Rule-Based Explainability for Quantum Machine Learning in Thyroid Cancer Prediction

Daniel Martín-Pérez[1*], Francesc Rodríguez-Díaz[1], Angela del Robledo Troncoso-García[1], David Gutiérrez-Avilés[2], Alicia Troncoso[1], and Francisco Martínez-Álvarez[1]

[1] Data Science and Big Data Lab, Pablo de Olavide University, 41013 Seville, Spain
{dmarper2,froddia,artrogar,atrolor,fmaralv}@upo.es
[2] Department of Computer Science, University of Seville, 41012 Seville, Spain
dgutierrez3@us.es

Abstract. This paper presents an exploratory study on the use of rule-based explainability to support clinical decision-making in thyroid cancer recurrence prediction. Quantum machine learning models are compared within a unified experimental workflow, and a model-agnostic explainability technique is applied to extract interpretable rules. These rules are consistent with established clinical risk factors and are further used for explainability-driven feature selection, leading to improved model behavior, particularly in terms of sensitivity. The results provide useful information on the applicability of explainable quantum models and show consistent behavior in both ideal and noisy computational environments, highlighting their potential for interpretable decision support in healthcare.

Keywords: Quantum machine learning · Association rules · Explainability · Feature selection · Thyroid cancer.

1 Introduction

Machine learning (ML) techniques are increasingly being used in the healthcare sector to support clinical decision-making by analyzing complex patient data. These methods have yielded promising results in tasks such as diagnosis, prognosis estimation, and risk prediction. However, as predictive models become more sophisticated, their internal decision-making processes often prove difficult to interpret, raising concerns about transparency, trust, and clinical usability. In medicine, predictive performance alone is not sufficient, as healthcare professionals also need to understand the factors driving a model's decisions to assess whether they are plausible, reliable, and consistent with medical knowledge. For this reason, explainable artificial intelligence (XAI) has become an important field of research aimed at making ML models more understandable and applicable [1,2].

* Corresponding author.

At the same time, quantum machine learning (QML) [3] has attracted growing interest as a paradigm that combines machine learning techniques with the computational principles of quantum mechanics. Although it is still in the exploratory phase, several quantum approaches have been proposed as alternatives or complements to classical models for classification and prediction tasks. Among them, quantum support vector machines (QSVM) [4] and quantum neural networks (QNN) [5] are two of the most representative architectures. However, the use of quantum circuits and high-dimensional feature mappings makes QML models particularly difficult to interpret, which further limits their applicability in domains where explainability is essential [6].

This study addresses the prediction of thyroid cancer recurrence in a realistic clinical setting [7]. Although well-differentiated thyroid cancer typically has a favorable prognosis, recurrence remains a clinically significant event that can occur years after treatment. Therefore, predicting the risk of recurrence is important for follow-up, treatment planning, and personalized monitoring, making transparent and reliable predictive models particularly valuable [8].

This article presents an exploratory study on the application of QML models to a real-world dataset on thyroid cancer for the prediction of recurrence. Specifically, QSVM and QNN models are evaluated in both an ideal simulation and in a noisy emulation based on a real IBM quantum device. The dataset includes several categorical clinical variables that were appropriately encoded for model processing, while the target variable corresponds to cancer recurrence.

Beyond predictive performance, interpretability is analyzed by extracting association rules from the best-performing model [9]. These rules reveal relevant relationships between clinical features and support an attribute classification inspired by prior explainability criteria. Based on this classification, the models are retrained using both the full set of features and reduced subsets containing the most relevant variables, allowing for an evaluation of the usefulness of the extracted rules as a feature selection mechanism.

The main contributions of this study are summarized as follows. First, we present an exploratory study on predicting thyroid cancer recurrence using a real-world clinical dataset with categorical and clinical variables encoded for QML models. Second, we extract association rules from the best-performing model to obtain interpretable insights into the relationships between clinical attributes associated with thyroid cancer recurrence. Third, we define a feature relevance ranking based on these rules, which serves as a feature selection mechanism to construct reduced subsets of variables. Fourth, we evaluate the impact of retraining the models with different feature subsets to determine if this rule-guided selection improves predictive performance.

The rest of the manuscript is structured as follows. Section 2 reviews previous research on explainability in classical and QML, with a particular focus on applications in the healthcare field. Section 3 describes the dataset, the preprocessing phase, the learning models, and the feature analysis based on association rules. Section 4 presents and analyzes the experimental results. Finally, Section 5 summarizes the main conclusions, limitations, and future research directions.

2 Related work

Recent research on quantum computing in the healthcare sector has focused primarily on potential applications, technical limitations, and the feasibility of its clinical use in the short term. Described use cases include diagnosis, treatment optimization, medical data processing, and decision support, although current noisy intermediate-scale quantum (NISQ) hardware continues to face significant limitations related to noise and scalability [10,11,12]. In this context, QML has become one of the most active areas of research, particularly for classification tasks based on hybrid quantum-classical workflows.

Several studies have investigated quantum classifiers in controlled environments. Suzuki et al. [13] analyzed classification and regression using quantum support vectors under both noisy and noise-free conditions, as well as on trapped-ion hardware, highlighting the importance of evaluating quantum models beyond idealized simulations. In the healthcare field, QML has also been applied to disease prediction tasks, such as cardiovascular risk detection, where quantum-enhanced classifiers have been proposed as proof-of-concept alternatives to classical approaches [14]. However, most of these studies remain exploratory and are often based on relatively small or benchmark datasets.

At the same time, explainability in QML has attracted growing interest. Since classical XAI methods cannot be directly applied to quantum models, recent work has proposed specific frameworks to address this challenge. ExQUAL combines LIME and SHAP with a Pegasos-based QSVM to obtain local and global explanations [15], while QRLaXAI extends explainability to quantum representation learning using a quantum autoencoder and a variational quantum classifier with post-hoc explanation methods [16]. Recent studies have also highlighted the importance of having interpretable and clinically reliable quantum-based processes in the healthcare field [17].

In thyroid cancer, most studies continue to rely on classical machine learning. Wang et al. [18] developed models to predict recurrence of papillary thyroid cancer using multiple clinical-pathological predictors, and Borzooei et al. [19] demonstrated that machine learning models can support recurrence risk stratification and follow-up planning in patients with well-differentiated thyroid cancer. More recent studies have incorporated strategies focused on explainability and features. Atay et al. [20] combined association rule mining with classification algorithms for thyroid cancer recurrence, while Yapici and Uzun Arslan [21] addressed class imbalance and feature redundancy through data balancing and feature selection.

Despite these advances, the overlap between these lines of research remains limited. The prediction of thyroid cancer recurrence has been studied primarily using classical machine learning, while the interpretability of quantum learning has been investigated largely using benchmark datasets or in other clinical tasks. Therefore, the combination of quantum models for predicting thyroid cancer recurrence with explanatory analysis based on association rules and feature classification remains largely unexplored, which motivates the present study.

3 Methodology

This study follows an exploratory experimental workflow designed to assess both predictive performance and interpretability of QML models in a clinical prediction task. The workflow integrates model training, post-hoc explainability, and feature selection in a unified framework and is depicted in Figure 1.

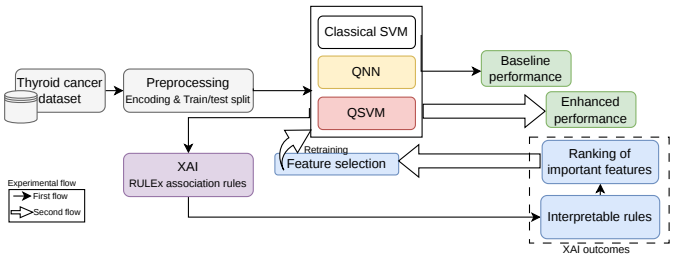


Fig. 1. Experimental workflow adopted in this study. QML models were trained using the same preprocessed clinical dataset to establish baseline performance. The RULEx explainability technique is then applied to extract interpretable rules and derive feature importance scores, which are subsequently used for explainability-driven feature selection and model retraining.

The methodology is organized as a two-stage experimental workflow. The objective is not only to evaluate the predictive capacity of QML models for thyroid cancer recurrence, but also to analyze whether the explanations extracted from these models can guide a subsequent feature selection process.

The workflow starts from a clinical dataset related to well-differentiated thyroid cancer. The prediction target is recurrence, that is, whether cancer reappears after the initial treatment. The dataset contains demographic and clinical variables such as age, gender, smoking status, physical examination, risk category, tumor stage, lymph node involvement, metastasis, and recurrence status. Since the dataset contains no missing values, the preprocessing stage mainly focuses on preparing the variables for quantum model training. Categorical variables are encoded into numerical values, and all input features are normalized to the interval $[0, \pi]$, which allows them to be mapped into quantum circuits.

After preprocessing, the dataset is divided into training and test partitions. The same preprocessed data are used to train all models, ensuring that their results are directly comparable. Three types of classifiers are considered: a classical SVM, used as a reference baseline, and two quantum models, namely QSVM and QNN. The QSVM model utilized in this study is the implementation of Pegasos (Primal Estimated sub-GrAdient SOlver for SVM) [22] adapted for quantum kernels [23], while the QNN architecture implemented is based on a Variational Quantum Classifier (VQC) paradigm [24].

Both quantum models are evaluated under two computational settings: an ideal, noise-free simulation and a noisy emulation based on the IBM Fez quantum processor. This design makes it possible to compare theoretical performance with a more realistic noisy scenario. The first part of the workflow therefore

establishes the baseline performance of each model using the complete set of 13 input features. Accuracy, sensitivity, specificity, and training time are used to assess model behavior. This baseline stage is important because it shows how the models perform before applying any explainability-guided reduction of the input space.

Once the initial models have been trained, the best-performing model (the classical SVM, which achieved the highest baseline accuracy and sensitivity) is analyzed using RULEx [25], a model-agnostic explainability method based on quantitative association rules. RULEx extracts rules that relate combinations of input variables to the predicted class. Since RULEx is model-agnostic, it constructs this rule-based mapping by explaining the relationship between the input features and the predictions generated by the classifier, rather than relying on its internal parameters. These rules provide an interpretable description of the decision patterns learned by the model.

To prevent data leakage during the feature selection process, the RULEx explainability analysis and feature ranking were performed strictly on the training partition of the first-stage models. The resulting Top-9 and Top-4 feature subsets were then used to filter both the training and test sets for the second-stage model training and evaluation.

The explainability stage produces two outputs. The first output is a set of human-readable clinical rules, which helps assess whether the model's behavior is consistent with known clinical knowledge. The second output is a feature importance ranking. This ranking is obtained by counting how frequently each variable appears in the antecedents of the extracted rules.

This feature ranking is then used to build reduced datasets. In particular, the study defines a Top-9 subset, which includes the variables with non-zero relevance according to RULEx, and a Top-4 subset, which retains only the four most important variables. The original models are then retrained using these reduced feature subsets. This second training stage evaluates whether the explanations obtained from RULEx can improve model performance by removing irrelevant or redundant variables. Finally, the performance on the full dataset is compared with that on the Top-9 and Top-4 subsets.

4 Results

This section presents the experimental results obtained following the workflow described in Section 3. First, the input dataset and preprocessing steps are described in Section 4.1. The experimental setup is described in Section 4.2. Then, the baseline performance of the QML models is reported and analyzed in Section 4.3, followed by the explainability analysis carried out with RULEx in Section 4.4. Finally, the impact of explainability-driven feature selection on model performance is evaluated in Section 4.5.

4.1 Input data

The input data describe well-differentiated thyroid cancer. The objective is to predict recurrence, defined as the reappearance of cancer cells after initial treatment, which has important implications for patient prognosis and management [26]. The dataset consists of a cohort of 383 instances (patient records) and includes 13 clinicopathological features collected over 15 years, with each patient followed for at least 10 years. It is publicly available as the Differentiated Thyroid Cancer Recurrence dataset in the UCI Machine Learning Repository, developed within AI–medicine research in the United States [19]. The dataset contains no sensitive information and has no missing values. Table 1 lists the dataset features, including demographic and clinical variables relevant to thyroid cancer recurrence prediction.

Table 1. Clinicopathological features present in the input data (383 instances total).

Feature	Description	Type	Categories / Range	Frequencies / Statistics
Age	Patient age at diagnosis	Numerical	[15, 82]	Mean: 40.87 ± 15.13
Gender	Gender classification	Binary	0: Female, 1: Male	Female: 81.5% (312), Male: 18.5% (71)
Smoking	Current smoking status	Binary	0: No, 1: Yes	No: 87.2% (334), Yes: 12.8% (49)
Smoking history	Past smoking history	Binary	0: No, 1: Yes	No: 92.7% (355), Yes: 7.3% (28)
Radiotherapy history	Prior radiation therapy	Binary	0: No, 1: Yes	No: 98.2% (376), Yes: 1.8% (7)
Physical examination	Clinical neck assessment findings	Categorical	Normal, Diffuse goiter, Single nodular-left/right, Multinodular goiter	Multinodular goiter: 36.6% (140), Single nodular goiter-right: 36.6% (140), Single nodular goiter-left: 23.2% (89), Normal: 1.8% (7), Diffuse goiter: 1.8% (7)
Focality	Number of tumor foci	Binary	0: Uni-Focal, 1: Multi-Focal	Uni-Focal: 64.5% (247), Multi-Focal: 35.5% (136)
Risk	Clinical recurrence risk stratification	Binary	0: Low, 1: Intermediate/High	Low: 65.0% (249), Intermediate/High: 35.0% (134)
Tumor	Tumor size (T staging)	Categorical	T1a, T1b, T2, T3a, T3b, T4a, T4b	T2: 39.4% (151), T3a: 25.1% (96), T1a: 12.8% (49), T1b: 11.2% (43), T4a: 5.2% (20), T3b: 4.2% (16), T4b: 2.1% (8)
Lymph nodes	Lymph involvement (N staging)	Binary	0: No (N0), 1: Yes (N1a/N1b)	No (N0): 70.0% (268), Yes (N1a/N1b): 30.0% (115)
Metastasis	Distant metastasis (M staging)	Binary	0: No (M0), 1: Yes (M1)	No (M0): 95.3% (365), Yes (M1): 4.7% (18)
Stage	AJCC cancer stage	Categorical	Stage I to IVB	Stage I: 86.9% (333), Stage II: 8.4% (32), Stage IVB: 2.9% (11), Stage III: 1.0% (4), Stage IVA: 0.8% (3)
Recurred	Patient recurrence status (Target)	Binary	0: No, 1: Yes	No (No recurrence): 71.8% (275), Yes (Recurrence): 28.2% (108)

The first step involved preprocessing the input data. Categorical variables were numerically encoded and mapped to integer representations according to established clinical criteria to ensure model compatibility. As the dataset was already clean, no missing value imputation was required. To prevent data leakage, the scaling parameters for the normalization of continuous variables (i.e., Age) to the range $[0, \pi]$ were computed strictly on the training partition and subsequently applied to scale the test partition. The dataset was split into training (80%) and test (20%) partitions. The target variable, "Recurrence", shows a mild class imbalance, with approximately 28.2% of samples in class 1 (recurrence) and 71.8% in class 0 (no recurrence). No correction techniques were initially applied, as the imbalance was not severe. Nevertheless, its potential impact on model performance was considered, particularly regarding sensitivity and specificity.

4.2 Experimental setup

The experimental framework was implemented in Python using the Qiskit ecosystem [27], using its specialized libraries for quantum circuit construction. It was executed on the Hércules supercomputer, managed by the Centro de Informática

Científica de Andalucía (CICA). This high-performance computing (HPC) infrastructure provided the necessary computational resources to handle the intensive classical simulations of quantum circuits. For the noisy execution scenarios, hardware-specific noise profile of the 127-qubit IBM Fez processor was integrated. This noise model was accessed via the Qiskit Runtime Service, incorporating critical device parameters such as gate fidelity, thermal relaxation times and dephasing times. By combining HPC-grade classical power with real-world quantum noise data, we ensured a robust environment for evaluating the scalability and reliability of the proposed QML models.

4.3 General QML results

Table 2 presents the preliminary performance metrics for the classical SVM, and the ideal and noisy quantum models, using the full 13-feature set. Specifically, QSVM_{id} and QNN_{id} refers to the ideal conditions whereas $\text{QSVM}_{\text{noise}}$ and $\text{QNN}_{\text{noise}}$ to the model trained simulating real noise. Then, under ideal conditions, the QML models achieve slightly better performance than their noisy counterparts. A critical observation in all models is the impact of the class distribution, which is moderately imbalanced (70% for no recurrences, 30% for recurrence).

Table 2. Performance on the full 13-feature input dataset.

Model	Acc (%)	Sens (%)	Spec (%)	t_{train} (s)
SVM	85.5 ± 3.5	77.3 ± 9.3	88.6 ± 6.0	0.01 ± 0.00
QSVM_{id}	70.4 ± 16.5	64.7 ± 30.1	72.6 ± 30.0	26.95 ± 2.29
QNN_{id}	75.5 ± 1.9	16.3 ± 6.0	98.0 ± 1.2	1215.46 ± 18.06
$\text{QSVM}_{\text{noise}}$	65.8 ± 19.5	60.2 ± 32.0	68.1 ± 31.5	2173 ± 241
$\text{QNN}_{\text{noise}}$	72.6 ± 2.8	18.5 ± 7.5	95.4 ± 2.0	94612 ± 1438

Concerning the ideal models, QNN_{id} achieves a seemingly competitive accuracy of 75.5%. However, a closer inspection of its sensitivity (16.3%) and specificity (98.0%) reveals that the model is heavily biased towards the majority class. This indicates that the quantum kernel approach is more resilient to class imbalance than the QNN at this dimensionality. However, the high standard deviation in QSVM_{id} suggests instability when handling the full feature space without prior selection. With reference to training time, a significant bottleneck was observed for the QNN_{id}, requiring over 20 minutes (1215.46 s) compared to the nearly instantaneous classical SVM and the more efficient QSVM algorithm (26.95 s). Regarding the noisy models, $\text{QSVM}_{\text{noise}}$ shows a slight degradation in accuracy (65.8% vs. 70.4%) and sensitivity (60.2% vs. 64.7%) compared to its ideal counterpart, while maintaining a comparable level of variance. This suggests that the kernel-based approach retains a degree of noise resilience, at the cost of a significant increase in training time, rising from 26.95 s to 2173 s. In contrast, $\text{QNN}_{\text{noise}}$ maintains the same majority-class bias observed under ideal conditions, with a sensitivity of 18.5% and a specificity of 95.4%, indicating that quantum noise does not correct the variational circuit's tendency to predict

the majority class. Moreover, the associated training time increases to 94612 s, considerably limiting its practical feasibility when using the full feature set.

4.4 Explainability analysis: RULEx

To better understand the observed model behavior, RULEx was applied to analyze the relationship between input features and recurrence predictions. The extracted QARs provide interpretable insights into disease recurrence based on patient characteristics. These QARs are consistent with established clinical risk factors for thyroid cancer recurrence, such as larger tumor size, lymph node involvement, and older age, which have been reported as significant predictors in clinical studies [28].

R1: $\text{Gen}[0] \wedge \text{Smok}[0] \wedge \text{Met}[0] \implies \text{Rec}[0]$
R2: $\text{Risk}[0] \implies \text{Rec}[0]$
R3: $\text{LN}[1] \implies \text{Rec}[1]$
R4: $\text{PhysExam}[1] \wedge \text{Risk}[1] \implies \text{Rec}[1]$
R5: $\text{Gen}[1] \implies \text{Rec}[1]$
R6: $\text{Stage}[2-4] \implies \text{Rec}[1]$
R7: $\text{Age}[51-75] \wedge \text{Foc}[1] \implies \text{Rec}[1]$
R8: $\text{Age}[67-82] \wedge \text{Risk}[1] \implies \text{Rec}[1]$

In detail, the patients identified by the rules as low-risk are those who do not smoke and are usually female (gender 0). Moreover, patients with no metastasis are less likely to experience recurrence. In contrast, certain factors increase the probability of recurrence, such as having affected lymph nodes, being male (gender 1) or being categorized as high risk level.

Clinical indicators also play a crucial role. Patients with a positive physical examination and high-risk classification, as well as those in cancer stages 2 to 4, have a higher likelihood of recurrence. Additionally, age-related factors are significant: individuals aged 51 or older are more prone to recurrence. Beyond interpretability, the rules have been used to assess the importance level of the input features.

Feature importance is estimated based on the frequency with which variables appear in the rule antecedents, under the assumption that features occurring more frequently across multiple rules exert a greater influence on the model's predictions. Table 3 reports the resulting importance levels for each input feature.

Table 3. Feature importance grouped by level.

Importance	Features
0.231	Risk
0.154	Age, gender
0.077	Smoking, physical examination, focality, lymph nodes, metastasis, stage
0.000	Smoking history, radiotherapy history, tumor

4.5 Performance after explainability-driven feature selection

The models were retrained using two reduced feature subsets: Top-9, comprising the features with the three highest importance levels (risk, age, gender, smoking,

physical examination, focality, lymph nodes, metastasis, and stage), and Top-4, formed by the four features with the highest individual importance scores. Table 4 presents the resulting performance metrics and training times for each configuration.

Table 4. Performance after RULEx-driven feature selection using different subsets: full set (All), Top-9, and Top-4.

Model	All				Top-9				Top-4			
	Acc. (%)	Sens. (%)	Spec. (%)	t (s)	Acc. (%)	Sens. (%)	Spec. (%)	t (s)	Acc. (%)	Sens. (%)	Spec. (%)	t (s)
SVM	85.5 ± 3.5	77.3 ± 9.3	88.6 ± 6.0	0.01 ± 0.00	84.1 ± 2.9	72.5 ± 10.9	88.7 ± 5.5	0.03 ± 0.04	84.5 ± 3.0	87.4 ± 6.3	83.3 ± 5.0	0.06 ± 0.03
QSVM_{id}	70.4 ± 16.5	64.7 ± 30.1	72.6 ± 30.0	27.0 ± 2.3	67.5 ± 15.0	76.4 ± 23.0	63.9 ± 22.0	13.2 ± 2.0	66.7 ± 18.5	68.3 ± 31.4	66.1 ± 30.0	5.2 ± 0.8
QNN_{id}	75.5 ± 1.9	16.3 ± 6.0	98.0 ± 1.2	1215.5 ± 18.1	62.8 ± 3.8	98.4 ± 1.8	48.6 ± 5.0	281.3 ± 2.7	64.8 ± 7.8	94.2 ± 9.6	53.1 ± 10.5	89.47 ± 1.73
QSVM_{noi}	65.8 ± 19.5	60.2 ± 32.0	68.1 ± 31.5	2173 ± 241	63.4 ± 17.5	71.8 ± 25.0	60.0 ± 23.0	1087 ± 143	62.5 ± 21.0	64.5 ± 32.0	61.7 ± 31.0	443 ± 57
QNN_{noi}	72.6 ± 2.8	18.5 ± 7.5	95.4 ± 2.0	94612 ± 1438	60.1 ± 5.5	94.8 ± 4.0	46.3 ± 6.5	21834 ± 412	62.0 ± 9.0	90.5 ± 11.0	50.6 ± 12.0	6921 ± 193

Regarding the classical SVM, the Top-4 subset leads to an improvement in sensitivity (87.4% vs. 77.3% with all features), while accuracy remains largely stable (84.5%) and specificity is balanced at 83.3%. For QSVM_{id}, the Top-9 subset yields a moderate increase in sensitivity (76.4% vs. 64.7%), at the cost of a slight drop in accuracy and a decrease in specificity to 63.9%, while reducing training time from 27.0 s to 13.2 s. A similar trend is observed for QSVM_{noi}, where sensitivity improves from 60.2% to 71.8% with Top-9, and specificity is 60.0%, with training time reduced by approximately half.

The most notable changes are observed for both QNN variants. With the full feature set, QNN_{id} exhibits a strong bias towards the majority class, with a specificity of 98.0% and a sensitivity of only 16.3%, essentially acting as a majority-class predictor. After explainability-driven feature selection, this bias is substantially corrected: sensitivity reaches 98.4% with Top-9 and 94.2% with Top-4. However, a closer look at the specificity reveals that it drops significantly to 48.6% (Top-9) and 53.1% (Top-4). A comparable pattern is observed under noisy conditions, where QNN_{noi} improves from 18.5% to 94.8% sensitivity with Top-9, but its specificity falls to 46.3%.

These results suggest that while the high-dimensional feature space was a primary contributor to the majority-class bias of the variational model, feature selection based on explainability rules did not simply yield a balanced model; instead, it shifted the decision boundary in the opposite direction (toward predicting the positive class). This indicates that class imbalance is indeed a severe challenge for the variational quantum circuit, and that feature selection alone is insufficient to fully correct it. Future work must investigate dedicated class imbalance correction techniques, such as synthetic oversampling or class-weighted loss functions, specifically tailored for quantum algorithms.

Furthermore, when comparing classical and quantum models, the classical SVM baseline consistently outperforms the quantum classifiers (QSVM and QNN) in terms of overall accuracy and computational efficiency, with training times that are orders of magnitude faster. This absence of a clear "quantum advantage" is expected given the limitations of NISQ-era simulated environments, device noise profiles, and the small dataset size (383 instances). Nevertheless, the results demonstrate that explainability-driven feature selection can narrow

the performance gap and reduce the computational overhead, which is critical for making QML models more practical in clinical settings.

Additionally, while traditional rule-based classifiers (such as Decision Trees or RIPPER) could be applied directly to the dataset to provide interpretable baselines, the primary focus of this exploratory study is to assess the capacity of hybrid classical-quantum models and show how post-hoc model-agnostic rule-based explainability can guide feature selection in quantum workflows, rather than finding the single best classifier.

5 Conclusions and future work

This study presents an exploratory analysis and does not introduce new methodological developments. The primary contribution lies in the application and joint analysis of existing QML and explainability techniques within a healthcare context. The results showed that both QSVM and QNN exhibit limitations when trained on the full feature set, particularly in terms of sensitivity, and that these limitations persist under noisy simulation conditions. The application of the model-agnostic XAI technique RULEx provided interpretable rule-based explanations consistent with established clinical risk factors for thyroid cancer recurrence. Moreover, explainability-driven feature selection led to improved sensitivity across classical and quantum models, including under noise, with a concurrent reduction in training time.

Despite these findings, this study has several limitations. First, the study is based on a single retrospective dataset with a limited sample size of 383 instances, which may restrict the generalizability of the results. Second, the quantum model is evaluated using a simulated quantum environment rather than real quantum hardware, and therefore, the reported computational costs and performance do not reflect the full physical constraints of actual quantum processors. Third, the feature importance ranking is based on the simple frequency of feature occurrences in the antecedents of the extracted rules. While effective, this metric does not capture the complex co-occurrence and synergistic interactions between subsets of clinical variables.

Future work will focus on validating these results on larger and independent clinical cohorts, as well as exploring alternative XAI techniques to further assess their suitability for QML models. From a quantum computing perspective, future studies may investigate different circuit architectures, noise-aware simulations, and experiments on real quantum hardware. We also plan to explore more sophisticated rule-based importance measures that account for feature co-occurrence and rule strength (e.g., support, confidence, lift) to refine the feature selection process. Incorporating feedback from medical experts into the explainability analysis also represents a promising direction to enhance clinical relevance.

Data and code availability

The datasets analyzed, the models, and the complete source code used are available at: `https://github.com/Data-Science-Big-Data-Research-Lab/XAI_QML_CAEPIA2026`. This repository supports full reproducibility of the experiments and facilitates independent validation of the results.

Acknowledgments

The authors would like to thank the Spanish Ministry of Science and Innovation for the support under the projects PID2023-146037OB-C21 and PID2023-146037OB-C22 funded by MICIU/AEI/10.13039/501100011033. The Junta de Andalucía's Centro de Informática Científica de Andalucía (CICA) is also acknowledged for providing the computing resources on the Hércules supercomputer, which were instrumental in performing the simulations/calculations presented in this work. We also acknowledge Pablo de Olavide University for funding the Q-Resilience project.

References

1. S. S. Band, A. Yarahmadi, C. C. Hsu, et al. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286, 2023.
2. Z. Sadeghi, R. Alizadehsani, M. A. Cifci, et al. A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, 118:109370, 2024.
3. F. Rodríguez-Díaz, D. Gutiérrez-Avilés, A. Troncoso, and F. Martínez-Álvarez. A survey of quantum machine learning: Foundations, algorithms, frameworks, data and applications. *ACM Computing Surveys*, 58:91, 2026.
4. F. Rodríguez-Díaz, J. F. Torres, D. Gutiérrez-Avilés, A. Troncoso, and F. Martínez-Álvarez. An Experimental Comparison of Qiskit and Pennylane for Hybrid Quantum-Classical Support Vector Machines. *Lecture Notes in Artificial Intelligence*, 14640:121–130, 2024.
5. A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner. The power of quantum neural networks. *Nature Computational Science*, 1:403–409, 2021.
6. L. Pira and C. Ferrie. On the interpretability of quantum neural networks. *Quantum Machine Intelligence*, 6(52), 2024.
7. G. Hu, B. Huang, and L. Cai. Machine learning prediction of thyroid cancer recurrence for early screening and clinical decision pathways: a retrospective cohort study. *Discover Oncology*, 17(1), 2026.
8. A. Schindele, A. Krebold, U. Heiß, et al. Interpretable machine learning for thyroid cancer recurrence predicton: Leveraging XGBoost and SHAP analysis. *European Journal of Radiology*, 186:112049, 2025.
9. P. Patel, B. Sivaiah, R. Patel, and R. Choudhary. Association Rule Mining for Healthcare Data Analysis. *Computational Intelligence in Healthcare Informatics*, 1132:127–139, 2024.
10. V. K. Bukkarayasamudram, P. C. R. Shaker, K. A. Kumar, et al. Quantum computing revolution in healthcare: a systematic review of applications, issues and future directions. *Artificial Intelligence Review*, 58(389), 2025.

11. S. Alrashed and N. Min-Allah. Quantum computing research in medical sciences. *Informatics in Medicine Unlocked*, 52:101606, 2025.
12. S. C. Fairburn, L. Jehi, B. T. Bicknell, B. G. Wilkes, and B. Panuganti. Applications of quantum computing in clinical care. *Frontiers in Medicine*, 12:1573016, 2025.
13. T. Suzuki, T. Hasebe, and T. Miyazaki. Quantum support vector machines for classification and regression on a trapped-ion quantum computer. *Quantum Machine Intelligence*, 6(31), 2024.
14. S. Prabhu, S. Gupta, G. M. Prabhu, et al. QuCardio: Application of Quantum Machine Learning for Detection of Cardiovascular Diseases. *IEEE Access*, 11:136122–136135, 2023.
15. K. Kadian, S. Garhwal, and A. Kumar. ExQUAL: an explainable quantum machine learning classifier. *Applied Intelligence*, 55(13), 2025.
16. A. K. Kankanamge Don and I. Khalil. QRLaXAI: quantum representation learning and explainable AI. *Quantum Machine Intelligence*, 7(24), 2025.
17. S. S. Singh, S. Kumar, R. Ahuja, and J. Barua. Fusion of quantum computing and explainable AI: A comprehensive survey on transformative healthcare solutions. *Information Fusion*, 122:103217, 2025.
18. H. Wang, C. Zhang, Q. Li, et al. Development and validation of prediction models for papillary thyroid cancer structural recurrence using machine learning approaches. *BMC Cancer*, 1:427, 2024.
19. S. Borzooei, G. Briganti, M. Golparian, et al. Machine learning for risk stratification of thyroid cancer patients: a 15-year cohort study. *European Archives of Oto-Rhino-Laryngology*, 281:2095–2104, 2024.
20. F. F. Atay, F. H. Yagin, C. Colak, et al. A hybrid machine learning model combining association rule mining and classification algorithms to predict differentiated thyroid cancer recurrence. *Frontiers in Medicine*, 11:1461372, 2024.
21. I. S. Yapici and R. U. Arslan. Predictive analytics for thyroid cancer recurrence: a feature selection and data balancing approach. *The European Physical Journal Special Topics*, 234:4751–4771, 2025.
22. S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated subgradient solver for SVM. In *Proceedings of the International Conference on Machine learning*, pages 807–814, 2007.
23. 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.
24. M. Schuld and N. Killoran. Quantum machine learning in feature Hilbert spaces. *Physical Review Letters*, 122(4):040504, 2019.
25. A. R. Troncoso-García, M. Martínez-Ballesteros, F. Martínez-Álvarez, and A. Troncoso. A new approach based on association rules to add explainability to time series forecasting models. *Information Fusion*, 94:169–180, 2023.
26. A. R. Shaha. Recurrent differentiated thyroid cancer. *Endocrine Practice*, 18(4):600–603, 2012.
27. G. Aleksandrowicz, T. Alexander, P. Barkoutsos, et al. Qiskit: An open-source framework for quantum computing. *Zenodo*, 2019. https://zenodo.org/records/2562111, last accessed on May 2026.
28. M. Rahman, M. Anwar Hossain, T. Rahman, and A.M. Jewel. Thyroid cancer recurrence: Risk factors and preventive strategies. *International Journal of Otorhinolaryngology*, 10(2):58–64, 2024.