PIMA DIABETES PREDICTION USING MACHINE LEARNING AND QUANTUM MACHINE LEARNING TECHNIQUES

ABSTRACT

Quantum Machine Learning (QML) which is the integration of quantum mechanics and machine learning comes with immense computational capacity due to principles of Superposition and entanglement. The PIMA diabetes dataset is used in this research as a test bed for comparing the results obtained from ML and QML approaches. The assessment also shows that from the list of the original ML classifiers, SVM has better results compared to LR with an accuracy of 0. The optimized performance achieved 76 accuracy as well as comparable recall and F-measure which is 0. 75, 0. 76, and 0. 77, respectively. On the other hand, among QML algorithms; QSVC has a higher accuracy rate than VQC with 0. 74 and the precision, recall, and F-measure of 0. 75, 0. 73, and 0. 74, respectively. The findings of the given research imply that the use of QML methods may be valuable for further developments in diabetes prediction in the future.

Keywords - Machine Learning, Quantum Computing, LR, SVM, VQC, QSVC

1. INTRODUCTION

Diabetes is among the primary causes of death globally as well as in India. As per the report of WHO [1], the death rate due to diabetes mellitus by the age group 55-59 is the lowest and most in 85 and above as depicted in Figure 1. Whereas it is very low in the age group of 0-29 and low to moderate in the age group of 30-54. Diabetes, or metabolic disorder, is a multi-faceted category that includes several sub-types that have their unique cause, pathophysiology, and presentation.

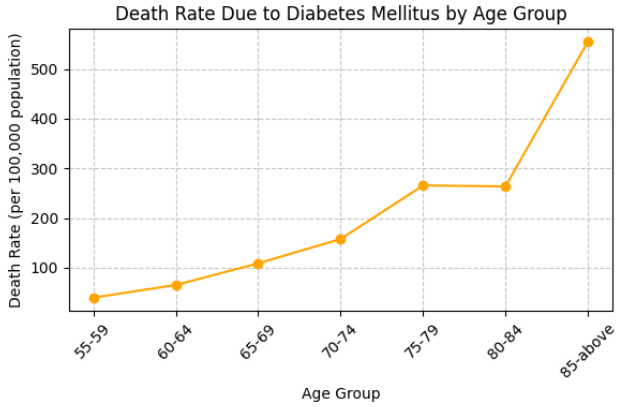

Figure 1 – Death Rate due to diabetes mellitus by age group.

Type I diabetes evokes the stage of autoimmune destruction of pancreatic cells of islets - beta-cells after which type I diabetes is the most common type of diabetes in childhood or adolescence. On the contrary, type 2 diabetes, as the most widely ramified one in the world, anticipates both insulin resistance and relative insulin deficiency, which are usually connected with irregular lifestyle and overweight tendency [2]. Pregnancy-induced diabetes mellitus (GDM) is also caused by hormonal changes that make the maternal insulin resistance prevalent, as these pose danger to both mother and fetus.

Besides this, there are several types of diabetes that are distinct, such as MODY [3] that gifts diabetes at a young age, diabetes secondary to pancreatic pathologies and induced drug effects, and endocrine disruptions that affect insulin secretion or action. Accurate classification of the types of diabetes and their origin is fundamental for the use of individualized management strategies, and timely interventions for the prevention and mitigation of risk of consequences. These strategies aim at improving patients' prognosis. By investigating the pathogenesis, subtle differentiations of clinical features, and idiosyncrasies between types of diabetes, healthcare specialists can then give patients care matching their unique needs, help them get their blood sugar under control, and empower them to live healthier lives. Figure 2 depicts the Prevalence rate of diabetes in adults according to WHO.

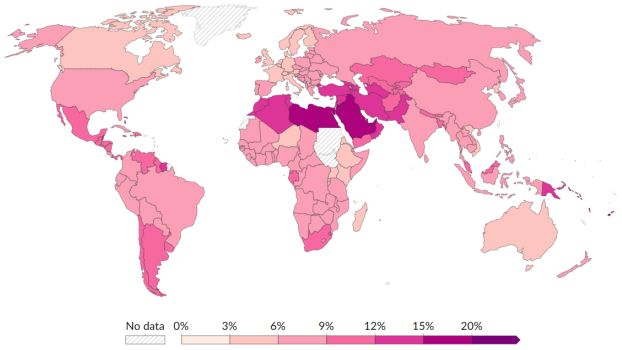

Figure 2 – WHO, Prevalence rate of diabetes in adults, 2014.

The world is facing increasing numbers of people suffering from diabetes, so the search for better-predicting tools and efficient methods of treatment becomes imperative. Machine learning and consequently QML technologies being among the most advanced ones, are expected to play a significant role in increasing the accuracy and the level of personalization in the respective field. This introduction will describe the importance of prediabetes screening technology and the role

of ML and QML in this new domain. Also, this technology would be able to change the healthcare management of diabetes for the better.

An expanded set of algorithms including logistic regression, SVM, decision trees, and random forests has been applied to diabetes prediction by means of labeled data that allows to build models capable of constructing classes reflecting various degrees of diabetes risk. They are really good in realizing the interconnections between inputs and outcomes of disease by which precise risk assessments can be carried out and then direct interventions are taken. Besides, ensemble learning approaches that exploit the cooperation of many models for settlement of superiority have well proven their considerable role in elevating the competence and general correlation of predictors of diabetes. The Contribution of this paper is as follows:

- Conducted experiments on the Indian PIMA diabetes dataset to evaluate the algorithms.
- Addressed data imbalance using Synthetic Minority Over-sampling Technique (SMOTE) for improved model accuracy.
- Applied Principal Component Analysis (PCA) for feature extraction, enhancing computational efficiency and predictive performance.
- Analyzed the performance of Machine Learning (ML) and Quantum Machine Learning (QML) algorithms for diabetes prediction.

The rest of the paper consists of the following sections: Section 2 discusses machine learning techniques; Section 3 includes basics of quantum computing; Section 4 consists of various Quantum Machine Learning techniques; Section 5 discusses implementation; Section 6 consists of Results and discussion and finally conclusion and future work is discussed in section 7 and 8 respectively.

2. MACHINE LEARNING TECHNIQUES

Through data mining and AI-based research, diabetes prediction is an important component. The researchers employ machine learning and deep learning techniques [4] to name the common diabetes characteristics making use of the gigantic datasets, highly precisely and stably.

These methods help to determine the pathogenic factors and physiological elements. They offer innovative chances for diabetes screening and diagnosis as well. Imaging for this critically lethal condition being prominent in medical reports makes diabetes still a priority for research and hence a huge data generator. Data mining is a valuable tool in data discovery and utilization, taking the form of both descriptive and predictive mining methods. Machine learning which symbolizes artificial intelligence becomes better and better to learn how to perform tasks like the human brain in automatic pattern recognition. Deep learning being a subset of machine learning behaves exactly as we process data in our brains

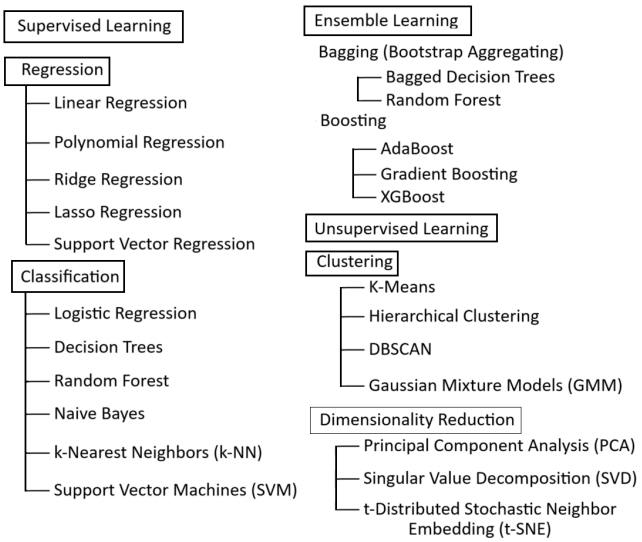

Figure 3 – Machine Learning Techniques Classification.

capability of identifying intricate patterns has completely changed the face of AI applications. The classification techniques of ML are depicted in Figure 3.

2.1 Logistic Regression

Logistic Regression(LR) as a tool for binary classification is a statistical method. It approximates the probability that an input belongs to a categorical class using the logistic function, which is essentially a mapping between input features with the domain $(0, 1)$. The model learns the coefficients to yield the best fit, hence it can establish the relationship between input and output given the data set. In terms of mathematics, the probability of this prediction ($\theta(\mathbf{x})$) is given by

$$\theta(\mathbf{x}) = \frac{1}{1 + e^{-\beta^T \mathbf{x}}} \tag{1}$$

where β are coefficients, and $\mathbf{x}$ is the input vector. These coefficients are then optimized using methods such as Maximum Likelihood Estimation, and the model is trained to accurately predict the output based on input variables.

2.2 Support Vector Machine

The Support Vector Machine (SVM) is a supervised machine learning algorithm designed for the purpose of classifying cases into two distinct classes. In binary classification, we have a dataset of n feature vectors $\mathbf{x}_i$ and corresponding targets y_i together with labels, and SVM aims to find the hyperplane represented by $\mathbf{w}$ and b which maximizes the margin between the classes. Here, let's denote $\mathbf{w}$ as the weight vector, b for the bias, and $\mathbf{x}$ the input feature vector, formulated as -

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \tag{2}$$

that are to satisfy $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$ for all i, and its decision function can be written as -

$$f(\mathbf{x}) = |\mathbf{w} \cdot \mathbf{x} + b| \tag{3}$$

which, if $f(\mathbf{x}) \geq 0$, is the prediction of the class label for input $\mathbf{x}$.

2.3 Principal Component Analysis

PCA is a method for dimensionality reduction and for visualising data, transforming original variables into orthogonal vectors called principal components, and maximizing data variance. Given an $n \times p$ data matrix $\mathbf{X}$, PCA computes eigenvectors and eigenvalues of its covariance matrix. The first principal component, $\mathbf{PC}_1$, is the linear combination of variables maximizing variance, with subsequent components $\mathbf{PC}_2, \mathbf{PC}_3, \ldots$ orthogonal to preceding ones, capturing remaining variance. PCA's essence lies in expressing data in terms of these components, effectively reducing dimensionality while preserving the most significant information. Mathematically, PCA computes $\mathbf{X}_{\text{pca}} = \mathbf{XV}$, where $\mathbf{X}_{\text{pca}}$ contains principal component scores, $\mathbf{V}$ comprises eigenvectors, and $\mathbf{X}$ represents original data.
.

3. QUANTUM COMPUTING

Quantum computing [5] represents a rapidly evolving domain harnessing the principles of quantum mechanics to execute computational tasks. In contrast to classical computers reliant on binary bits (0s and 1s), quantum computers utilize quantum bits, or qubits. Qubits possess the unique ability to occupy multiple states simultaneously through superposition and entanglement phenomena, thereby empowering quantum computers to handle extensive data volumes and execute specific calculations with remarkable efficiency compared to classical counterparts. To define the concept of quantum computing mathematically, Let $\mathcal{H}$ denote the Hilbert space associated with the quantum computing system. A quantum computer operates by manipulating qubits, which are represented as vectors in $\mathcal{H}$. Each qubit can be in a superposition of basis states, denoted by $|0\rangle$ and $|1\rangle$, where $|0\rangle$ represents the state corresponding to the logical value 0, and $|1\rangle$ represents the state corresponding to the logical value 1.

3.1 Superpostion

Quantum states can be represented as linear combinations of basis states, allowing qubits to exist in a superposition of states $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. Mathematically, for a single qubit $|\psi\rangle$, superposition is expressed as:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{4}$$

where α and β are complex probability amplitudes satisfying $|\alpha|^2 + |\beta|^2 = 1$, enabling the representation of both 0 and 1 simultaneously.

3.2 Entanglement

Entanglement means that when two qubits are correlated,regardless of their physical separation,one qubit's state depends on the other's state. Mathematically, for two qubits $|\psi\rangle$ an entangled state can be represented as:

$$\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle) \tag{5}$$

3.3 Quantum Gates

Quantum gates are unitary operators that manipulate qubits to perform specific operations. Analogous to classical logic gates, quantum gates serve as the building blocks of quantum algorithms. Mathematically, a quantum gate U operates on a qubit $|\psi\rangle$ as $U|\psi\rangle = |\psi'\rangle$.

4. QUANTUM MACHINE LEARNING (QML)

QML is a widening field of study where quantum computing unites with machine learning. In a nutshell, QML puts the properties/concepts of quantum mechanics into use to design and invent new machine-learning algorithms with associated techniques. The classification techniques of QML are depicted in Figure 4.

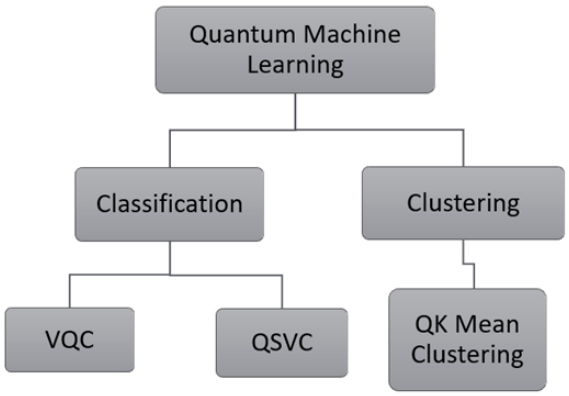

Figure 4 – Major QML Techniques Classification.

As far as mathematically QML might use the quantum computer's unique characteristics, which are superposition, entanglement, and interference in overcoming the classic learning tasks more efficiently. With implementation of quantum algorithms and quantum data representations in QML, it targets to address hard problems of this domain, which include optimization, pattern recognition and data analysis. The primary purpose of QML is to open up the marine of mind-blowing technologies that quantum computing application might contribute to enhancing the efficiency and precision of machine learning systems.

4.1 Variational Quantum Classifier (VQC)

VQC [6] is a QML algorithm that uses regularized PQCs (parameterized quantum circuits) to classify inputs. Let a set of input features denoted by $\mathbf{x}$ and the corresponding class labels given by $\mathbf{y}$, the VQC encodes the single feature into a quantum state with the trainable system parameters $\boldsymbol{\theta}$ via parameterized quantum circuit $U(\boldsymbol{\theta})$. It [10] encodes data in classical form into quantum representations that are used next for quantum computations. Lastly, a quantum measuring

procedure is performed, generating results which are used as classifiers. The optimization of the parameters $\boldsymbol{\theta}$ is done to get a minimal cost function $J(\boldsymbol{\theta})$, which characterizes the disparity between the true labels and the predicted ones. Generally speaking, neural networks employ classical optimization algorithms like gradient descent. Quantum computers are powerful machines that can be trained to solve classification problems characterized by a large number of measurements.

4.2 Quantum Support Vector Classifier (QSVC)

The QSVC [7] is a method that is explicit in the following way: the quantum kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ is defined as a quantum circuit $U(\mathbf{x}_i, \mathbf{x}_j)$ mapping input data points $\mathbf{x}_i$ and $\mathbf{x}_j$ into the higher-dimensional quantum feature space. Allegorically, this transformation can be designated as $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ mathematically. In this quantum feature space, the inner product $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$ is actually the quantum kernel $K(\mathbf{x}_i, \mathbf{x}_j)$, which is responsible for expressing the similarity between input points in terms of their corresponding quantum states. Next, QSVC [11] chooses a certain hyperplane in a quantum feature space, which aims to minimize classification errors and obtain the greatest distance between classes. Thus, we may formulate it as a quadratic programming problem where we seek to minimize the objective function given the established constraints.

4.3 Quantum K-Mean Clustering

Quantum K-Mean Clustering (QKMC) [8] enables the laws of quantum mechanics to perform clustering operations. In QKMC, the data vectors are mapped to quantum states by phases using a circuit of quantum interconnections, with each data point being represented by a quantum state in a higher-dimensional Hilbert space. This wiring is conveyed mathematically as $|\psi_i\rangle = U(\mathbf{x}_i)|0\rangle$, where $\mathbf{x}_i$ represents the i-th data set, and $U(\mathbf{x}_i)$ is a quantum circuit that encodes data $\mathbf{x}_i$ into states. The QKMC teams carry out the process of quantum operations periodically to group the quantum states into clusters, but the distance between the intra-states and distance between the inter-states have to be minimized and maximized respectively. In QKMC, the optimization process is highly complicated, where the position of the centroids of clusters occurs by adjusting the parameters of the quantum circuit in order to minimize a distance function.

5. PERFORMANCE EVALUATION AND IMPLEMENTATION

In classification problems, performance evaluation metrics [9] are essential for defining the success of predictive models by the fact of dividing data into previously set classes. Metrics which are usually used comprise accuracy, precision, recall, and F1 score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$\text{Precision}(P) = \frac{TP}{TP + FP} \tag{7}$$

$$\text{Precision}(N) = \frac{TN}{TN + FN} \tag{8}$$

$$\text{Sensitivity} = \text{Recall}(P) = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Specificity} = \text{Recall}(N) = \frac{TN}{TN + FP} \tag{10}$$

$$\text{F-Measure} = \text{F-1 Score} = \frac{2 \cdot \text{Precision}(P) \cdot \text{Recall}(P)}{\text{Precision}(P) + \text{Recall}(P)} \tag{11}$$

The Methodology followed in this paper is written in the below Algorithm, which indicate the several steps for the Classification of PIMA Diabetes Dataset using ML and QML techniques.

Algorithm : PIMA Diabetes Prediction using ML and QML

Step 1: $D \leftarrow$ Load PIMA Diabetes Dataset

Step 2: $D_{\text{balanced}} \leftarrow \text{SMOTE}(D)$

Step 3: $\text{EDA} \rightarrow D_{\text{balanced}}$

Step 4: $D_{\text{reduced}} \leftarrow \text{PCA}(D_{\text{balanced}})$

Step 5: ML and QML Classification Model Selection

Step 6: Model Training and Validation

$D_{\text{train}}, D_{\text{val}} \leftarrow \text{Split}(D_{\text{reduced}})$

$M_{\text{trained}} \leftarrow \text{Train}(M, D_{\text{train}})$

$\text{Validation} \rightarrow M_{\text{trained}}, D_{\text{val}}$

Step 7: Model Evaluation using Performance Metrics:

$\text{Accuracy}(A): \quad A \leftarrow \text{Accuracy}(M_{\text{trained}}, D_{\text{val}})$

$\text{Precision}(P): \quad P \leftarrow \text{Precision}(M_{\text{trained}}, D_{\text{val}})$

$\text{Recall}(R): \quad R \leftarrow \text{Recall}(M_{\text{trained}}, D_{\text{val}})$

$\text{F1 Score}(F): \quad F \leftarrow F1\_Score(M_{\text{trained}}, D_{\text{val}})$

Step 8: Comparative Analysis of Model Performance

As the first step, the algorithm takes in the PIMA Diabetes Dataset D. Later, it shows the way data imbalance is overcome by means of the Synthetic Minority Over-sampling Technique (SMOTE) which simulates a balanced dataset denoted as D_{balanced}. Then EDA is conducted on the balanced dataset so as to understand the description of the data. In order to reduce dimensions and improve computing elasticity, Principal Component Analysis (PCA) is performed to generate D_{reduced}. The model selection includes classical ones like Logistic Regression (LR) and Support Vector Machine (SVM) and on the other side quantum models such as Variational Quantum Circuit (VQC) and Quantum Support Vector Classifier (QSVC). Next, we separate the dataset into 2 sets, train (D_{train}) and validation (D_{val}) and then train the learners with the training data. Model validation involves evaluating the results of each model on a hold-out validation set. Performance metrics such as Accuracy (A), Precision (P), Recall (R), and F1 Score (F) are obtained to make a judgment about the performance. Furthermore, the algorithm generates a comparison between model outputs and then

wraps up the summary of the major findings and possible research directions or applications.

We have used two Python libraries for the implementation of the above algorithm: sklearn library [12] for Machine Learning models (LR and SVM) and Qiskit library [13] for Quantum Machine Learning models (VQC and QSVC). VQC uses quantum feature map, like ZZFeatureMap to encode input features to quantum states, and ansatz (initial guess) is used in the Quantum circuit, referring real amplitude for shaping quantum states as depicted in Figure 5. The objective function value over iterations is shown in Figure 6.

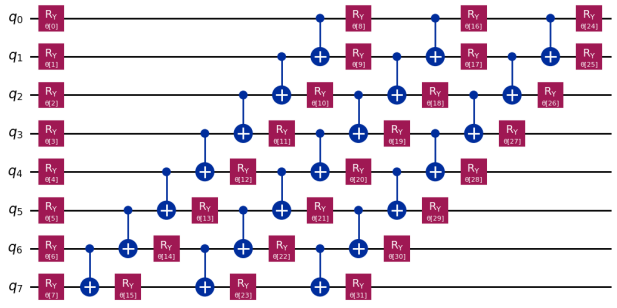

Figure 5 – VQC Real Amplitude.

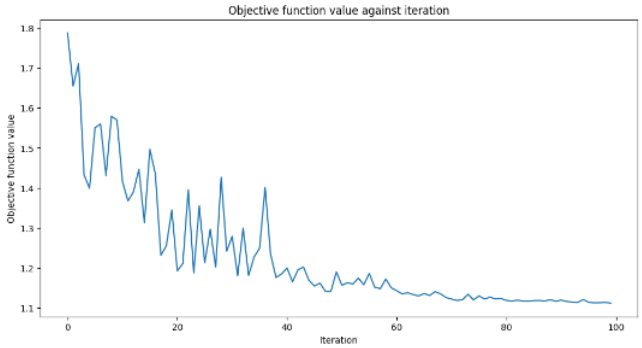

Figure 6 – Training over iterations.

Quantum Support Vector Classifier (QSVC) uses a feature map and kernel that are specified to match the dimensions of the input features for specific classification problems. The classifier employs quantum computing methods to learn and make accurate predictions.

6. RESULT AND DISCUSSION

A performance comparison analysis of various classifiers on the Indian PIMA Diabetes Dataset indicates some distinct behaviors and consequences, according to the experimental assessment carried out on the dataset. The results of the performance matrices show that Support Vector Machine (SVM), with accuracy values of 0.76 and matching precision, recall, and F-measure results of 0.75, 0.76, and 0.77, respectively, is the best combination of the classification methods examined. This suggests that SVM has good capability in identifying cases in the dataset that are either diabetic or not.For the most part, LR (Logistic Regression) produces results that are equivalent to SVM, lagging behind it by a small margin with an accuracy of 0.74 and balanced precision (0.73), recall (0.73), and F-measure (0.75), respectively. In particular, the F-measure in LR is superior to its precision, suggesting that the number of accurately detected positive cases and the number of false alarms that are avoided are in balance as depicted in Figure 7.

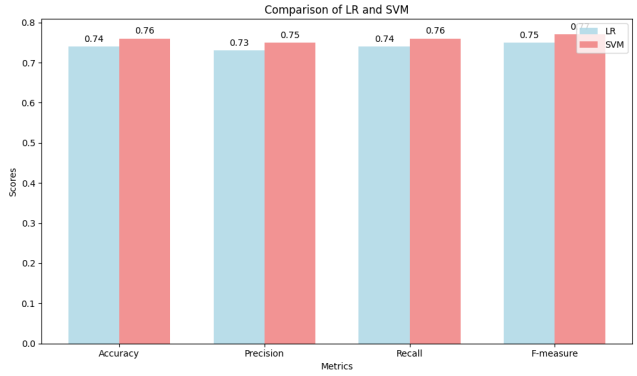

Figure 7 – Comparative Analysis of LR and SVM.

The experimental results reveals that among QML algorithms,Quantum Support Vector Classifier (QSVC) showed a relatively good performance over indian PIMA diabetes dataset with an accuracy of 0.74 and the precision,recall,F-measure of 0.75,0.73,0.74 respectively (Figure 8). Whereas the Variational Quantum Classifier (VQC) has the lowest accuracy of 0.61 and precision,recall, F-measure of 0.60,0.61 and 0.59 respectively which poses a lesser value compared with the other classifiers.

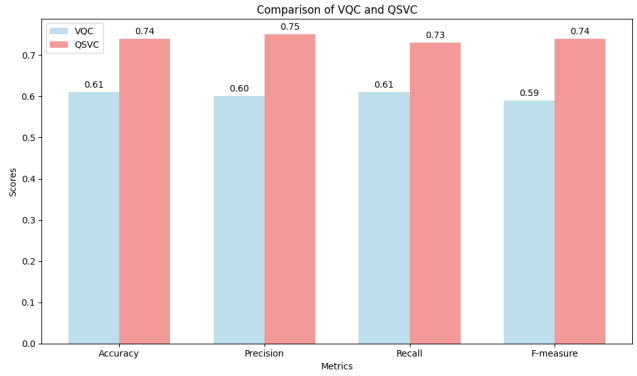

Figure 8 – Comparative Analysis of VQC and QSVC.

Therefore, it is inferred that VQC is somehow less effective in identifying whether a case of diabetes exists in the given data set. These results demonstrate SVM's superiority in accuracy and, in particular, the practicality of LR and QSVC for the diabetes classification problem, while also emphasizing the necessity for VQC to keep enhancing its accuracy.

7. CONCLUSION

This article has presented a comprehensive overview of ML and QML techniques.The nexus of quantum computing with machine learning is known as QML. The study evaluated the potential of QML for the classification of diabetes using the PIMA Indian Diabetes Dataset by comparing quantum-enhanced algorithms (QSVC, VQC) with their classical counterparts (LR, SVM). Future directions is to Explore more sophisticated QML algorithms for diabetes classification. Investigate the impact of larger datasets on QML performance compared to classical methods. Analyze the computational efficiency of QML algorithms for real-world applications.

8. FUTURE WORK

Further research prospects touching on the use of QML algorithms for diabetes classification are to pursue modified algorithms with higher effectiveness. As such, it is also necessary to study the work of QML algorithms when processing larger datasets than the classical approach. Besides, answers to the questions regarding the computational complexity of QML algorithms for real-world applications will be obtained. Another significant branch of the subjected research is considering methodology of applying hybrid classical-quantum models combining the advantages of two approaches. Thus, the application of both approaches simultaneously could increase predictive accuracy and computational velocity, so QML would be more suitable to solve big medical data analysis problem.

REFERENCES

[1] "Diabetes," https://www.who.int/news-room/fact-sheets/detail/diabetes, Accessed: Apr. 05, 2023.

[2] "Diabetes Cure for Type1, Type 2 and Type 3 (LADA)," *International Journal of Diabetes & Metabolic Disorders*, 2018.

[3] "Diabetes: Mody," https://www.nationwidechildrens.org/conditions/diabetes-mody#, Accessed: Apr. 14, 2024.

[4] R.I. Mukhamediev et al., "Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges," *Mathematics*, 2022.

[5] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2010.

[6] D. Maheshwari et al., "Variational Quantum Classifier for Binary Classification: Real vs Synthetic Dataset," *IEEE Access*, vol. 10, pp. 3705-3715, 2022.

[7] M. Aly et al., "Experimental Benchmarking of Quantum Machine Learning Classifiers," in *2023 6th International Conference on Signal Processing and Information Security (ICSPIS)*, pp. 240-245, 2023.

[8] Z. Shao, S. Zhang and S. Kumar, "Quantum K-Means Model based on Optimization Framework," in *2023 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC)*, Dalian, China, pp. 372-376, 2023, doi: 10.1109/IPEC57296.2023.00071.

[9] I. M. De Diego et al., "General Performance Score for classification problems," *Applied Intelligence*, vol. 52, pp. 12049-12063, 2022.

[10] G. Laskaris et al., "Comparison between Tensor Networks and Variational Quantum Classifier," 2023.

[11] S. Sridevi, B. Indira, S. S. Dutta, S. Sandeep and A. Sreenivasan, "Quantum Enhanced Support Vector Machine with Instantaneous Quantum Polynomial Encoding for Improved Cyclone Classification," in *2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)*, Chennai, India, pp. 748-752, 2023, doi: 10.1109/ICRTAC59277.2023.10480791.

[12] "Scikit-learn: Machine learning in Python — Scikit-learn 1.4.2 documentation," https://scikit-learn.org/stable/.

[13] "IBM Quantum Computing | Qiskit," https://www.ibm.com/quantum/qiskit.