Journal of Science and Technology – Smart Systems and Devices
Volume 36, Issue 3, September 2026, 001–007

A Hybrid Dimensionality Reduction and Quantum Support Vector Machine for Breast Cancer Diagnosis

Hang Dang[1], My Nguyen [2], Van Tran[3], Dinh Do Van[4,*]
[1]Le Quy Don Technical University, Ha Noi, Vietnam
[2]Military Hospital 7, Hai Phong, Vietnam
[3]Hospital of University of Medicine and Pharmacy,
University of Medicine and Pharmacy, Ha Noi, Vietnam
[4]Sao Do University, Hai Phong, Vietnam
[*]Corresponding author email: dinh.dv@saodo.edu.vn

Abstract

Quantum Support Vector Machines (QSVMs) have recently emerged as a promising approach for biomedical classification tasks. However, their practical deployment remains constrained by limited qubit availability and sensitivity to high-dimensional feature spaces. These challenges often result in increased computational overhead and reduced model reliability. To mitigate these issues, this study proposes a hybrid framework integrating Pearson correlation–based feature selection and Principal Component Analysis (PCA) with QSVMs for breast cancer diagnosis. First, Pearson correlation analysis is employed to remove redundant and weakly relevant features, thereby improving data quality and reducing dimensionality at an early stage. Subsequently, PCA projects the selected attributes into a compact subspace while preserving most of the original data variance. This dimensionality reduction strategy decreases the number of qubits required for quantum encoding and improves computational efficiency. Experiments conducted on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset demonstrate that the proposed hybrid QSVM achieves 98% classification accuracy, outperforming or matching existing classical and quantum-based approaches. The results confirm that combining classical preprocessing techniques with quantum classifiers provides a robust and resource-efficient solution for biomedical data analysis. Overall, the findings of this study underscore the effectiveness of combining classical preprocessing techniques with quantum machine learning models. By bridging the gap between classical data optimization and quantum computational power, the proposed approach offers a resource-efficient, scalable, and high-performance solution for complex biomedical data analysis, paving the way for more feasible applications of quantum computing in healthcare domains.

Keywords: Breast cancer diagnosis, feature selection, principal component analysis, quantum machine learning, quantum support vector machine.

1. Introduction

Breast cancer remains one of the leading causes of cancer–related mortality among women worldwide, with approximately 1.7 million patients and more than 500,000 deaths each year [1]. Early detection significantly improves survival rates. However, late-stage diagnosis remains prevalent, particularly in developing countries (40–90%) [2, 3]. Consequently, the development of accurate computer-aided diagnostic systems has become an essential research objective.

Quantum Machine Learning (QML) has gained increasing attention due to its potential computational advantages over classical algorithms [4]. Among QML methods, QSVMs leverage quantum kernel estimation to enhance classification performance [5-7]. Despite promising theoretical advantages, Quantum Support Vector Machines (QSVMs) implementation is limited by current Noisy Intermediate–Scale Quantum (NISQ) hardware constraints, especially the restricted number of available qubits. Since each feature typically corresponds to a qubit in quantum encoding schemes, high–dimensional datasets pose significant challenges.

In classical machine learning, feature selection and dimensionality reduction techniques-such as Pearson correlation filtering and Principal Component Analysis (PCA)–have proven effective in improving model performance and reducing computational complexity. Inspired by these findings, this study integrates classical preprocessing methods with QSVMs to construct a hybrid classification framework. Ibrahim *et al*. applied this method to the WBCD set and achieved an accuracy of 98.24%, surpassing methods using only SVM, RF, or DT [15].

p-ISSN 3093-3285
e-ISSN 3093-3315
https://doi.org/10.51316/jst.xxx.ssad.2026.36.x.x
Received: Mar 4, 2026; Revised: Apr 5, 2026;
Accepted: Apr 7, 2026; Online: May 6, 2026

Table 1. Description of characteristics

Feature	Description
Radius	Average value of the distance from the center to points on the tumor circumference.
Texture	Standard deviation of gray level values (gray–scale)
Perimeter	
Area	
Smoothness	The degree of local variation of radius length
Compactness	
Concavity	The degree of indentation of the sections within the contour.
Concave points	The number of indentations on the contour
Symmetry	
Fractal dimension	"coastline approximation" – 1

Motivated by these findings, this study combines classical preprocessing techniques with QSVMs to develop a hybrid classification framework. The main contributions of this work are:

- A two-stage dimensionality reduction strategy combining Pearson correlation and PCA prior to quantum encoding.

- Evaluation of the impact of qubit number on QSVM performance.

- Comparative analysis with previously reported quantum and classical approaches.

2. Related works

Recent studies have explored QSVMs applications in biomedical classification. Vashisth *et al*. implemented QSVM with quantum kernels for breast cancer detection and reported competitive performance compared to classical SVM models. Premanand *et al*. compared QSVM (93.8% accuracy) with classical SVM (97.3%), demonstrating feasibility but limited superiority [9]. Akpinar *et al*. investigated different quantum feature maps and circuit configurations, highlighting the influence of circuit architecture on classification performance in high-dimensional biomedical datasets [10].

Unlike prior work focusing primarily on quantum kernel design or circuit optimization, this study emphasizes classical preprocessing integration. By reducing feature dimensionality before quantum encoding, the proposed method mitigates qubit constraints and enhances classification stability.

3. Materials and Methods

3.1 Dataset

The experiments utilize the Wisconsin Diagnostic Breast Cancer (WDBC) dataset from the UCI Machine Learning Repository. The dataset consists of 569 samples obtained from fine–needle aspiration (FNA) of breast masses. Each instance contains 30 real–valued features describing cellular nucleus characteristics [11, 12].

The target variable represents two classes: malignant (M) and benign (B). Features are categorized into three groups: mean, standard error, and worst (largest) values of the following attributes: radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The correlation matrix among all variables is presented in Fig. 1.

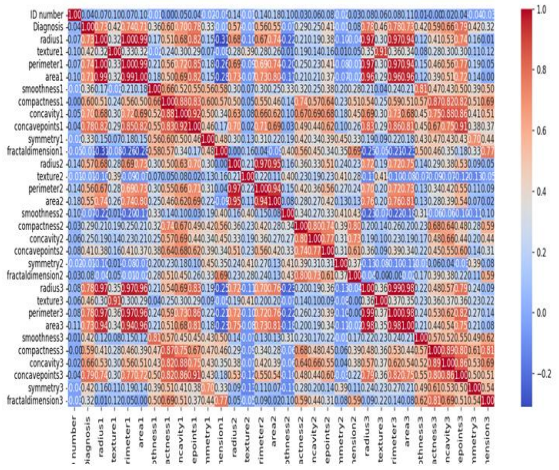

Fig. 1. Correlation matrix between variables

3.2. Data Preprocessing and Feature Selection

All features were normalized using Min–Max scaling to ensure stable model convergence.

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

Min–Max normalization contributes to improved convergence, especially when dealing with features that have large, unevenly sized values.

Pearson correlation analysis was then applied to evaluate linear relationships between features and the target variable [13, 14].

Given two features X and Y, the Pearson correlation coefficient is:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \quad (2)$$

Highly correlated features introduce redundancy and amplify noise during quantum encoding. Let $\Sigma \epsilon \mathbb{R}^{dxd}$ denote the covariance matrix. If Σ is ill-conditioned due to multicollinearity, PCA eigen–decomposition becomes unstable. Therefore, correlation filtering improves numerical conditioning before eigenvalue decomposition.

We retain features satisfying:

$$\left|\rho X_{k,y}\right| > \tau \quad (3)$$

where τ is a predefined threshold.

Highly correlated and redundant attributes were filtered, resulting in a subset of 12 significant features, *'concavepoints3', 'perimeter3', 'concavepoints1', 'radius3', 'perimeter1', 'area3', 'radius1', 'area1', 'concavity1', 'concavity3', 'compactness1'* and *'compactness3'*. The correlation of the selected attributes with the diagnosis variable is shown in Fig. 2.

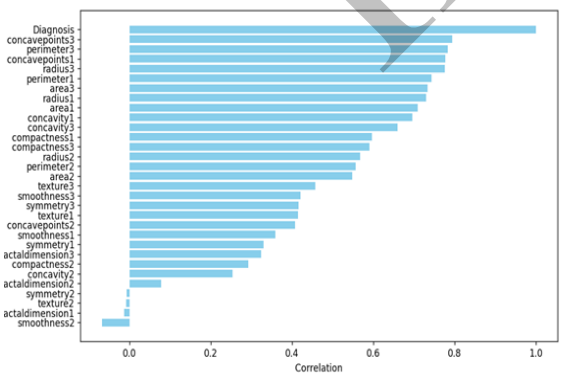

Fig. 2. Correlation with diagnosis

To further reduce dimensionality, PCA was performed on the selected features. The explained variance analysis indicated that the first six principal components preserved the majority of data variance. Consequently, these six components were retained for quantum encoding, reducing qubit requirements without substantial information loss as illustrated in Fig. 3.

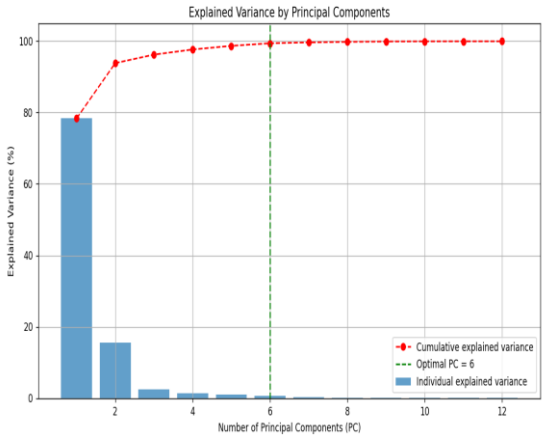

Fig. 3. Variance plot by principal components

During the experiment, we constructed variance maps for each principal component to assess their contribution to the total variance of the data. The results showed that with only the first six principal components, we were able to retain most of the important information, thereby significantly reducing the number of features to be processed without significant information loss. Therefore, features such as *'radius1', 'area1', 'concavity1', 'concavity3', 'compactness1'*, and *'compactness3'* were removed.

3.3. Quantum Support Vector Machine

The dataset was split into training and testing sets using a 70:30 ratio. The QSVM implementation was developed using Qiskit's quantum kernel framework.

Each feature was encoded into a qubit using a custom quantum feature map composed of:

- Hadamard gates
- Parameterized rotation gates (Rx and Rz)
- A chain entanglement layer using controlled–NOT (CX) gates

This architecture was experimentally selected to balance circuit depth and classification performance. The precomputed quantum kernel was then supplied to a classical SVM optimizer.

For a clearer comparison of circuit complexity and performance, the quantum circuit configurations corresponding to 4, 6, and 8 qubits are depicted in the following figures.

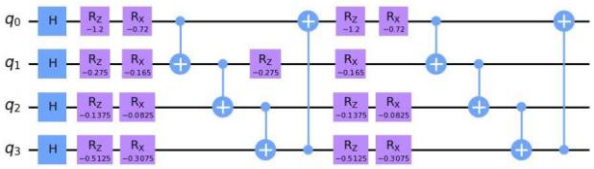

Fig. 4. Quantum circuit with 4 qubits

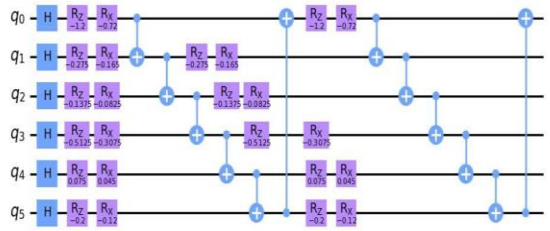

Fig. 5. Quantum circuit with 6 qubits

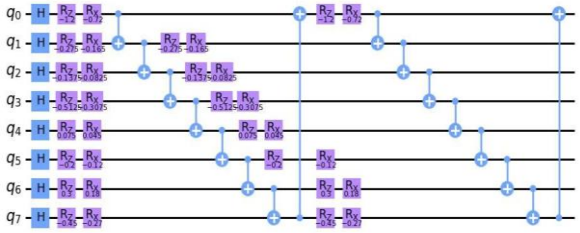

Fig. 6. Quantum circuit with 8 qubits

We employed a custom quantum feature mapping, comprising Hadamard gates, parametric R*x* and R*z* rotations, and a chain–like quantum entanglement layer utilizing CX gates. This structure, is inspired by ZZFeatureMap, was selected experimentally due to its balance between circuit depth and classification performance. Compared to standard feature mappings, this structure achieved higher accuracy on our dataset, demonstrating its suitability for biomedical data with moderate feature correlation.

4. Experimental Results

In this study, we used a 'precomputed' kernel, as this is the only type suitable for the QSVM algorithm. Other kernel types such as 'linear' or 'rbf' are only suitable for classical classification models.

The experimental results of the QSVM algorithm are presented in Table 2, Table 3, Table 4 and Fig. 7, Fig. 8, and Fig. 9.

Table 2. Results of running the QSVM algorithm with 4 Qubits

	Precision	**Recall**	***F1*-score**	**Support**
B	0.96	0.95	0.95	111
M	0.90	0.93	0.92	60
Accuracy			0.94	171
Macro avg	0.93	0.94	0.94	171
Weighted avg	0.94	0.94	0.94	171

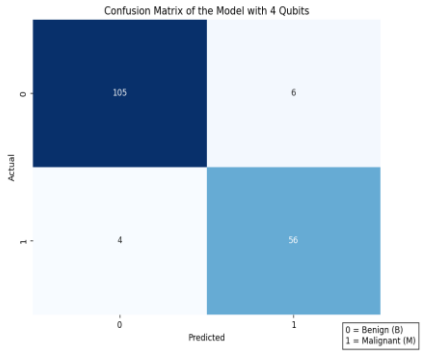

Fig. 7. Confusion matrix with 4 qubits

The model correctly predicted 94% of the 171 samples, which is a high figure, indicating reliable classification. The Benign class had a Precision of 0.96, Recall of 0.95, and *F1*-score of 0.95, showing that the model performed very well, with low positive mispredictions (high precision) and detection of the majority of B cases (high recall). The M (Malignant) class had a Precision of 0.90, Recall of 0.93, and *F1*-score of 0.92, with slightly lower performance than the B class. The lower precision suggests some cases were misdiagnosed as malignant, but the relatively high recall (0.93) is positive because the model still detected the majority of M cases–which is particularly important in biomedical diagnosis.

Table 3. Results of running the QSVM algorithm with 6 Qubits

	Precision	**Recall**	***F1*-score**	**Support**
B	0.98	0.99	0.99	111
M	0.98	0.97	0.97	60
Accuracy			0.98	171
Macro avg	0.98	0.98	0.98	171
Weighted avg	0.98	0.98	0.98	171

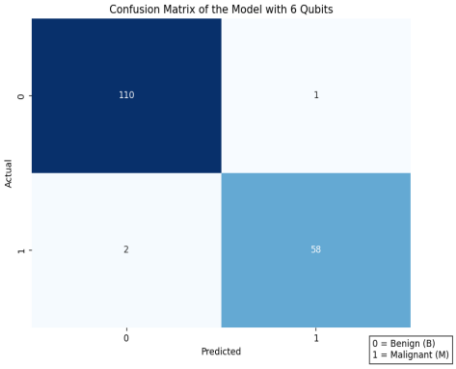

Fig. 8. Confusion matrix with 6 qubits

As illustrated in Table 3, the model achieved very high performance with an Accuracy of 0.98, indicating near-perfect classification accuracy. Both classes B and M had precision, recall, and *F1*-scores of approximately 0.97–0.99, demonstrating stable and balanced performance across classes. In particular, the high recall in class M indicated good detection of critical cases. Overall, this is a very good and reliable result.

Table 4. Results of running the QSVM algorithm with 8 Qubits

	Precision	**Recall**	***F1*-score**	**Support**
B	0.98	0.97	0.98	111
M	0.95	0.97	0.96	60
Accuracy			0.97	171
Macro avg	0.97	0.97	0.97	171
Weighted avg	0.97	0.97	0.97	171

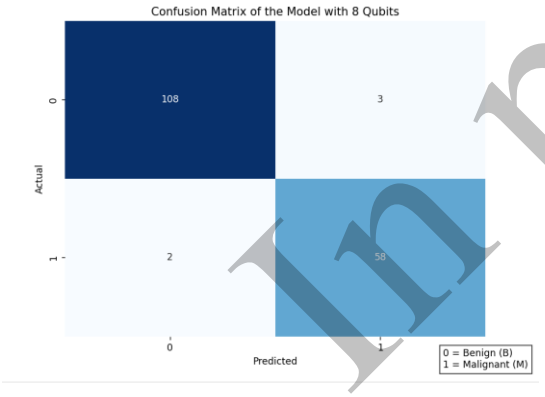

Fig. 9. Confusion matrix with 8 qubits

Table 4 shows that the model achieved high performance with Accuracy = 0.97, indicating very good classification ability. Class B had an *F1*-score = 0.98, while Class M achieved an *F1*-score = 0.96; both precision and recall indices were high, especially the recall of Class M (0.97), showing that the model still maintained good detection of important cases. Compared to the previous result (Accuracy = 0.98), the model's performance decreased slightly. Specifically, the precision of Class M decreased from 0.98 to 0.95, leading to a lower *F1*-score for this class, indicating an increase in false predictions. In addition, the recall of Class B decreased slightly (from 0.99 to 0.97).

Table 5. Summarizes classification performance across different qubit configurations

Qubits	**Accuracy**	**Precision**	**Recall**	***F1*- score**
4	94%	93%	94%	94%
6	98%	98%	97%	97%
8	97%	97%	97%	97%

Table 5 shows that, the optimal performance was achieved with 6 qubits, corresponding to six principal components. Increasing to 8 qubits slightly reduced performance, suggesting potential overfitting. Reducing to 4 qubits resulted in information loss and decreased accuracy.

Table 6. Comparison of results with other studies

References	**Feature mapping function**	**Classification**	**Acc (%)**
S. Vashisth *et al.* [8]		Linear support vector machine	89
		Nonlinear support vector machine	95
	H, U	Quantum support vector	92
Varsha Premanand *et al.* [9]	*ZZ* Feature map	Quantum support vector classifier	93.8
E. Akpinar *et al.* [10]	R_x	Quantum support vector machine	98
	R_x and CX		98
Our method	H, R_x, R_z and CX	Quantum support vector machine	98

Comparative analysis indicates that the proposed hybrid QSVM achieves performance equal to or better than previously reported QSVM implementations while requiring fewer effective features.

The experimental results show that the QSVM model achieves different performance levels when varying the number of qubits (4, 6, and 8 qubits), with the best accuracy of 98% obtained at 6 qubits, compared to 94% at 4 qubits and 97% at 8 qubits.

This trend indicates the existence of an optimal balance between representational dimensionality and model generalization capability. The lower performance with 4 qubits reflects information loss due to excessive dimensionality reduction, as the reduced feature space is no longer sufficient to optimally separate the benign and malignant classes. Conversely, increasing to 8 qubits does not improve performance and slightly reduces accuracy, which may result from a poorly conditioned quantum kernel matrix, or a higher risk of overfitting given the relatively small training dataset of 569 samples. This finding highlights that QSVM performance is not directly proportional to the number of qubits but strongly depends on the quality of feature representation. Importantly, the balanced Precision, Recall, and *F1*–scores across both classes indicate that the model does not exhibit significant bias toward the majority class, which is particularly critical in medical diagnosis tasks.

Furthermore, a key scientific contribution of this study lies in the two–stage dimensionality reduction strategy applied prior to quantum encoding. The first stage employs Pearson correlation filtering to remove multicollinearity and redundant features, thereby improving the numerical conditioning of the covariance matrix. The second stage applies (PCA) to project the data into an orthogonal subspace while preserving maximal variance. This combined approach reduces the original 30 features to 6 principal components, consequently decreasing the number of required qubits, reducing quantum circuit depth, enhancing kernel stability. These results reinforce the argument that, in the current NISQ era, classical preprocessing plays a decisive role in improving the effectiveness and feasibility of quantum machine learning models.

5. Conclusion

This study presents a hybrid Pearson–PCA enhanced QSVM framework for breast cancer classification. By combining classical feature selection and dimensionality reduction with quantum kernel methods, the proposed approach achieves 98% accuracy on the WDBC dataset using only six qubits. The results confirm that dimensionality reduction plays a critical role in quantum machine learning applications under hardware constraints. The integration of Pearson correlation filtering and PCA not only reduces qubit requirements but also improves classification robustness and suggest that classical preprocessing remains essential in hybrid quantum–classical workflows, particularly for biomedical datasets characterized by correlated features and moderate dimensionality.

Future research will explore applications to larger–scale electronic health record datasets and the development of adaptive hybrid quantum-classical architectures.

References

[1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, Globocan 2012 v1.0: Cancer incidence and mortality worldwide: IARC CancerBase No. 11 [Internet], Lyon, France: International Agency for Research on Cancer, 2013. [Online]. Available: http://globocan.iarc.fr, Accessed on: Oct. 10, 2017.

[2] L. A. Torre, R. L. Siegel, E. M. Ward, and A. Jemal, Global cancer incidence and mortality rates and trends–An update, Cancer Epidemiol, Biomarkers & Prevention, vol. 25, no. 1, pp. 16–27, Jan. 2016. http://doi.org/10.1158/1055-9965.EPI-15-0578

[3] O. D. Balogun and S. C. Formenti, Locally advanced breast cancer - Strategies for developing nations, Frontiers in Oncology, vol. 5, Apr. 2015, Art. no. 89. https://doi.org/10.3389/fonc.2015.00089

[4] Y. Gujju, A. Matsuo, and R. Raymond, Quantum machine learning on near-term quantum devices: current state of supervised and unsupervised techniques for real-world applications, 2023. https://doi.org/10.48550/arXiv.2307.00908

[5] R. Zhang, J. Wang, N. Jiang, H. Li, and Z. Wang, Quantum support vector machine based on regularized Newton method, Neural Networks, vol. 151, pp. 376–384, Jul. 2022. https://doi.org/10.1016/j.neunet.2022.03.043

[6] M. Schuld, R. Sweke, and J. J. Meyer, Effect of data encoding on the expressive power of variational quantum-machine-learning models, Physical Review A, vol. 103, no. 3, Mar. 2021, Art. no. 032430. https://doi.org/10.1103/physreva.103.032430

[7] J. E. Park, B. Quanz, S. Wood, H. Higgins, and R. Harishankar, Practical application improvement to quantum SVM: theory to practice, 2020. https://doi.org/10.48550/arXiv.2012.07725

[8] S. Vashisth, I. Dhall, and G. Aggarwal, Design and analysis of quantum powered support vector machines for malignant breast cancer diagnosis, Journal of Intelligent Systems, vol. 30, iss. 1, Sep. 2021. https://doi.org/10.1515/jisys-2020-0089

[9] V. Premanand, S. M. B. Snavya, S. Srinivas, and S. Reddy, Quantum machine learning for breast cancer detection: a comparative study with conventional machine learning methods, Indian Journal of Natural Sciences, vol. 14, iss. 78, pp. 57728–57733, Jun. 2023.

[10] E. Akpinar, S. M. N. Islam, and M. Oduncuoglu, Evaluating the impact of different quantum kernels on the classification performance of support vector machine algorithm: a medical dataset application, Jul. 2024. [Online] Available: https://arxiv.org/abs/2407.09930, Accessed on: Feb. 27, 2026.

[11] W. H. Wolberg, O. L. Mangasarian, W. N. Street, and W. Street, Breast cancer Wisconsin (diagnostic) dataset,

UCI Machine Learning Repository, 1995. [Online] Available: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic, Accessed on: Feb. 27, 2026.

[12] Breast cancer database, University of Wisconsin-Madison, 1995. [Online] Available: https://pages.cs.wisc.edu/~olvi/uwmp/cancer.html, Accessed on: Feb. 27, 2026.

[13] Qiskit machine learning overview, IBM Quantum, Qiskit Community, 2024. [Online] Available: https://qiskit-community.github.io/qiskit-machine-learning/, Accessed on: Feb. 27, 2026.

[14] G. Aleksandrowicz, T. Alexander, P. Barkoutsos, L. Bello, Y. Ben-Haim, D. Bucher, F. Cabrera-Hernández, J. Carballo-Franquis, A. Chen, C.-F. Chen, J. M. Chow, A. D. Córcoles, A. Cross, A. J. Ferris, J. A. Gambetta, J. Gunnels, I. Hamamura, S. King, A. Mezzacapo, S. M. Nation, J. Perez, M. Pistoia, E. Rieffel, D. Rodríguez, M. Ross, M. Suárez, S. Temme, and H. Abraham, Qiskit: An open-source framework for quantum computing, pp. 55–63, 2019. https://doi.org/10.1145/3360307.3414934

[15] S. Ibrahim, S. Nazir, and S. A. Velastin, Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis, Journal of Imaging, vol. 7, iss. 11, Oct. 2021, Art. no. 225. https://doi.org/10.3390/jimaging7110225