dc-qml: data-centric quantum machine learning

E. S. Flores[1], K. D. R. C. Barbosa[1], B. V. Guterres[2], S. S. C. Botelho[1], M. R. Pias[1]
(1) Universidade Federal do Rio Grande – FURG; (2) Technological University of Uruguay – UTEC

INTRODUCTION and MOTIVATION

Quantum machine learning (QML) has traditionally been developed from a model-centric perspective with a focus on circuit design and algorithm optimization [1]. However, current hardware is constrained by a limited number of qubits. Circuit depth and noise tolerance also remain restricted [2]. These constraints increase the importance of data preparation before quantum processing. High-dimensional datasets should often be preprocessed before quantum encoding. The selected encoding strategy also affects circuit depth, qubit usage and model performance [3]. Data representation can therefore be improved to reduce computational cost without changing the quantum model. A data-centric perspective is thus needed to complement the prevailing model-centric approach.

DATA-CENTRIC QML DESIGN: MODELS and DATA

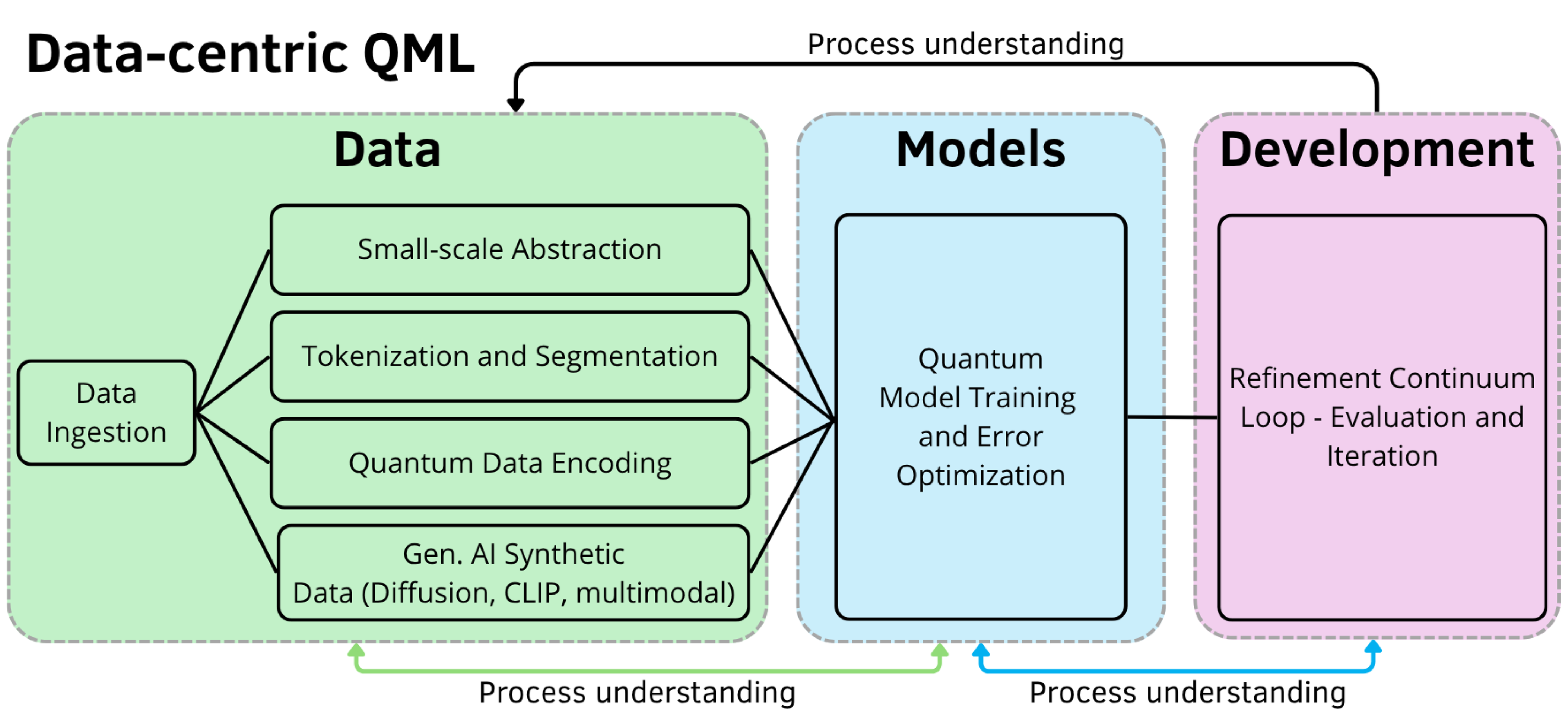

Figura 1: *Iterative data-centric quantum machine learning workflow.*

The *dc-qml* workflow (Fig. 1) links data ingestion with tokenization. Encoding and synthetic data generation are then integrated into model training and evaluation. Results are returned to the data preparation stage through a continuous refinement cycle, rather than being used only to update model parameters.

MAIN CONTRIBUTIONS

We introduce *dc-qml* as a data-to-model workflow in which data quality and representation are treated as main elements of QML development. Both are refined iteratively alongside model optimization.

PRELIMINARY RESULTS

Two related studies provide results for the *dc-qml* thesis using the SKAB benchmark for industrial anomaly detection. First, a factorial analysis of quantum (qIT) and classical (RBF) kernels shows that F1-Score varies non-monotonically with encoding resolution for both kernel types. Encoding should therefore be treated as an experimental rather than a fixed parameter [7a]. Second, detector behavior was found to depend on data characteristics. In a broader comparison with reconstruction-based baselines, the quantum kernel was the only non-degenerate detector in the most topologically complex scenario (F1 $= 0.733$). The autoencoder-based models failed to detect anomalies in this scenario [7b]. Together, these results provide initial indications that performance is determined by data properties as well as model choice.

FUTURE WORK

Open questions are still valid regarding standardized evaluation protocols. Further work is also needed on synthetic data validation and bias control. Encoding strategies should also be coordinated with hardware constraints [4]. Multiprocessor hybrid environments [6] and model trainability limits such as barren plateaus [5] should be at the top of the research issues.

CONCLUSION

dc-qml shifts the focus of QML from model optimization alone to an integrated data-to-model workflow. Preliminary results show that data preparation affects scalability and behavior under NISQ constraints. Future work will validate the workflow in industrial QML applications.

REFERENCES

[1] Hong, Y.-Y.; Lopez, D. J. *A Review on Quantum Machine Learning in Applied Systems and Engineering*. IEEE Access, 13:144607–144631, 2025.

[2] Yang, Z. et al. *A Survey of Important Issues in Quantum Computing and Communications*. IEEE Communications Surveys & Tutorials, 25(2):1059–1094, 2023.

[3] Raisa, M. et al. *A Parallel Quantum Feature Encoding Scheme for Effective Classical Data Classification in Quantum Convolutional Neural Networks*. Proc. IEEE TENCON, 2023.

[4] Di Meglio, A. et al. *Quantum Computing for High-Energy Physics: State of the Art and Challenges*. PRX Quantum, 5(3):037001, 2024.

[5] McClean, J. R. et al. *Barren Plateaus in Quantum Neural Network Training Landscapes*. Nature Communications, 9:4812, 2018.

[6] Carrera Vazquez, A. et al. *Combining Quantum Processors with Real-Time Classical Communication*. Nature, 636(8041):75–79, 2024.

[7a] Barbosa, K. D. R. C. et al. (a) *Data-centric Sensitivity Analysis of Quantum Kernels for Industrial Time Series*. CBCTQ, 2026, to appear

[7b] Barbosa, K. D. R. C. et al. (a) *Unsupervised Quantum Kernels for One-Class Fault Detection in Industrial Time Series*. IEEE Quantum Week (QCE), 2026, to appear

INSTITUTIONAL SUPPORTERS