Quantum Data Science (STA9075) Final Project Report

2025314249 최윤아

Quantum Support Vector Machines for High-Dimensional Biomedical Data Classification

1. Problem Definition

- We consider the problem of medical data classification under high-dimensional, low sample size and nonlinear conditions. To address this problem setting, we employ a classification model based on quantum kernel methods. Quantum kernels enable an implicit embedding of classical data into high-dimensional Hilbert spaces, providing the potential to represent complex nonlinear structures more effectively. Accordingly, this project aims to clearly define the structural characteristics of medical data and the associated classification challenges, and to investigate the potential advantages of quantum kernel–based approaches over existing classical methods under this problem formulation.
- Throughout this work, we make this following assumptions. First, we assume that the input features are real-valued and have been appropriately preprocessed and normalized prior to quantum data embedding. Second, we assume a supervised learning setting in which labeled data are available, and the primary objective is classification performance rather than causal inference. Finally, we assume access to quantum simulators or near-term quantum devices capable of implementing parameterized quantum feature maps with a limited number of qubits and circuit depth.

2. Motivation and Significance

- Medical AI is a critical field in which classification accuracy directly affects clinical decision-making. Degradation in the performance of diagnostic or predictive models may lead to misdiagnosis or delayed treatment. Therefore, constructing reliable classification models for medical data is an important task with both scientific and societal significance.
- However, medical and biological data pose inherent challenges from a machine learning perspective. Such data are typically high-dimensional, often containing thousands of features, while the number of available samples remains limited.
  · Leukemia microarray dataset : 72 samples with 3,572 gene-expression features
  In addition, medical data are generated through complex interactions among multiple biological factors, resulting in strongly nonlinear and unstructured patterns. These characteristics make high-dimensional embeddings essential for capturing the intrinsic structure of the data.
- These data properties impose significant limitations on classical machine learning classification models when learning stable decision boundaries. In particular, the support vector machine is a representative kernel-based method capable of handling nonlinear classification problems through the kernel trick. Nevertheless, in high-dimensional, low

sample size settings, classical SVMs exhibit several limitations. Previous studies applying classical SVMs to real-world medical datasets, such as leukemia microarray data and Parkinson's disease voice data, have reported constrained classification performance.

- classical kernel-based SVMs also suffer from scalability issues. The construction and training of the kernel matrix typically require computational complexity of at least $O(N^2)$, and in some cases up to $O(N^3)$, where N denotes the number of training samples. Moreover, model performance is highly sensitive to the choice of kernel function, while selecting an appropriate kernel for complex medical data remains a nontrivial task. These factors limit both the scalability and generalization capability of classical SVMs in biomedical applications.
- QSVM extends the classical SVM framework by encoding classical input data into quantum states and constructing kernels via similarity evaluations performed by quantum circuits. In this setting, the classical inner product is replaced by the quantum fidelity between two quantum states, $K_q(x_i,x_j) = |\langle \psi(x_i)|\psi(x_j)\rangle|^2$ where $|\psi(x)\rangle = U_\phi(x)|0\rangle^{\otimes n}$ is a quantum state, $U_\phi(x)$ is a feature map which is a data encoding unitary circuit. As a result, QSVMs offer the potential for enhanced expressive power in representing complex nonlinear structures.
- Recent advances in quantum hardware and simulation platforms have made kernel evaluation using small-scale quantum circuits practically feasible. Investigating whether quantum feature spaces can provide advantages over classical kernel methods in high-dimensional, low sample size medical data settings is therefore an important research.

3. Related Work

[QSVM]

- Havlíček, V. et al. "Supervised learning with quantum-enhanced feature spaces"
  This paper proposes a supervised learning framework that utilizes quantum-enhanced feature spaces, introducing quantum kernel methods and quantum classifiers. Classification is performed in high-dimensional Hilbert spaces through quantum feature maps, and kernel values are estimated using quantum computers. They theoretically and experimentally demonstrate that quantum kernels can achieve greater expressive power than classical kernels. [1]
- Park, J.-E. et. al. "Practical application to Quantum SVM: theory to practice"
  This paper proposed practical improvements to enable the application of QSVM to real world data analysis. By comparing classical SVM and QSVM across datasets of varying complexity, they showed that as data complexity increases, the performance of classical SVM degrades, whereas QSVM maintains consistent performance when appropriate quantum feature maps are regularization are applied. In particular, they experimentally demonstrated that QSVM can outperform classical SVM on datatsets characterized by complex decision boundaries. [2]

[Applied to Biomedical Area]

- Saranya et al., "A Quantum-Based Machine Learning Approach for Autism Detection Using Common Spatial Patterns of EEG Signals"
  This paper proposed a QSVM-based approach for classifying EEG signals to distinguish children with autism spectrum disorder (ASD) from typically developing children. By employing an amplitude embedding, they efficiently encoded high-dimensional EEG feature vectors into $\log_2 N$ qubits, demonstrating the feasibility of processing medical data under limited quantum resources. It provides experimental evidence for the practical applicability of quantum embeddings to high-dimensional biomedical data. [3]
- Walid El Maouaki et al. "Quantum Support Vector Machine for Prostate Cancer Detection: A Performance Analysis"
  This paper applied QSVM to prostate cancer diagnosis and conducted a comparative performance analysis against classical SVM. Using the Kaggle prostate cancer dataset, their results showed that QSVM achieves comparable accuracy to classical SVM while demonstrating superior sensitivity and F1-score, particularly in reducing false negatives. This paper highlights the practical advantage of QSVM in medical diagnostic settings, where high sensitivity is crucial for reliable disease detection. [4]
- Existing studies on QSVM have demonstrated the expressive power of quantum feature spaces and their potential performance advantages. However, many prior studies construct QSVM models using a single dataset or a fixed quantum feature map, and systematic analyses of the relationship between data characteristics and feature map selection remain limited. To address this, we apply QSVM to multiple biomedical datasets (Leukemia, Parkinson, EGFR kinase target data), and conducts comparative experiments using various quantum feature maps. Through this analysis, we provide empirical results on which feature maps yield meaningful performance for different types of data, and aim to clarify the practical applicability of quantum kernel methods in medical data classification tasks.

4. Quantum Advantage Justification

- QSVM is expected to improve upon classical SVM primarily due to its enhanced expressive power and the efficiency of its data embedding mechanism. Medical data often exhibit a high dimensional low sample size structure together with nonlinearity, making expressive capacity a crucial factor for effective classification. However, classical approaches face limitations in designing suitable kernel functions as data complexity increases.
- Quantum kernel methods implicitly embed classical data into high dimensional Hilbert spaces through quantum feature maps, enabling a natural representation of complex nonlinear structures. This property offers the potential for improved performance in problems characterized by complex data distributions, such as medical datasets. In particular, circuit based quantum feature maps can encode high order interactions involving entanglement, which may provide greater expressive power.

- In QSVM, kernel values are computed intrinsically through quantum circuit measurements, allowing the model to bypass the explicit high dimensional inner product calculations required in classical kernel methods. As a result, theoretical studies suggest that the computational complexity of kernel evaluation and classification can be improved from polynomial to logarithmic scaling. Although the experiments in this study are conducted in a limited simulation setting, this computational structure may become a significant advantage when scaling to larger datasets in the future.
- The objective of this project is to examine whether QSVM can maintain competitive classficiation performance compared to classical SVM, or achieve improved performance under specific conditions, in high dimensional medical data settings. In particular, we evaluate the practical applicability of quantum kernel methods by comparing training and test accuracy across different datasets and feature map configurations.

5. Proposed Approach

- In this project, we analyze the classification performance of quantum kernel-based QSVM on medical and biological datasets characterized by high dimensionality and strong nonlinearity. In particular, we conduct comparative experiments across multiple biomedical datasets, focusing on how the choice of quantum feature map affects QSVM performance.

[Overall Methodology]

- First, classical biomedical data are loaded and preprocessed. Principal Component Analysis (PCA) is applied to extract principal features, and the input values are normalized to the range $[-\pi, \pi]$ to match the domain of quantum circuits.
- Next, the data are embedded into quantum states using a selected quantum feature map. Kernel values between data points are then computed via quantum circuits, and the resulting kernel matrix is used to train a QSVM classifier. Finally, the classification performance of QSVM is compared with that of classical SVM based on training and test accuracy.

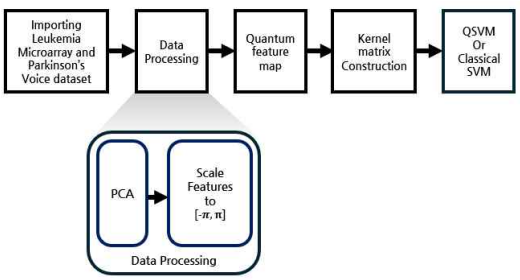

Fig 1. Leukemia, Parkinson dataset QSVM Workflow

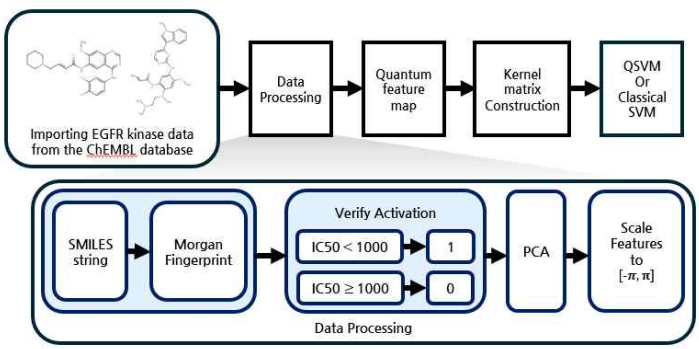

Fig 2. EGFR kinase dataset QSVM Workflow

[Biomedical Datasets]

- This study employs three biomedical datasets with distinct characteristics.
- ① Leukemia Microarray Dataset

  This dataset addresses a cancer classfication problem based on gene expression data and represents a typical biomedical dataset with a high-dimensional, low sample size structure. The Leukemia small dataset from the classic Golub et al. (1999) paper is used. Leukemia microarray dataset is used to distinguish acute lymphoblastic leukemai(ALL) from acute myeloid leukemia(AML). The first row of the dataset contains class labels, while the remaining rows correspond to gene expression features. The labels werer mapped to −1(ALL) and +1(AML).
- ② Parkinson's Voice Dataset

  This dataset is based on voice signal features for neurological disease classification and exhibits strong nonlinear characteristics. The parkinson dataset from OpenML is used. This dataset contains biomedical voice features extracted from individuals with and without Parkinson's disease. The class labels were converted into a binary format, mapping healthy subjects to −1 and Parkinson's patients to +1.
- ③ EGFR Kinase Target Dataset

  This dataset focuses on drug response prediction based on molecular features and contains a mixture of biological and chemical characteristics. EGFR-related drug data are collected from ChEMBL, and SMILES strings are converted into Morgan fingerprints to construct the input feature set. Activity labels are generated based on IC50 values. The class labels were converted into a binary format, mapping active (IC50<1,000 nM) to 1 and inactive (IC50≥1,000 nM) to 0.

[Feature Map Variations]

- To embed classical data into quantum states, several types of quantum feature maps are employed and their performance is compared.
- ① Pauli Feature Map

  The Pauli feature map is a generalized feature map composed of combinations of Pauli operators X, Y, Z. It allows flexible design of rotations and interactions depending on the characteristics of the data.
- ② Z Feature Map

  The Z feature map is a special case of the Pauli feature map and encodes features directly into the phase of quantum states using single-qubit Z-rotation gates. It has a simple circuit structure without entanglement, resulting in shallow circuit depth and improved robustness to noise.
- ③ ZZ Feature Map

  The ZZ feature map is another special case of the Pauli feature map and generates entanglement between qubits using $Z \otimes Z$ interaction gates. This feature map is well suited for capturing pairwise interactions between features and effectively represents complex dependencies in data such as biomedical signals and gene expression profiles.

[Performance Evaluation]

- The performance of QSVM is evaluated based on training accuracy and test accuracy. For each dataset, QSVM models using different quantum feature maps are compared, and their performance is analyzed against that of a classical SVM on the same data.

6. Experiments Results (Comment #4)

- ① Leukemia Microarray Dataset

The input features were standardized using StandardScaler, followed by PCA-based dimensionality reduction to a maximum of 10 principal components. The PCA outputs were then resclaed to the range $[-\pi, \pi]$ to match the input domain of quantum circuits. The dataset was split into training and test sets with an 80/20 ratio, preserving class balance. For the leukemia dataset, each sample was represented using 3,571 gene expression features. After standard scaling, PCA was applied while retaining all 10 principal components, which explained nearly 100% of the variance. The QSVM employed 10 qubits, corresponding to the feature dimension, and the number of feature map layers was controlled via the reps parameter, ranging from 1 to 3.

Classical SVM Result – Leukemia dataset		
Kernel	Train Accuracy	Test Accuray
Linear	0.68	0.67
Poly	0.72	0.67
RBF	0.72	0.67
Sigmoid	0.60	0.67

Quantum SVM Result – Leukemia dataset			
Feature map	reps	Train Accuracy	Test Accuray
Z	1	0.81	0.53
Z	2	0.86	0.53
ZZ	1	1.00	0.67
ZZ	2	1.00	0.67

Table 1. Result of Classical / Quantum SVM in Leukemia dataset

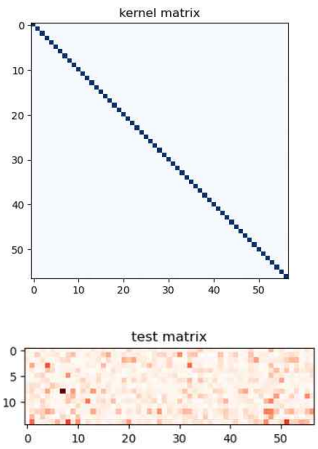

Fig3. Kernel and test matrix of Best case in QSVM (ZZ, reps=1)

Table 1 summarizes the classification performance of classical SVM and QSVM on the leukemia dataset. Among classical SVM models, the best-performing kernels achieve a test accuracy of approximately 0.67, indicating limited generalization improvement under the high-dimensional, small-sample setting. In contrast, QSVM performance depends strongly on the quantum feature map. The Z feature map yields relatively high training accuracy (up to 0.86) but lower test accuracy (0.53), suggesting overfitting. The entangling ZZ feature map achieves perfect training accuracy (1.00) while maintaining a test accuracy of 0.67, comparable to the best classical SVM results.

Increasing the number of repetitions (reps) does not further improve test performance for either Z or ZZ feature maps, suggesting that shallow quantum circuits are sufficient for this dataset. Overall, these results indicate that QSVM with entangling feature maps can achieve competitive generalization performance relative to classical SVM, while highlighting the importance of feature map selection in avoiding overfitting.

- ② Parkinson's Voice Dataset

Except for the change in the dataset, the remaining experimental procedures were conducted following the same setup as described for the leukemia dataset. For the Parkinson's dataset, the original data consisted of 195 samples with 22 features. To address class imbalance, a balanced subset of 90 samples (45 per class) was selected. After standard scaling, PCA was applied to reduce the feature dimension to 15, retaining approximately 99.9% of the total variance. The dataset was then split into 60 training samples and 30 test samples. The QSVM employed 15 qubits, corresponding to the PCA-reduced feature dimension, and the number of feature map layers was controlled via the reps parameter, which was varied across experiments.

Classical SVM Result – Parkinson dataset		
Kernel	Train Accuracy	Test Accuray
Linear	0.85	0.67
Poly	0.98	0.73
RBF	0.93	0.70
Sigmoid	0.72	0.70

Quantum SVM Result – Parkinson dataset			
Feature map	reps	Train Accuracy	Test Accuray
Z	1	0.95	0.73
Z	2	0.95	0.80
Z	3	0.87	0.60
ZZ	2	1.00	0.77
X, Z, ZZ	2	1.00	0.67
X, Y, Z	2	0.90	0.63
X, Y, Z, ZZ	2	1.00	0.50
Z, ZZ, XX	2	1.00	0.63

Table 2. Result of Classical / Quantum SVM in Parkinson dataset

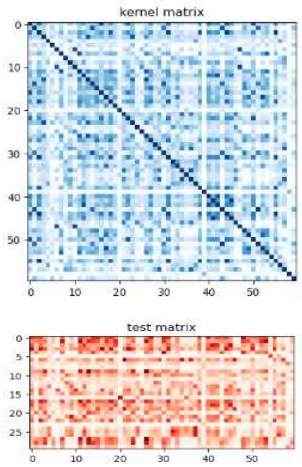

Fig4. Kernel and test matrix of Best case in QSVM (Z, reps=2)

Table 2 presents the classification results of classical SVM and QSVM on the Parkinson's dataset. Among classical SVM models, the best test accuracy ranges between 0.67 and 0.73, indicating limited generalization performance despite high training accuracy. In contrast, QSVM performance depends strongly on the quantum feature map and circuit depth. Using the Z feature map, QSVM achieves the highest test accuracy of 0.80 at reps = 2, outperforming all classical SVM baselines. Increasing the circuit depth further leads to reduced test performance, suggesting overfitting. Entangling ZZ feature maps also show competitive performance, with test accuracy up to 0.77, while more expressive Pauli-based feature maps achieve perfect training accuracy but poor generalization.

Overall, the Parkinson dataset results indicate that QSVM can outperform classical SVM under appropriate feature map configurations, while also emphasizing that careful feature map and circuit-depth selection is critical to avoid overfitting. These findings further support the role of QSVM as a competitive and flexible alternative to classical kernel methods in biomedical classification tasks.

- ③ EGFR Kinase Target Dataset

Except for the change in the dataset and Preprocessing step, the remaining experimental procedures were conducted following the same setup as described for the Parkinson dataset. For the EGFR kinase target dataset, the original data consisted of 24,344 samples with 2,048-dimensional Morgan fingerprint features, obtained from the ChEMBL database. To control class imbalance and computational cost, a balanced subset of 120 samples was constructed, with 60 active and 60 inactive compounds based on an IC50 threshold of 1000 nM. The dataset was then split into 60 training samples and 30 test samples. QSVM employed 20 qubits, corresponding to the PCA-reduced feature dimension. The number of feature map layers was controlled via the reps parameter, which was varied across experiments to adjust the circuit depth.

Classical SVM Result – EGFR Kinase Target dataset		
Kernel	Train Accuracy	Test Accuray
Linear	0.78	0.73
Poly	0.93	0.67
RBF	0.82	0.77
Sigmoid	0.47	0.50

Quantum SVM Result – EGFR Kinase Target dataset			
Feature map	reps	Train Accuracy	Test Accuray
Z	1	0.87	0.73
Z	2	0.80	0.63
Z	3	0.90	0.63
ZZ	1	0.95	0.57
ZZ	2	1.00	0.63
ZZ	3	1.00	0.63
Z, ZZ	2	1.00	0.73
Z, ZZ	2	1.00	0.70

Table 3. Result of Classical / Quantum SVM in EGFR Kinase Target dataset

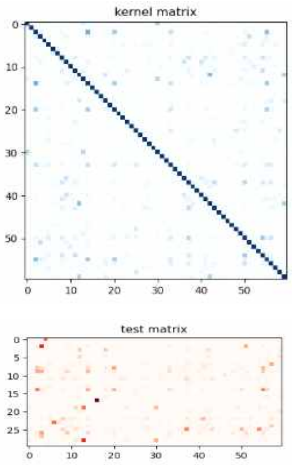

Fig5. Kernel and test matrix of Best case in QSVM (Z, ZZ reps=2)

Table 3 reports the classification performance of classical SVM and QSVM on the EGFR kinase target dataset. Among classical SVM baselines, the best test accuracy is achieved by the RBF kernel (0.77), while other kernels show comparable or inferior performance.

QSVM performance depends strongly on the quantum feature map and circuit depth. Non-entangling Z feature maps achieve moderate test accuracy (up to 0.73), comparable to classical linear and polynomial kernels. In contrast, purely entangling ZZ feature maps exhibit perfect training accuracy but poor test accuracy (0.57–0.63), indicating overfitting. The best QSVM performance is obtained using combined Z–ZZ feature maps with moderate circuit depth (reps = 2), achieving a test accuracy of 0.73 with stable generalization.

Overall, the EGFR results show that QSVM can achieve competitive performance with classical SVM when appropriately configured, while highlighting that excessive circuit expressivity does not necessarily improve generalization in chemically complex, small sample datasets.

7. Feasibility and Limitations

- **Feasibility**

  The proposed QSVM approach is feasible using the Qiskit Aer simulator, which provides quantum feature map library and supports classical data preprocessing and loading. The method follows a hybrid framework, where quantum circuits are used only for kernel estimation, while SVM training and classification are performed classically. This design makes the approach practically implementable in the NISQ era. Previous studies indicate that quantum feature maps can achieve competitive performance on high-dimensional, small-sample biomedical datasets, motivating the application of QSVM in this work.

- **Limitations**

  QSVM implementations are constrained by circuit depth, quantum noise, and qubit requirements, as the number of qubits scales with the feature dimension. Even in simulation, kernel matrix estimation is computationally expensive. (using approximately 19 features required several thousand minutes of runtime.) These limitations make large-scale experiments and real-hardware execution challenging, highlighting the need for embedding optimization and error mitigation techniques.

8. Future works

- **① Neural Quantum Embedding Based Extension** (Comment #1)

  Our experimental results show that QSVM performance is highly sensitive to the choice of quantum feature map and circuit depth, and that increased circuit expressivity often leads to overfitting rather than improved generalization. This means that the quality of data embedding into the quantum feature space is a more critical factor. To address this issue, we can introduce Neural Quantum Embedding (NQE), which augments quantum feature maps with a trainable classical neural network. [5]

  The generalization performance of quantum machine learning models is governed not by circuit depth or parameter count, but by the classification margin in the quantum feature space. In particular, this margin is directly related to the trace distance between class-dependent quantum state ensembles, indicating that maximizing class separability at the embedding stage is a key principle of quantum kernel methods. [6]

  In NQE, the embedding is defined as $\Phi_{NQE} : x \mapsto |x\rangle = V[g(x,w)]|0\rangle^{\otimes n}$, where V is a general quantum-embedding circuit and g is a classical neural network that transforms the input data x using r trainable parameters. By learning a data-dependent transformation prior to quantum embedding, NQE aims to increase class separability directly in the quantum feature space. The training objective of NQE is motivated by maximizing the trace distance between quantum state ensembles corresponding to different classes. Since direct computation of the trace distance is expensive, an implicit fidelity-based loss is used. $l_{fid}[(x_i, y_i), (x_j, y_j)] = [|\langle x_i|x_j\rangle|^2 - 1/2(1 + y_i y_j)]^2$. This loss encourages enhancing class separation in the quantum feature space.

NQE is particularly relevant for entangling feature maps such as ZZ embeddings, which are commonly used due to their expressive power but often suffer from heuristic parameter choices and overfitting. In conventional QSVM implementations, the functions $\phi_i(x)$ and $\phi_{i,j}(x)$ are fixed and manually chosen. As future work, NQE can be integrated into the QSVM pipeline in place of PCA to mitigate overfitting, improve generalization, and increase kernel value variance, thereby reducing the cost of quantum kernel estimation.

- **② Extension to Multi-Class Classification** (Comment #2)

The proposed QSVM-based framework can be naturally extended beyond binary classification to multi-class problems by adopting standard strategies from classical support vector machines, in particular the one-vs-rest (OvR) scheme. In this approach, a set of binary QSVM classifiers is trained, each distinguishing one target class from all remaining classes. During inference, the final prediction is obtained by selecting the class with the highest decision score among all binary classifiers. Such OvR-based multi-class extensions of QSVM have been demonstrated in recent work. [7]

As future work, the binary QSVM pipeline can be extended using the OvR strategy to handle multi-class biomedical classification tasks. This direction enables the investigation of how quantum kernel expressivity and computational cost scale with the number of classes, while remaining compatible with near-term quantum devices and simulators.

- **③ Extension to Imbalanced Dataset** (Comment #3)

In medical and biological data, the ratio between disease-positive and normal samples is generally imbalanced. This class imbalance can form a decision boundary favorable to majority classes in classical SVMs, which can reduce sensitivity for minority classes.

QSVM provides the possibility of effectively separating fine but important structural differences contained in minority classes by embedding data into high-dimensional Hilbert space through a quantum feature map. In particular, since the quantum feature map containing entanglement can express interactions between features, it can emphasize complex patterns that exist only in minority classes. In addition, since kernel is defined through quantum circuit design, it provides the possibility of effectively separating by introducing weighted kernel that reflects class-specific importance or using custom feature maps specialized for data structure.

However, QSVM does not automatically solve the class imbalance problem, and existing adjustment strategies such as data sampling or cost-sensitive learning are still needed. Nevertheless, the high expressive power of the quantum feature space implies that QSVMs can have potential benefits in Imbalanced high-dimensional medical data environments.

8. Reference

[1] Havlíček, V., Córcoles, A.D., Temme, K. et al. Supervised learning with quantum-enhanced feature spaces. Nature 567, 209–212(2019).https://doi.org/10.1038/s41586-019-0980-2

[2] Park, J.-E., Quanz, B., Wood, S., Higgins, H., & Harishankar, R., “Practical application to Quantum SVM: theory to practice”, pp. 1-9, 2020. https://arxiv.org/abs/2012.07725

[3] S. Saranya and R. Menaka, "A Quantum-Based Machine Learning Approach for Autism Detection Using Common Spatial Patterns of EEG Signals," in IEEE Access, vol. 13, pp. 15739-15750, 2025, doi: 10.1109/ACCESS.2025.3531979.

[4] Walid El Maouaki et al. “Quantum Support Vector Machine for Prostate Cancer Detection: A Performance Analysis”, https://arxiv.org/pdf/2403.07856

[5] Tak Hur, Israel F. Araujo, Daniel K. Park. “Neural quantum embedding: Pushing the limits of quantum supervised learning”, Phys. Rev. A 110, 022411 (2024), doi: 10.1103/PhysRevA.110.022411

[6] Hur, T. & Park, D.K.. (2025). “Understanding Generalization in Quantum Machine Learning with Margins”, Proceedings of the 42nd International Conference on Machine Learning 267:26338-26360 Available from https://proceedings.mlr.press/v267/hur25a.html.

[7] Gabriela Pinheiro, Donovan, M. Slabbert, Luis Kowada, Francesco Petruccione, “Quantum kernel and HHL-based support vector machines for multi-class classification”, https://doi.org/10.48550/arXiv.2509.10190