Empirical Quantum Advantage Analysis of Quantum Kernel in Gene Expression Data

Arpita Ghosh
Computer Science & Engineering
Shahjalal University of Science & Technology
Sylhet, Bangladesh
ghosharpita423@gmail.com

Seemanta Bhattacharjee
Computer Science & Engineering
Shahjalal University of Science & Technology
Sylhet, Bangladesh
babune99@gmail.com

MD Muhtasim Fuad
Computer Science & Engineering
Shahjalal University of Science & Technology
Sylhet, Bangladesh
muhtasimfuad625@gmail.com

***Abstract*—The incorporation of quantum ansatz with machine learning classification models demonstrates the ability to extract patterns from data for classification tasks. However, taking advantage of the enhanced computational power of quantum machine learning necessitates dealing with various constraints. In this paper, we focus on constraints like finding suitable datasets where quantum advantage is achievable and evaluating the relevance of features chosen by classical and quantum methods. Additionally, we compare quantum and classical approaches using benchmarks and estimate the computational complexity of quantum circuits to assess real-world usability. For our experimental validation, we selected the gene expression dataset, given the critical role of genetic variations in regulating physiological behavior and disease susceptibility. Through this study, we aim to contribute to the advancement of quantum machine learning methodologies, offering valuable insights into their potential for addressing complex classification challenges in various domains.**

***Index Terms*—Empirical Quantum Advantage, Kernel Trick, Quantum Machine Learning, Quantum Annealers.**

I. Introduction

In the face of rapid mutations and variations in diseases, the ever-expanding volume of biological data presents unparalleled opportunities for unraveling intricate biological phenomena. Among these, gene expression analysis emerges as a cornerstone tool for comprehending the molecular mechanisms underlying diverse physiological and pathological processes. This comprehension leads to the accurate classification of cancer subtypes, which is essential for guiding individualized treatment strategies [1]. In this endeavor, the Golub et al. gene expression dataset has been serving as an instrumental resource facilitating numerous studies aimed at precisely categorizing different subtypes of leukemia based on their distinctive gene expression profiles.

However, the sheer magnitude and intricacy of gene expression data pose formidable challenges to extracting meaningful insights. Traditional classification methods often face computational limitations when tasked with discerning patterns amidst the noise and high-dimensional feature spaces inherent in gene expression data. In this context, harnessing these principles of quantum mechanics, quantum computing offers to exponentially accelerate data processing tasks, resulting in a paradigm shift in computational power and efficiency [2].

In this paper, we conduct an experiment investigation on gene expression data, exploring and evaluating the efficacy of both classical and quantum computing approaches for solving key challenges in gene expression classification. Our approach integrates innovative methodologies in feature selection, classification algorithms, and complexity analysis to advance our understanding of biological systems. Specifically, as part of our gene expression data analysis, we utilize quantile normalization [3] to preprocess the Golub et al. dataset, ensuring uniformity and reliability of data in our subsequent analyses. After preprocessing, it's significant to filter out relevant features to simplify the model and improve computational efficiency. To conduct a comparative study, we adopt both quantum and classical approaches for this task. The paper [4] highlights the effectiveness of Lasso in identifying relevant features from high-dimensional datasets, particularly in the field of genomics where the number of features often exceeds the number of samples. So we have employed LASSO Regularization (L1) [5] for classical means of feature selection. For the quantum approach to feature selection, we utilized D-Wave's hybrid quantum-classical framework [6] leveraging D-Wave's quantum annealers [7] to formulate the Quantum Unconstrained Binary Optimization (QUBO) problem. This hybrid approach allows us to explore alternative solutions for feature selection tasks in high-dimensional datasets.

We have classified the data using both quantum and classical kernels, utilizing the features selected by both approaches. To evaluate the performance of both the classical and quantum kernels using multiple metrics, including the F1 score, balanced accuracy, and Phase Terrain Ruggedness Index (PTRI),

geometric difference [8]. Finally, we conduct a comparative analysis of the computational complexity of quantum and classical kernels to evaluate their practical feasibility in large-scale gene expression analysis.

II. Methodology

According to a genome-wide association study (GAWS), detecting the precise biomarker contributes vitally to diagnosing specific diseases. Gene expression data with large feature space also enables the exploitation of the exponential transformation capability of quantum embedding. The gene expression dataset (Golub et al.) generated by a proof-of-concept study for cancer subtyping task (AML vs ALL) comprises 7129 gene expression profiles for each of the 38 training and 34 test samples. As gene expression values vary in a wide range, quantile normalization [3] is performed to make the values comparable across different samples.

Then from this large normalized data, 20 important features are extracted using [9] the L1 regularization(Lasso) method. Here are the gene accession number of the selected genes [‘AB000466_at’, ‘D17391_at’, ‘D38524_at’, ‘D78134_at’, ‘HG3945-HT4215_at’, ‘J04990_at’, ‘J05158_at’, ‘L01664_at’, ‘M26602_at’, ‘M60047_at’, ‘M63904_at’, ‘S67156_at’, ‘S77094_at’, ‘U30828_at’, ‘U47011_cds1_at’, ‘U63289_at’, ‘U66580_at’, ‘U70981_at’, ‘M15169_at’, ‘J00268_s_at’]. Simultaneously, the resultant genes from QUBO are [‘D17391_at’, ‘HG1148-HT1148_at’, ‘HG4188-HT4458_at’, ‘L09717_at’, ‘M77810_at’, ‘S82240_at’, ‘U14550_at’, ‘U16997_at’, ‘U19878_at’, ‘U34360_at’, ‘U49248_at’, ‘U63289_at’, ‘X07820_at’, ‘X14046_at’, ‘X17042_at’, ‘U31556_at’, ‘M27783_s_at’, ‘M63438_s_at’, ‘HG3731-HT4001_r_at’, ‘U84388_at’].

To build a more robust model against the outliers, min-max scaling is performed. After tuning the parameter, it’s inferred that the range from 0 to π works well for this specific experiment. Support vector machine(SVM) is a widely used tool for classification tasks in supervised machine learning that applies kernel tricks when the data is not linearly separable in their original feature space. It transforms data into higher dimensional feature space where the data is linearly separable. The efficient kernel approximation is crucial for resource optimization and performance enhancement of a model. For this experiment, we have adopted the quantum kernel estimation method implemented by Havlicek et al [10] where the kernel function is calculated using a quantum circuit and then it’s passed to the classical SVM for drawing a decision boundary. This model is implemented using qiskit library.

For solving certain tasks, utilizing quantum properties provides exponential speedup over the best-known classical approach, which is termed as quantum supremacy. But in near-term quantum hardware, the quantum advantage is yet to be attained for all kinds of operations. Due to the vulnerable nature of Noisy intermediate-scale quantum(NISQ) devices and the decoherence effects of qubits, the problem is critical to the size of the feature space. To verify the feasibility of the quantum approach, some heuristic metrics are applied. This paper [8] introduced a novel approach for this type of measurement named empirical quantum advantage(EQA). It’s a framework that analyzes different performance metrics. For instance, in this experiment, the following measures have been used at various configuration spaces: F1 score, balanced accuracy, and the phase space terrain ruggedness index (PTRI). The paper [8] also suggests geometric difference as a framework to evaluate the kernel’s efficiency where the quantum advantage is potential. Here, the metric is applied to determine the kernel’s capability to generate the optimized decision boundary for a classification problem. It provides insight into which kernel outperforms the other kernel based on the shape of the decision boundary.

III. Result And Discussion

After performing all the pre-processing tasks, the dataset is split into train and test subsets with an $80 : 20$ ratio. The configuration space of this experiment consists of 57 training samples. Nine sub-configuration spaces are chosen for that configuration space. The sub-configuration spaces consist of sample size $[25, 41, 57]$ and features $[2, 8, 14]$, each feature denoted by a single qubit. Then F1 score and balanced accuracy are calculated for nine configurations.

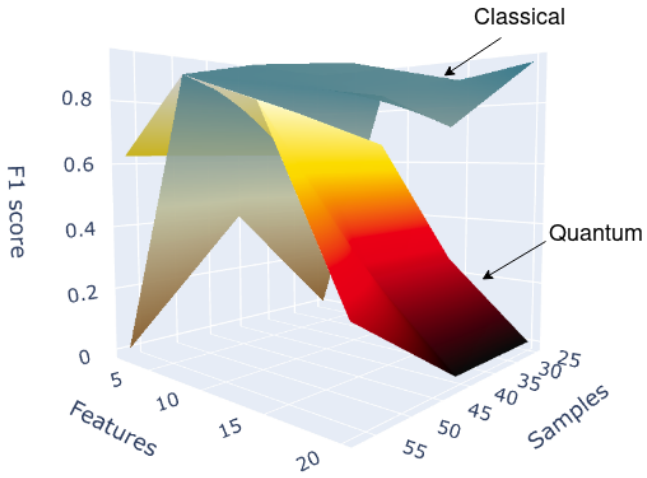

Fig. 1. F1 score of data selected by lasso.

It is observed from the Fig. 1 that, for 14 features and 25 samples, classical kernel has the highest F1 score of $.93$ and for 8 features and 57 samples quantum kernel has $.85$. Classical and quantum surfaces intersect at $(8, 57)$ point. And in Fig.2 for 14 features and 25 samples, classical kernel has the highest F1 score of $.93$ and for 8 features and 41 samples quantum kernel has $.44$. Classical and quantum surfaces intersect at $(8, 41)$ point.

Similarly, from Fig. 3, it is observed that for 20 features and 25 samples classical kernel has the highest balanced accuracy of $.93$ and for 8 features and 57 samples quantum kernel has balanced accuracy of $.86$. Classical and quantum surfaces intersect at $(8, 57)$ point. In Fig. 4, for 8 features and 57 samples classical kernel has the highest balanced accuracy

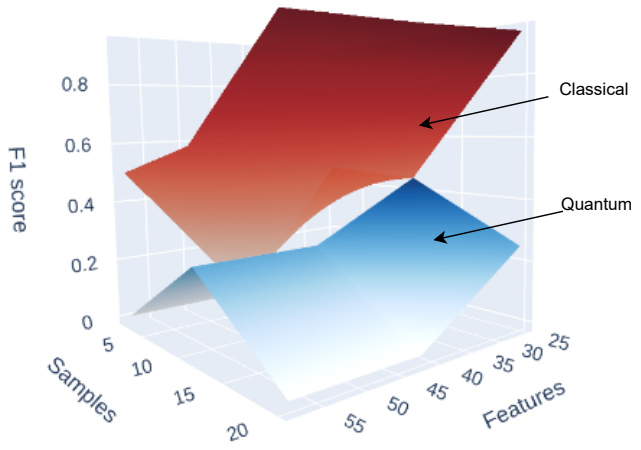

Fig. 2. F1 score of data selected by QUBO.

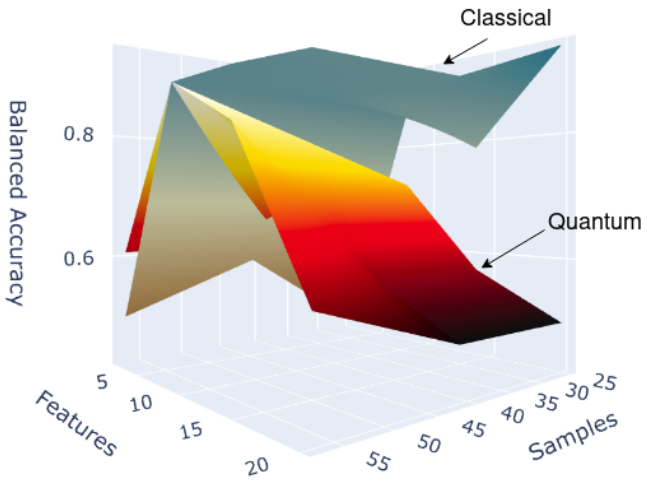

Fig. 3. Balanced accuracy of data selected by lasso.

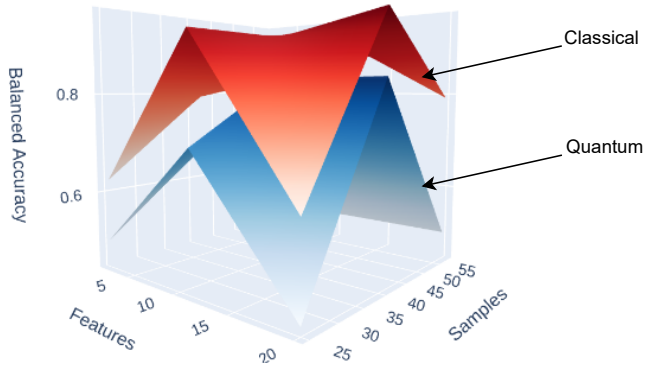

Fig. 4. Balanced accuracy of data selected by QUBO.

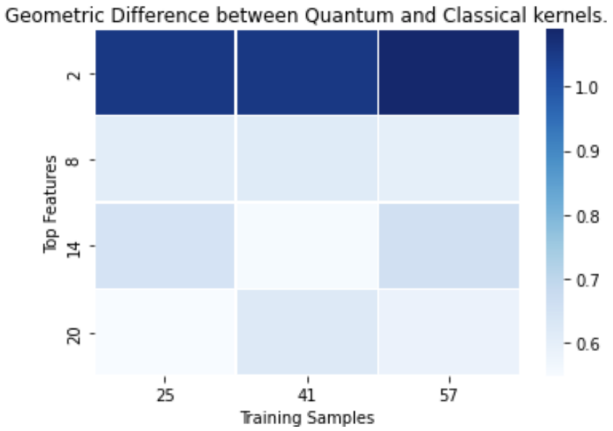

Fig. 5. Heatmap for the geometric difference of both kernels.

of .96 and for 2 features and 41 samples quantum kernel has balanced accuracy of .64. Classical and quantum surfaces intersect at $(2, 41)$ point.

For different configuration spaces, the geometric difference between classical(linear) and quantum kernel(Pauli Z feature map) is analyzed. Fig. 5 shows that potential quantum advantage is more likely in the configuration space with 2 features and 57 samples.

The Phase terrain ruggedness index (PTRI) is a metric that helps to identify the configuration space where quantum advantage is potential for a specific problem. The flattest region in classical landscape helps to consider quantum configuration for that problem to attain privilege over the classical. As the flattest region indicates stagnation of performance, the ruggedness of the quantum landscape helps to get insights if there is any quantum advantage possible.

The ruggedness of the PTRI landscape for the F1 score in Fig.6 indicates a point of advantage for the quantum kernel in the configuration of $(8, 41)$, where the classical kernel performs poorly. Whereas, it can be observed from the PTRI landscape for Balanced Accuracy that classical kernel performs better in configuration of $(8, 25)$ in Fig.7 than the quantum counterpart.

It is also to be considered the PTRI values resulting in calculations with different metrics should indicate which performance metric is better with accuracy for a problem in the analysis of EQA. The experiment has better performance accuracy with F1 score metric facilitating quantum advantage in the configuration of $(8, 25)$. It also shows that balanced accuracy may seem a reasonable metric to proceed initially with much prediction accuracy. But, considering PTRI to evaluate its potentiality towards quantum advantage does provide much insight into its unlikeliness of outperforming.

IV. Quantum Resource Estimation

Proceeding toward solving a problem requires the consideration of the efficiency of the implementation. So it is important to analyze the complexity of an algorithm along with enhancing its performance. For this type of evaluation, resource estimation is vital in the quantum field. Quantum resource estimation is a method to determine the number

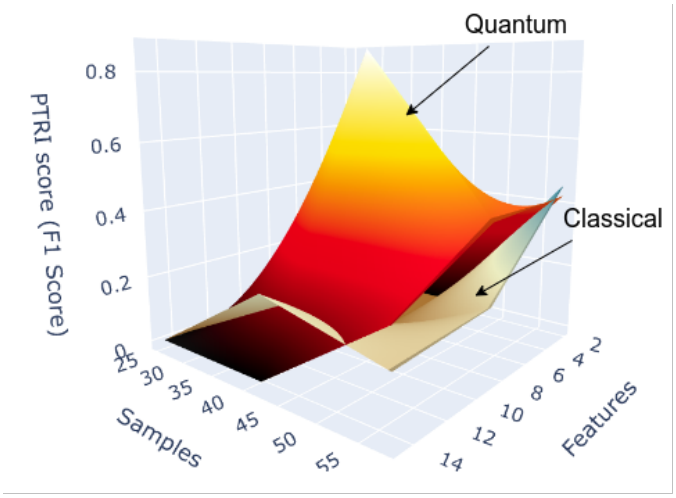

Fig. 6. PTRI landscape for F1 Score of data selected by lasso.

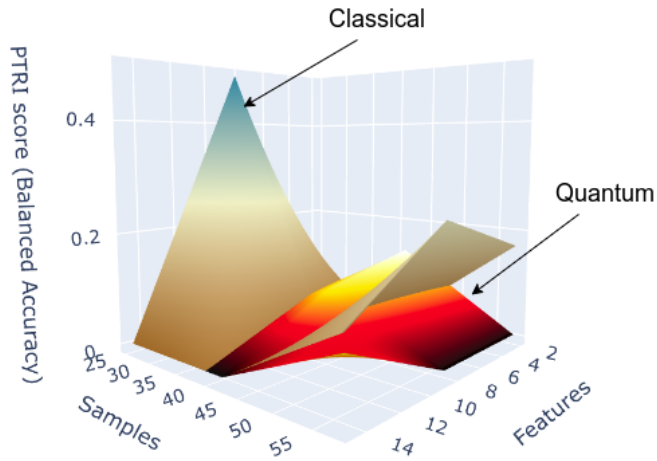

Fig. 7. PTRI landscape for Balanced Accuracy Score of data selected by lasso.

of qubits, unitary gates, quantum processing unit(QPU) utilization, and other resources required for algorithmic implementation.

In the quantum kernel estimation approach, the kernel is estimated using quantum unitary circuit. The quantum kernel transforms the classical state into a quantum state. Then the classical SVM is applied to draw the separating hyperplane among classes. Here, the classical data $\vec{x} \in \Omega$ is converted into a quantum state $|\Phi(\vec{x})\rangle$ by applying the unitary circuit $\mathcal{U}_{\Phi(\vec{x})}$, where the quantum state,

$$|\Phi(\vec{x})\rangle = U_{\Phi(\vec{x})} H^{\otimes n} U_{\Phi(\vec{x})} H^{\otimes n} |0\rangle^{\otimes n}$$

$$U_{\Phi(\vec{x})} = \exp\left(i \sum_{S \subseteq [n]} \phi_S(\vec{x}) \prod_{i \in S} Z_i \right)$$

Considering maps with low-degree expansions where $|S| \leq 2$, two types of feature maps are possible. For $d = 2$, the feature map that results in,

$$U_{\phi\{k,l\}}(\vec{x}) = \exp\left(i\phi_{\{k,l\}}(\vec{x}) Z_k Z_l\right)$$

And the circuit in Fig. 8 is drawn for this mapping for 3 qubits and linear entanglement. This circuit is repeated 2 times.

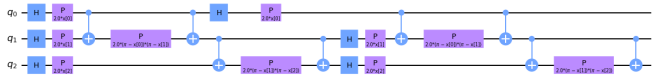

Fig. 8. Circuit for ZZfeaturemap

Using linear entanglement, for different numbers of qubits and repetitions, a depth analysis is shown in the following table,

TABLE I
DEPTH ANALYSIS OF ZZFEATUREMAP WITH LINEAR ENTANGLEMENT.

n \ r	1	2	3	4
2	5	10	15	20
3	8	16	24	32
4	11	19	27	35
5	14	22	30	38
6	17	25	33	41

Here, n is the number of qubits and r is the number of repetitions. After analyzing the above table, it can be concluded that ZZfeaturemap has $\mathcal{O}(5 \times r)$ complexity for $n = 2$, and for $n \geq 3$ it is $\mathcal{O}(8 \times r + 3 \times (n-1))$, both are equivalent to $\mathcal{O}(n)$.
For $d = 1$, the resulting feature map is,

$$U_{\phi\{k\}}(\vec{x}) = \exp\left(i\phi_{\{k\}}(\vec{x}) Z_k\right)$$

And the circuit in Fig. 9 is drawn for this mapping with two qubits and two repetitions.

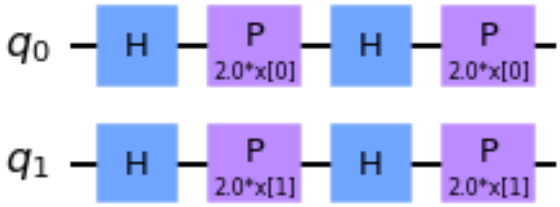

Fig. 9. Circuit for Pauli Z featuremap

Observing the circuit architecture, it is evident that the circuit has $\mathcal{O}(2 \times r)$ which is equivalent to $\mathcal{O}(r)$ where r is the number of repetitions.

Here is the summary of estimated gates for each feature map.

TABLE II
INDIVIDUAL GATE ESTIMATION FOR EACH FEATUREMAP

Unitary gate / Featuremap	Hadamard Gate	Z gate	CNOT gate
ZZfeaturemap	$\mathcal{O}(n \times r)$	$\mathcal{O}(r \times (2 \times n - 1))$	$\mathcal{O}(2 \times (n-1) \times r)$
PauliZfeaturemap	$\mathcal{O}(n \times r)$	$\mathcal{O}(n \times r)$	-

REFERENCES

[1] Yeang, Chen-Hsiang, et al. "Molecular classification of multiple tumor types." ISMB (Supplement of Bioinformatics) 2001 (2001): 316-322.

[2] Maheshwari, Danyal, Begonya Garcia-Zapirain, and Daniel Sierra-Sosa. "Quantum machine learning applications in the biomedical domain: A systematic review." Ieee Access 10 (2022): 80463-80484.

[3] Zhao, Yaxing, Limsoon Wong, and Wilson Wen Bin Goh. "How to do quantile normalization correctly for gene expression data analyses." Scientific reports 10.1 (2020): 15534.

[4] Kang, Chuanze, et al. "Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine." Journal of theoretical biology 463 (2019): 77-91.

[5] Analytics Vidhya, "Feature Selection Techniques in Machine Learning," Analytics Vidhya, Oct. 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/. [Accessed: April 12, 2023].

[6] Ferrari Dacrema, Maurizio, et al. "Towards feature selection for ranking and classification exploiting quantum annealers." Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022.

[7] D-Wave Systems, "D-Wave scikit-learn Plugin," GitHub. [Online]. Available: https://github.com/dwavesystems/dwave-scikit-learn-plugin/tree/main. [Accessed: February 2, 2024].

[8] Krunic, Zoran, et al. "Quantum kernels for real-world predictions based on electronic health records." IEEE Transactions on Quantum Engineering 3 (2022): 1-11.

[9] Sarder, M. Alamgir, Md Maniruzzaman, and Benojir Ahammed. "Feature selection and classification of leukemia cancer using machine learning techniques." Machine Learning Research 5.2 (2020): 18.

[10] Havlíček, Vojtěch, et al. "Supervised learning with quantum-enhanced feature spaces." Nature 567.7747 (2019): 209-212.