Quantum and Classical Kernels Applied to Classification of Unbalanced Datasets

Fernando Mondragon-Sampedro and Diego Alvarez-Estevez

Centro de Investigación CITIC, Universidade da Coruña, 15071 A Coruña, Spain
Correspondence: fernando.mondragon@udc.es

DOI: *https://doi.org/10.17979/spu.23.c28*

Abstract: This work investigates quantum kernel methods for classification on unbalanced datasets. The study considers the effect of techniques commonly used in classical machine learning, such as kernel centering and cost-sensitive learning, when applied to quantum models. Experiments are conducted on five datasets, including MNIST-1D, the real-world dataset Glass6, and synthetic datasets tailored for quantum classifiers. Results indicate that quantum kernel methods can address unbalanced classification tasks albeit none of the techniques studied showed a statistically significant improvement of the scoring metrics. Results also highlight the influence of hyperparameterization and in particular, quantum bandwidth is identified as a critical hyperparameter for the performance of quantum models.

1 Introduction

The steadfast growth of Quantum Computing has fueled the search for fields where its unique properties can provide an advantage. Such is the field of Quantum Machine Learning (QML). In this work, we explore the application of QML to solve classification tasks on unbalanced datasets. These datasets are prevalent in critical domains such as medical diagnosis or fraud detection, motivating the development of more effective solutions.

With that goal in mind, we carry out a benchmarking study comparing the performance of classical support vector machines (SVM) with that of quantum support vector machines (qSVM).We investigate the impact of techniques such as kernel centering and cost-sensitive learning, both individually and in combination, to improve the base models. Furthermore, this study aims to delve deeper the quantum kernel methods, focusing on the model's hyperparameters, such as the feature maps used to encode the data into quantum states, and the scaling factor, which tries to concentrate the data in the induced space to facilitate linear separation.

2 Methods

This section introduces the theoretical background of the study. It covers quantum kernels, quantum bandwidth hyperparameter, and the techniques studied to improve classifiers on unbalanced data.

2.1 Quantum kernels

Quantum kernel methods rely on the idea of embedding classical data into a high-dimensional Hilbert space through the use of parameterized quantum circuits that act as feature maps (Havlíček et al., 2019).

For this work we used two quantum feature maps, namely the *ZZFeatureMap* and the *CovariantFeatureMap*.

The *ZZFeatureMap* is a general-purpose and widely-used feature mapping that results in relatively shallow quantum circuits, ideal to execute on current quantum hardware. In addition, it has been conjectured to be hard to estimate classically (Havlíček et al., 2019).

The *CovariantFeatureMap* has been designed to solve the synthetic labeling cosets with error (LCE) problem discussed in Glick et al. (2024). It can provide a provable quantum advantage over classical machine learning algorithms for problems with specific symmetries or group structure. More specifically, in this work we use the modified circuit described in Alvarez-Estevez (2025).

2.2 Kernel bandwidth

This section introduces the quantum bandwidth hyperparameter for quantum kernels, recently developed in (Shaydulin and Wild, 2022). This hyperparameter acts as a scaling factor ($\lambda \in [0, 1]$) on the input data, concentrating it into certain regions of the induced Hilbert space so that they can be easier to separate linearly in that space (Canatar et al., 2022). In practice, models with bigger bandwidths tend to underfit while models with smaller bandwidth are prone to overfiting. The usefulness of such hyperparameter remains under debate, as recent studies like Slattery et al. (2023) points at a trade-off as models lose their quantum expressiveness and become easier to simulate classically.

2.3 Kernel Centering

Kernel centering is a technique used to adjust the kernel values in the transformed space with the objective to normalize the data, thereby mitigating the influence of a global phase as a consequence of class imbalance. It is commonly employed when working with non linear kernels (Schölkopf et al., 1998).

Kernel centering is defined as follows.

$$K_c = K - 1_n K - K 1_n + 1_n K 1_n \tag{19.1}$$

This adjustment helps to reduce bias when samples are not symmetrically distributed. As a consequence of that, models tend to generalize better, reducing overfitting. Moreover, models become more stable and less sensitive to little variations in the input data. It's important to note that the same transformation applied to the training kernel must be applied to the test kernel, such that:

$$K_{c\ test}(X, Y) = K_{test} - 1'_n K_{train} - K_{test} 1_n + 1'_n K_{train} 1_n \tag{19.2}$$

2.4 Cost-sensitive learning

Cost-sensitive learning addresses the problem of working with unbalanced datasets by changing the SVM loss function to take into account different penalties when penalizing misclassification of each class (Ling and Sheng, 2011). To achieve that, it suffices to modify the constraints of standard SVM loss function,

$$\max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right) \tag{19.3}$$

$$Subject\ to: \quad 0 \leqslant \alpha_i \leqslant C, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

resulting in:

$$\max_{\alpha} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \right) \tag{19.4}$$

$$Subject\ to: \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leqslant \alpha_i^+ \leqslant C^+, \quad 0 \leqslant \alpha_i^- \leqslant C^-, \quad \forall i = 1, \ldots, l$$

This leaves open the selection of C^+ and C^-. For the study at hand, we have chosen a balanced aproach where $C^{\pm}$ is assigned by multiplying each class frecuency by parameter C.

3 Experiments

To study the aspects mentioned in the section before, we conducted a benchmarking study comparing the quantum kernels to their classical counterparts. Specifically we train SVMs with with *linear*, *polynomial* and *radial basis function* (*RBF*) kernels. Adittionally we included a Logistic Regression classifier for further reference. All models were fitted by exploring the regularization parameter C in the range [0.01, 0.1, 1, 10, 100]. For polynomial kernels the selection of the degree was given by the following rule of thumb: set k to the closest natural number such that $\binom{n}{k} \leqslant 2^{2n}$, i.e., the space of density matrices for an n-qubit system. This hyperparameter was chosen to match the dimensionality of the induced space achieved by the quantum models. For RBF models, a range of [0.001, 0.001, 0.01, 1, 10] was tested to select the *gamma* parameter and for the quantum bandwidth, a range of scaling factors [0.001, 0.01, 0.1, 0.5, 1.0] was explored.

The selection of datasets required particular care as they needed to be meaningful , as discussed by Bowles et al. (2024), while also being sufficiently simple to fit the limitations of quantum computing simulations. The real datasets selected where the MNIST-1D (Greydanus and Kobak, 2020) which was transformed into 2 simplified versions by performing PCA and the Glass6 (Purnama, 2023), also simplified in the same way. Alongside the real datasets selected, we also run the experiment over some ad-hoc quantum datasets created to be easily classified by each one of the quantum feature maps. To evaluate the kernel centering and cost-sensitive techniques, four different experimental setups have been designed: baseline models, applying each of the techniques individually and applying both simultaneously. Each experiment followed the same procedure; first, we sampled the corresponding dataset to get an unbalanced subset to work with, then we trained and evaluated the models, with the appropriate techniques applied depending on the experiment iteration and model (e.g.,Logistic Regression is not compatible with kernel centering). Table 1 summarizes the characteristics of the samples taken. For a more detailed explanation of each dataset, see Fernando Mondragón (2025).

Cohen's Kappa was the evaluation metric selected to score the models as it takes into account the class imbalance. It does so by considering the probability of correctly guessing each class by chance.

Each experiment was then performed 30 times to gain a statistically significant result.

Table 1: Sample size for each run of the experiment and the class imbalance used.

Dataset	**Training sample size**	**Testing sample size**	**Class imbalance**
MNIST-1D-PCA-4	250	100	85%
MNIST-1D-PCA-8	250	100	85 %
Adhoc-ZZ	100	100	85%
Adhoc-COV	52	23	85%
Glass6-PCA-8	149	65	$\approx$ 86.5%

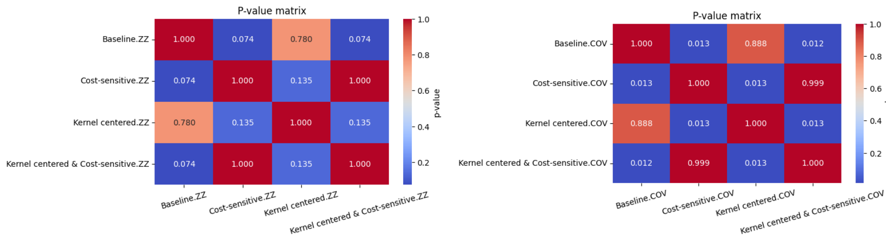

Figure 2: Paired t-test results (a) the *ZZFeatureMap* and (b) the *CovariantFeatureMap*.

4 Results

Figure 1 reports the average kappa for each experiment grouped by model. Due to space constraints, the results presented here aggregate the outcomes across all datasets. For a more detailed dataset-by-dataset analysis, refer to (Fernando Mondragón, 2025).

The results indicate that the kernel centering technique does not exhibit an improvement in the achieved performance, regardless of the model. For the cost-sensitive technique, the results are marginally better for the *CovariantFeatureMap* and the *polynomial* kernels, albeit the other models decreased their scores.

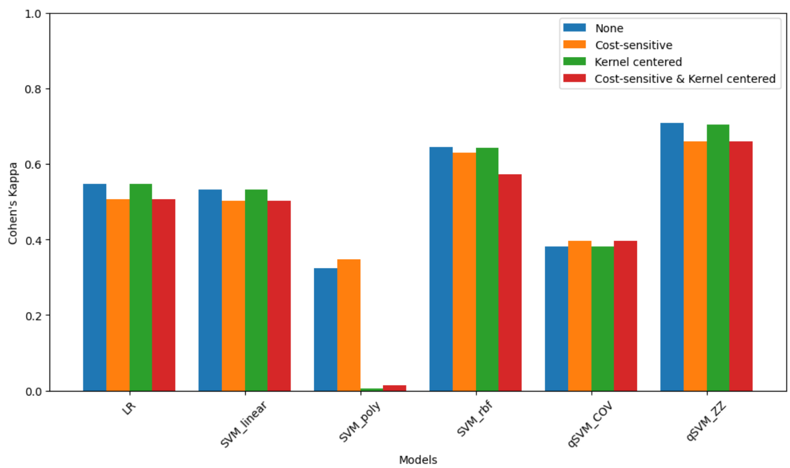

Figure 1: Average Kappa for each technique grouped by model.

Paired t-tests were performed to assess the influence of the techniques, and Figure 2 shows the p-values for each of the quantum models. The results indicate that none of the techniques resulted in significant changes in the achieved performance using a significance threshold of $p < 0.05$ as a reference.

Figure 3 represents the performance of the models, this time grouped by the explored technique. As shown in Figure 3, quantum models using the *ZZFeatureMap* outperform the rest, being closely followed by the classical models with the *RBF* kernel. The differences between the *ZZFeatureMap* and *CovariantFeatureMap* highlight the importance of choosing the right feature map to encode the data.

When examining the experiments without considering the quantum bandwidth (i.e. scaling factor $\alpha = 1$), quantum models showed no ability to learn. This is shown in Figure 4. These findings highlight the importance of tuning this hyperparameter in the training of quantum

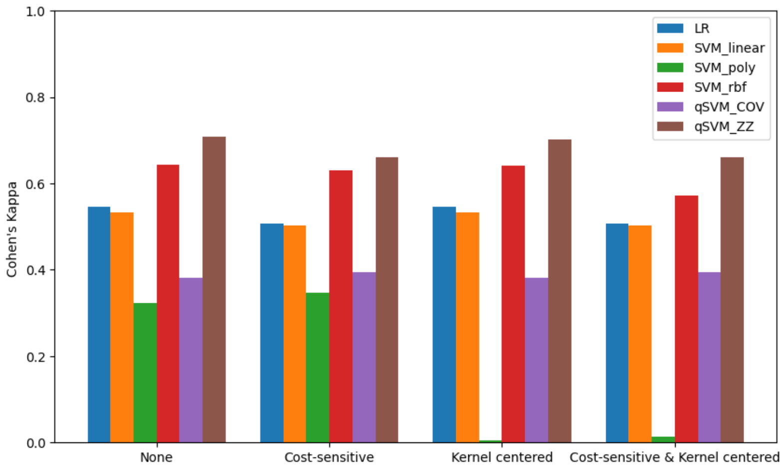

Figure 3: Average Kappa for each model and technique.

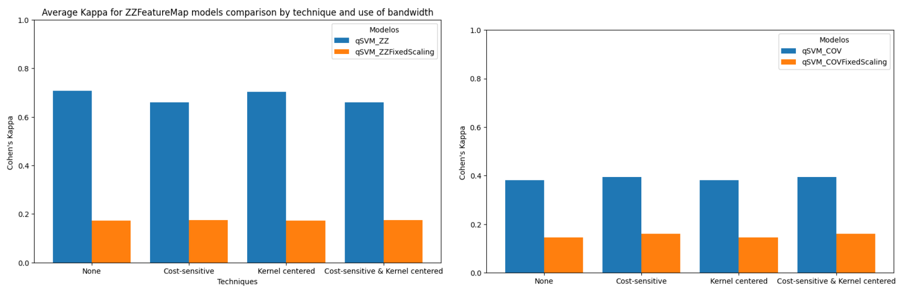

Figure 4: Performance of the quantum models with and without considering hyperparameterization of the kernel bandwidth. In (a) the results obtained by the *ZZFeatureMap*. In (b) the results for the *CovariantFeatureMap*.

models.

5 Conclusions

Although quantum kernels have proven competitive in the classification of unbalanced datasets, the techniques studied did not yield improvements in the overall classification score. Further research is needed to better understand these results, as the techniques also failed to improve the results with classical models. We cannot discard a possible bias in our results due to the limited sample sizes used for each dataset, which was constrained by the available computational resources to perform the quantum simulations. Future work could focus on datasets where the techniques are known to improve classical models, to properly assess their applicability to quantum models.

Importantly, our results have shown that the use of quantum bandwidth as a hyperparameter in quantum kernels can give place to significant improvements in the achieved classification. This indicates that, even with the caveat of limiting the ”quantumness” of the models, the trade-off can, and in this case does, result in better solutions, highlighting the importance of fine-tuning the value of this hyperparameter.

Acknowledgments

This study has been supported by project RYC2022-038121-I, funded by MCIN/AEI/10.13039/501100011033 and European Social Fund Plus (ESF+), project PID2023-147422OB-I00 funded by MCIU/AEI/10.13039/501100011033 and by the European FEDER program, and by project ED431F 2025/35 from Xunta de Galicia.

Bibliography

D. Alvarez-Estevez. Benchmarking quantum machine learning kernel training for classification tasks. *IEEE Transactions on Quantum Engineering*, 2025.

J. Bowles, S. Ahmed, and M. Schuld. Better than classical? the subtle art of benchmarking quantum machine learning models. *arXiv preprint arXiv:2403.07059*, 2024.

A. Canatar, E. Peters, C. Pehlevan, S. M. Wild, and R. Shaydulin. Bandwidth enables generalization in quantum kernel models. *arXiv preprint arXiv:2206.06686*, 2022.

Fernando Mondragón. Estudio comparativo de métodos de aprendizaje máquina basados en kernels cuánticos y clásicos aplicados a problemas de clasificación sobre conjuntos de datos desbalanceados. RUC *https://hdl.handle.net/2183/45578*, 2025.

J. R. Glick, T. P. Gujarati, A. D. Corcoles, Y. Kim, A. Kandala, J. M. Gambetta, and K. Temme. Covariant quantum kernels for data with group structure. *Nature Physics*, 20(3):479–483, 2024.

S. Greydanus and D. Kobak. Scaling down deep learning with mnist-1d. *arXiv preprint arXiv:2011.14439*, 2020.

V. Havlíček, A. D. Córcoles, K. Temme, A. W. Harrow, A. Kandala, J. M. Chow, and J. M. Gambetta. Supervised learning with quantum-enhanced feature spaces. *Nature*, 567(7747):209–212, 2019.

C. X. Ling and V. S. Sheng. Cost-sensitive learning. In *Encyclopedia of machine learning*, pages 231–235. Springer, 2011.

B. Purnama. Glass imbalanced. *https://www.kaggle.com/datasets/baguspurnama/glass-imbalanced*, 2023. Accessed: 2025-05-09.

B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

R. Shaydulin and S. M. Wild. Importance of kernel bandwidth in quantum machine learning. *Physical Review A*, 106(4):042407, 2022.

L. Slattery, R. Shaydulin, S. Chakrabarti, M. Pistoia, S. Khairy, and S. M. Wild. Numerical evidence against advantage with quantum fidelity kernels on classical data. *Physical Review A*, 107(6):062417, 2023.