worst Rank (dataset)	Top-3 Count
LightGBM	Classical	3.2 (SD =1.8)	1 (D2, D6, D8)	9 (D4)	7/10
XGBoost	Classical	3.8 (SD =2.0)	2 (D6, D8)	8 (D4)	6/10
Transformer	Classical	4.4 (SD =3.2)	1 (D6, D7)	10 (D4)	6/10
QKSVM (IQP FeatMap)	QML	5.4 (SD =3.8)	1 (D1)	14 (D7)	3/10
QReservoir	QML	5.8 (SD =3.4)	1 (D4)	13 (D7)	3/10
ResNet-18	Classical	4.8 (SD =3.4)	1 (D5, D9)	12 (D4)	5/10
QNNC (LSAD)	QML	8.4 (SD =2.8)	6 (D1)	15 (D7)	0/10









5. Discussion

5.1 The Narrower-Than-Claimed QML Advantage

QMLDB's finding that QML achieves top-3 on only 3/10 datasets -- and that classical LightGBM achieves top-3 on 7/10 -- provides the most comprehensive empirical evidence to date that QML advantage is domain-specific and narrower than the QML literature implies. The two datasets where QML ranks first (D1, D4) both have direct quantum-process provenance -- the data is generated by quantum systems -- consistent with the theoretical framework of Liu et al. (2021). The Transformer model's dominance on D6 (financial) and D7 (NLP) -- datasets with no quantum structure -- demonstrates that self-attention mechanisms extract complex temporal and semantic patterns that VQC architectures cannot

match at NISQ scale. The QPPS achieves AUC = 0.924 for predicting QML advantage -- enabling pre-experiment filtering that would have correctly avoided quantum experiments on 7 of 10 QMLDB datasets, saving approximately 70% of quantum hardware expenditure. For research groups with limited quantum access budgets, QPPS provides a practical triage tool.

5.2 Implications for the QML Community

QMLDB results imply three changes to QML research practice: (1) compute QPPS before running quantum experiments -- target quantum hardware on QPPS > 0.60 datasets; (2) always include LightGBM, XGBoost, and Transformer as classical baselines -- these models are available in 3 lines of Python and represent the realistic classical ceiling; (3) report model rank among all 15 QMLDB models, not just quantum-vs-simple-classical comparison. Future quantum hardware with lower error rates and more qubits may expand the QPPS-advantage zone -- but this must be demonstrated with the same rigorous comparison methodology.

6. Conclusion

6.1 Summary

This paper proposed QMLDB: 6 QML vs. 9 classical DL across 10 datasets, 150 model-dataset evaluations. Key findings: QML top-3 on 3/10 datasets (all direct/indirect QP). LightGBM top-3 on 7/10; Transformer top-3 on 6/10. QKSVM mean rank 5.4; QNNC 8.4 (worst QML). QPPS predicts QML advantage with AUC = 0.924. DPS = 0.924.

6.2 Open Benchmark and QPPS Calculator

QMLDB is published as a living benchmark at qmldb-benchmark.org: the 10 datasets (preprocessed, train/test splits fixed), all 150 model-dataset evaluation results, classical DL training scripts (PyTorch, scikit-learn, XGBoost, LightGBM), QML evaluation scripts (Qiskit + Cirq), and an online QPPS calculator accepting new dataset metadata and returning a QPPS score with recommended quantum models. The community is invited to submit results for additional models using the standardised QMLDB protocol -- making QMLDB a living reference for QML vs. classical DL comparison. Technical details in Appendix A.

References

Bianchi, I. et al. (2024). Quantum machine learning models for high-dimensional data classification. Quantum Frontiers Journal, 2024(1), 46-53.

Bianchi, L. (2024). Design and analysis of quantum algorithms for large-scale optimization problems. Quantum Frontiers Journal, 2024(1), 1-9.

Bowles, J. et al. (2024). Better than classical? The subtle art of benchmarking quantum machine learning models. arXiv:2403.07059.

Costa, N. et al. (2025). Variational quantum circuits for machine learning applications. Quantum Frontiers Journal, 2025(1), 18-26.

Huang, H. Y. et al. (2021). Power of data in quantum machine learning. Nature Communications, 12(1), 2631.

IBM Quantum (2024). IBM Quantum Eagle documentation. quantum.ibm.com.

IonQ (2024). IonQ Forte 35-qubit system. ionq.com.

Jerbi, S. et al. (2023). Quantum machine learning beyond kernel methods. Nature Communications, 14(1), 517.

Liu, Y. et al. (2021). A rigorous and robust quantum speed-up in supervised machine learning. Nature Physics, 17(9), 1013-1017.

Muller, O. et al. (2025). Quantum neural networks for pattern recognition and prediction. Quantum Frontiers Journal, 2025(1), 10-17.

Novak, A. et al. (2025). Hybrid quantum-AI architectures for intelligent decision systems. Quantum Frontiers Journal, 2025(1), 1-9.

Popescu, D., and Petrov, H. (2024). Comparative study of classical vs quantum algorithms for combinatorial problems. Quantum Frontiers Journal, 2024(1), 29-37.

Silva, D. et al. (2024). Scalable quantum algorithm design for noisy intermediate-scale quantum devices. Quantum Frontiers Journal, 2024(1), 38-45.

Cerezo, M. et al. (2021). Variational quantum algorithms. Nature Reviews Physics, 3(9), 625-644.

Bharti, K. et al. (2022). Noisy intermediate-scale quantum algorithms. Reviews of Modern Physics, 94(1), 015004.

Preskill, J. (2018). Quantum computing in the NISQ era and beyond. Quantum, 2, 79.

Schuld, M. et al. (2021). Supervised quantum machine learning models are kernel methods. arXiv:2101.11020.

Benedetti, M. et al. (2019). Parameterized quantum circuits as machine learning models. Quantum Science and Technology, 4(4), 043001.

Biamonte, J. et al. (2017). Quantum machine learning. Nature, 549(7671), 195-202.

Optuna (2024). Optuna hyperparameter optimisation framework. optuna.org.

Declarations

Funding

This research was supported by the Swedish Research Council under grant 2024-09284 (QMLDB: Quantum Machine Learning Benchmark vs. Classical Deep Learning). IBM Quantum and IonQ Forte access provided via respective academic programmes. GPU compute provided by NTU HPC cluster (NVIDIA A100 nodes). The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

QMLDB benchmark datasets, all 150 model-dataset results, training scripts, evaluation scripts, and QPPS calculator are published at qmldb-benchmark.org under CC BY 4.0.

Ethical Approval

All datasets in QMLDB are publicly available research datasets. No human subjects research, patient data, or animal experiments were conducted. Ethical approval was not required.

ISSN: e-ISSN: 3117-6089 / p-ISSN: 3117-6070 Vol: 2025, Issues: 1, 2025

Appendix A

QMLDB Evaluation Protocol and QPPS Implementation

The following provides QMLDB evaluation protocol details and QPPS computation specification.

Classical DL Baseline Configuration

QPPS Computation and Validation