Karshi Multidisciplinary International Scientific Journal

www.kmisj.reapress.com

K. Multidiscip. Int. Sci. J. Vol. 2, No. 4 (2025) 223–227.

Paper Type: Original Article

Hybrid Quantum–Classical Benchmarks for Synthetic DNA Risk Classification: A Fully Simulated Study

Rasim Abiyev*

Azerbaijan University of Languages, Labo Exchange Foundation Institute; abiyev.rasim@gmail.com.

Citation:

Received: 03 May 2025 *Revised: 25 July 2025* *Accepted: 12 October 2025*	Abiyev, R. (2025). Hybrid quantum–classical benchmarks for synthetic DNA risk classification: A fully simulated study. *Karshi multidisciplinary international scientific journal,* 2(4), 223-227.

Abstract

Hybrid quantum–classical architectures offer a promising direction for sequence modeling, yet their empirical behavior on realistic bioinformatics tasks remains insufficiently documented. Here, we present a fully simulated, reproducible benchmark for hybrid Quantum Machine Learning (QML) applied to synthetic Deoxyribonucleic Acid (DNA) disease-risk classification. A dataset of 5,000 sequences (200 bp) across five balanced classes was generated using motif-injection rules with controlled noise, from which 74 biological features were extracted. Baseline models (Convolutional Neural Network (CNN), attention, classical ensemble) were compared against two quantum-hybrid models incorporating a 4-qubit ZZFeatureMap+RealAmplitudes ansatz, executed exclusively on the Qiskit Aer Simulator with 1,024 shots. The attention model achieved the highest test accuracy (51.7%), while quantum-hybrid models produced comparable performance (51.1–51.3%), showing no measurable quantum advantage under these settings. The study establishes an honest, fully reproducible baseline for QML in genomics, highlighting current limitations of small-qubit encodings and motivating future work with trainable variational circuits and larger biological datasets.

Keywords: Quantum machine learning, Hybrid quantum-classical models, DNA sequence classification, Genomic risk prediction, Bioinformatics, Deep learning v.

1 | Introduction

The title "Machine-learning approaches for genomic sequence analysis" has advanced substantially, yet many tasks remain challenging due to high-dimensional interactions, motif entanglement, and nonlinear dependencies across long Deoxyribonucleic Acid (DNA) sequences. Quantum Machine Learning (QML) has been proposed as a potential means of enriching representation capacity through superposition and entanglement. However, practical evidence—especially for biologically relevant problems—has been limited, often relying on idealized toy datasets or assumptions that do not reflect realistic genomic structure [1], [2].

Corresponding Author: abiyev.rasim@gmail.com

https://doi.org/10.22105/kmisj.v2i4.98

This work addresses this gap by providing a fully simulated, entirely reproducible benchmark of hybrid quantum–classical learning applied to synthetic DNA risk classification. Unlike prior studies making strong claims of quantum benefit, our goal is explicitly conservative [3–5]:

I. Assess whether a shallow 4-qubit quantum feature extractor contributes measurable improvements over classical models.

II. Provide transparent reporting of all methods, data- generation rules, and training procedures.

III. Establish a baseline that can be extended to real genomic datasets and larger quantum circuits.

2 | Methods

2.1 | Synthetic Deoxyribonucleic Acid Dataset

A dataset of 5,000 DNA sequences (200 bp) was generated and balanced across five classes:

I. Healthy.

II. Type 2 diabetes risk.

III. Cardiovascular risk.

IV. Alzheimer's risk.

V. Cancer-associated patterns.

Disease motifs (12 bp) were embedded 3–5 times per sequence at random positions with 20–30% stochastic noise:

I. Diabetes: TCTAGCTAGCTA.

II. Cardiovascular: GCGCAATTGCGC.

III. Alzheimer: ATAGCGATAGCG.

IV. Cancer: CGCGATCGATCG dataset split:

V. 4,000 training.

VI. 1,000 test samples.

2.2 | Feature Engineering

Each sequence was encoded into 74 real-valued features, including:

I. Guanine–Cytosine (GC)/Adenine–Thymine (AT) ratios.

II. Cytosine-phosphate-Guanine (CpG)-island proportion.

III. Shannon entropy.

IV. Melting temperature.

V. DNA stability indices.

VI. 3-mer and 4-mer frequency vectors.

2.3 | Classical Models

Five classical/hybrid architectures were evaluated:

I. Convolutional Neural Network (CNN) (208,133 parameters).

II. Attention-based model.

III. Classical ensemble.

IV. Fast Hybrid (quantum feature extractor+Multi-Layer Perceptron (MLP)).

V. Ensemble Hybrid (quantum features + classical ensemble).

Training settings (shared):

I. Optimizer: Adam (lr = 0.001).

II. Batch size: 32.

III. Loss: CrossEntropyLoss.

IV. Epochs: Max 50.

V. Early stopping: Patience = 10.

VI. Hardware: Central Processing Unit (CPU).

2.4 | Quantum Feature Extractor (Qiskit Aer Simulation)

2.4.1 | Quantum circuit

I. Qubits: 4.

II. Feature map [6]: ZZFeatureMap (circular entanglement).

III. Ansatz: RealAmplitudes, entanglement depth = 2.

IV. Shots: 1,024.

V. Backend: Qiskit AerSimulator.

VI. No noise model.

VII. Optimizer inside quantum module: Constrained Optimization BY Linear Approximations (COBYLA).

VIII. Quantum parameters not trained.

2.4.2 | Pipeline

I. 74 classical features → 4D reduction.

II. Angle encoding into qubit rotations.

III. Entanglement via Controlled-X (CX) gates.

IV. Measurement over 1,024 shots.

V. Probability vector (16-dimensional) appended to classical features [7].

2.5 | Compute Cost

Total training time: ~45 minutes, broken down as:

I. CNN: ~10 min.

II. Attention: ~5 min.

III. Classical ensemble: ~3 min.

IV. Fast hybrid: ~2 min.

V. Ensemble hybrid: ~25 min (quantum simulation dominated).

Quantum overhead: ≈160,000 circuit executions × ~0.05 s each.

3 | Results

3.1 | Performance Summary

Table 1. Performance comparison of all evaluated models on the classification task.

Model	Epochs	Train Loss	Train Acc	Test Acc	F1 Score
CNN	49	0.8832	66.90%	42.00%	0.4180
Attention	28	1.1918	52.42%	51.70%	0.5169
Classical ensemble	17	1.1458	54.33%	46.50%	0.4651
Fast hybrid	21	1.2043	52.65%	51.10%	0.5102
Ensemble hybrid	28	1.1583	53.97%	51.30%	0.5125

3.2 | Interpretation

I. The attention model achieved the best test accuracy (51.7%).

II. The two quantum-hybrid models achieved comparable performance (51.1–51.3%) despite increased computational cost.

III. CNN exhibited strong overfitting (66.9% on the training set vs. 42.0% on the test set).

IV. The small 4-qubit circuit provided no observable advantage, consistent with expectations for shallow circuits and fixed, untrained parameters.

4 | Discussion

This work provides a transparent, fully simulated benchmark that illustrates the current limitations of hybrid QML for DNA-sequence classification. The absence of quantum advantage is expected and grounded in several factors [8], [9]:

I. Qubit count = 4, insufficient for representing 200-bp sequence complexity.

II. Circuit depth = 2, limiting expressivity.

III. Synthetic motifs, while structured, lack genuine genomic heterogeneity.

IV. Quantum parameters were not optimized, reducing representation capacity.

V. Quantum simulations increased runtime substantially.

Importantly, these limitations do not diminish the value of the benchmark: they underscore what present-day QML can realistically achieve and provide a reproducible baseline for future work with:

I. deeper parameterized circuits.

II. variational training of quantum parameters.

III. 20+qubit simulations or hardware.

IV. and real genomic datasets (e.g., Clinical Variants (ClinVar), the Database of Single Nucleotide Polymorphisms (dbSNP), and the Encyclopedia of DNA Elements (ENCODE)).

5 | Conclusion

We present a fully simulated hybrid QML benchmark for synthetic DNA risk prediction using small-scale quantum circuits. Classical attention models performed best, while quantum-hybrid models achieved similar accuracy without surpassing them. These results provide a realistic appraisal of QML capability under small-qubit constraints and establish a reproducible foundation for future genomic QML research.

Authors' Contributions

All aspects of the research and manuscript preparation were carried out by the author. The author has read and approved the final version of the manuscript.

Data Availability

All DNA sequences (5,000 samples, 200 bp) were generated synthetically using deterministic motif-injection rules. The dataset and generation scripts are available upon request and will be publicly released upon publication.

Funding

Not applicable.

Conflict of Interest

The author declares that they have no conflicts of interest.

Consent for Publication

The author confirms consent for the publication of this work

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., & Lloyd, S. (2017). Quantum machine learning. *Nature*, *549*(7671), 195–202. https://doi.org/10.1038/nature23474

[2] Peruzzo, A., McClean, J., Shadbolt, P., Yung, M. H., Zhou, X. Q., Love, P. J., & O'brien, J. L. (2014). A variational eigenvalue solver on a photonic quantum processor. *Nature communications*, *5*(1), 4213. https://doi.org/10.1038/ncomms5213

[3] Angermueller, C., Pärnamaa, T., Parts, L., & Stegle, O. (2016). Deep learning for computational biology. *Molecular systems biology*, *12*(7), MSB156651. https://doi.org/10.15252/msb.20156651

[4] Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning-based sequence model. *Nature methods*, *12*(10), 931–934. https://doi.org/10.1038/nmeth.3547

[5] Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., & Dalal, K. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature genetics*, *53*(3), 354–366. https://doi.org/10.1038/s41588-021-00782-6

[6] Havlíček, V., Córcoles, A. D., Temme, K., Harrow, A. W., Kandala, A., Chow, J. M., & Gambetta, J. M. (2019). Supervised learning with quantum-enhanced feature spaces. *Nature*, *567*(7747), 209–212. https://doi.org/10.1038/s41586-019-0980-2

[7] Mitarai, K., Negoro, M., Kitagawa, M., & Fujii, K. (2018). Quantum circuit learning. *Quantum circuit learning*. *Physical review A*, *98*, 32309. https://doi.org/10.1103/PhysRevA.98.032309

[8] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., & Fujii, K. (2021). Variational quantum algorithms. *Nature reviews physics*, *3*(9), 625–644. https://doi.org/10.1038/s42254-021-00348-9

[9] Schuld, M., Bergholm, V., Gogolin, C., Izaac, J., & Killoran, N. (2019). Evaluating analytic gradients on quantum hardware. *Physical review A*, *99*(3), 32331. https://doi.org/10.1103/PhysRevA.99.032331