WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Comparative Analysis of Hybrid Machine Learning Models for Early-Stage Diabetes and Cardiovascular Disease Prediction

Aman [a*], Rajender Singh Chhillar [b]

[a, b] *Department of Computer Science and Applications, M.D. University, Rohtak (124001), Haryana, India*

A B S T R A C T

Early detection of Type 2 Diabetes Mellitus (T2DM) and cardiovascular diseases (CVD) is critical for reducing global morbidity and mortality rates. This study presents a comprehensive comparative analysis of two advanced machine learning models designed for early-stage disease prediction. Model-1 employs a stacking ensemble architecture combining logistic regression, naïve Bayes, AdaBoost, support vector machines (SVM), artificial neural networks (ANN), and k-nearest neighbors (k-NN), aggregated via a random forest meta-classifier for T2DM prediction. Model-2 integrates a long short-term memory (LSTM) network with a quantum neural network (QNN), optimized using a self-improved aquila optimization (SIAO) algorithm for CVD prediction. The analysis evaluates performance metrics, computational efficiency, adaptability to diverse datasets, and practical implications for healthcare applications. Results demonstrate Model-1's exceptional accuracy (99.72%) and low false positive rate, while Model-2 achieves robust performance on imbalanced datasets (96.69% accuracy) despite higher resource demands. The study highlights trade-offs between model complexity, data requirements, and operational feasibility, offering actionable insights for medical practitioners and researchers.

Keywords: Ensemble Learning, Quantum-Inspired Neural Networks, Cardiovascular Diseases, Type 2 Diabetes Mellitus, Imbalanced Data

1. Introduction

1.1 Background and Motivation

Chronic diseases, particularly T2DM and CVD, represent significant public health challenges worldwide. Early diagnosis is pivotal for initiating timely interventions, improving patient outcomes, and reducing healthcare costs. Traditional diagnostic methods often rely on invasive procedures and clinical expertise, which may delay detection. Machine learning (ML) has emerged as a transformative tool in healthcare analytics, enabling automated, data-driven predictions (Aman and Chhillar 2020; Janiesch, Zschech, and Heinrich 2021). However, challenges such as dataset imbalance, high dimensionality, and variability in data quality persist, necessitating tailored solutions.

The global prevalence of T2DM has risen dramatically in recent decades, affecting approximately 422 million people worldwide according to the World Health Organization (Khan et al. 2020). Early detection allows for lifestyle interventions and pharmacological treatments that can significantly slow disease progression. Similarly, CVDs account for nearly 32% of global deaths annually, with early identification of at-risk patients enabling preventive measures and reducing the likelihood of severe cardiac events (Abidi et al. 2023; Aman and Chhillar 2021).

Machine learning offers a promising avenue for addressing these challenges by analyzing complex patterns in clinical data. Ensemble learning and hybrid architectures have gained prominence for their ability to enhance predictive accuracy and robustness. Ensemble methods combine multiple algorithms to leverage their individual strengths, while hybrid models integrate heterogeneous techniques to address specific data complexities (Abdollahi and Nouri-Moghaddam 2022; Darolia et al. 2024). The primary objectives of this research are:

- To develop a stacking ensemble model for T2DM prediction using symptom-based and clinical datasets.
- To design a quantum-inspired hybrid model for CVD prediction optimized for high-dimensional, imbalanced data.
- To conduct a comparative analysis of the models' performance, adaptability, and computational efficiency.
- To provide recommendations for selecting appropriate models based on healthcare resource availability and data characteristics.

The remainder of this paper is organized as follows: Section 2 presents the methodology, including detailed descriptions of both models, their architectures, and the datasets used. Section 3 reports the results of the comparative analysis, while Section 4 discusses the implications of these findings. Finally, Section 5 provides conclusions and suggestions for future research directions.

2. Methodology

2.1 Model-1: Stacking Ensemble for Diabetes Prediction

Model-1 (Aman and Chhillar 2023) adopts a hierarchical stacking ensemble framework to enhance predictive accuracy for T2DM. The architecture (Fig. 1) comprises two layers: base learners and a meta-classifier.

2.1.1 Base Learners

Five diverse algorithms were selected as base learners to capture varied patterns in the data, ensuring a comprehensive understanding of underlying relationships. Logistic Regression (LR), a widely used linear model, was included for its ability to estimate probabilistic outcomes and provide interpretability in classification tasks. Naïve Bayes (NB), a probabilistic classifier based on Bayes' theorem, was chosen for its efficiency and effectiveness despite its assumption of feature independence. AdaBoost combined with Support Vector Machine (AdaBoost + SVM) was incorporated as an ensemble method that enhances SVM's decision boundaries through adaptive boosting, improving its ability to handle complex patterns. Artificial Neural Network (ANN), implemented as a multi-layer perceptron with non-linear activation functions, was selected to capture intricate feature interactions and complex non-linear relationships in the dataset. Finally, k-Nearest Neighbors (k-NN), a non-parametric instance-based learning algorithm, was utilized to classify data points based on similarity, leveraging distance-based decision-making. Each of these base learners was trained independently on the diabetes dataset, generating intermediate predictions that contributed to the overall learning process. The selection of these algorithms was strategically made to encompass a broad spectrum of modeling techniques, ranging from linear to non-linear approaches and from parametric to non-parametric methods, ensuring diverse perspectives on the data and improving predictive performance.

2.1.2 Meta-Classifier

A random forest (RF) meta-classifier aggregated the predictions from the base learners. The RF algorithm constructs multiple decision trees and outputs the mode of their predictions, reducing overfitting and improving generalization. By combining the outputs of the base models, the meta-classifier leverages their collective strengths while mitigating individual weaknesses.

2.1.3 Preprocessing and Optimization

To address class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied to enhance the representation of diabetic cases. Feature normalization was performed using Z-score standardization to ensure consistent scaling across all features. Hyperparameter tuning was conducted through grid search with 10-fold cross-validation to optimize the parameters for each base learner as well as the meta-classifier. The model was implemented in the WEKA machine learning workbench, utilizing its integrated tools for ensemble construction and evaluation.

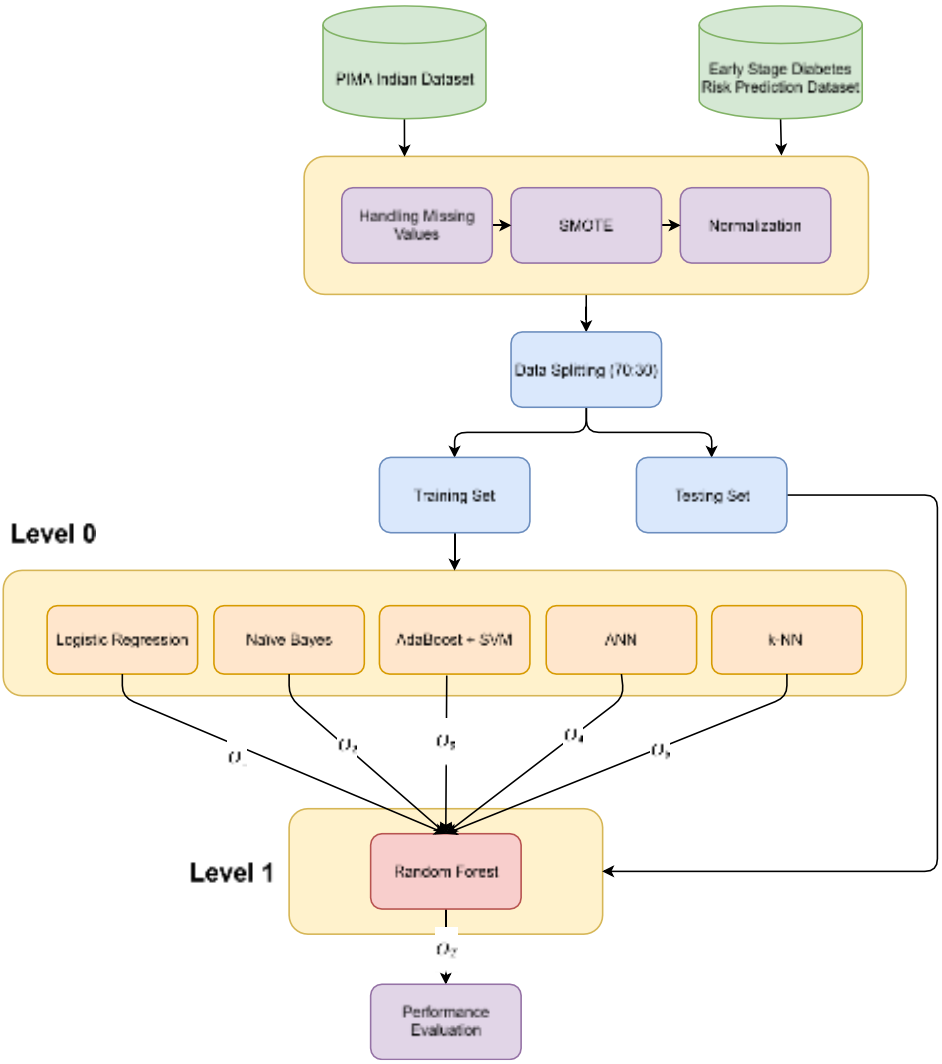

Fig. 1. Architecture of Model-1

2.2 Model-2: Quantum-Inspired Hybrid Model for CVD Prediction

Model-2 (Darolia et al. 2024) integrates sequential and quantum-inspired learning to handle the complexity of CVD datasets (Fig. 2).

2.2.1 LSTM Component

A long short-term memory (LSTM) network processed sequential patient data to capture temporal dependencies. The LSTM layer comprised 64 units with hyperbolic tangent and sigmoid activation functions, followed by dropout regularization to prevent overfitting. This component was designed to handle the temporal nature of medical records, where the sequence of clinical events can be crucial for accurate prediction.

2.2.2 Quantum Neural Network (QNN) Component

The QNN modeled feature correlations and uncertainties using quantum circuit principles. Parameterized quantum circuits were designed with rotation and entanglement layers, optimized for binary classification tasks. This quantum-inspired element aimed to enhance the model's ability to handle high-dimensional data by exploiting quantum parallelism and entanglement.

2.2.3 Optimization Algorithm

The self-improved aquila optimization (SIAO) algorithm optimized feature selection and weight tuning. SIAO, inspired by the hunting behavior of aquila birds, balances exploration and exploitation to converge on optimal solutions. This algorithm was chosen for its efficiency in navigating complex optimization landscapes, crucial for tuning the numerous parameters in the hybrid architecture.

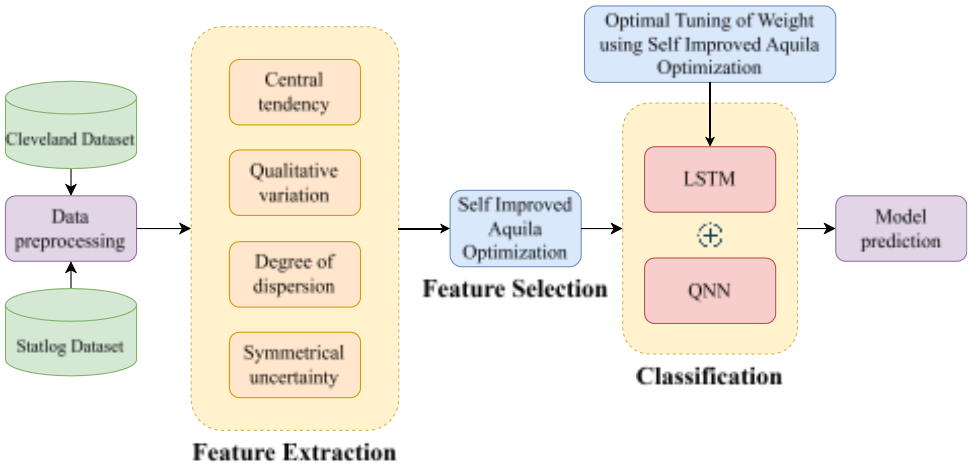

Fig. 2. Architecture of Model-2

2.2.4 Implementation

The hybrid model was developed in Python using TensorFlow for the LSTM component and Qiskit for quantum circuit simulations. Hyperparameters were tuned to balance computational cost and model performance. Comparative analysis of Model-1 and Model-2 depicted in Fig. 3.

2.3 Datasets and Preprocessing

2.3.1 Diabetes Datasets

Diabetes-Dataset-1 (Anon n.d.-a) consists of 768 female patient records, containing attributes such as plasma glucose, BMI, insulin levels, age, and pregnancies. The dataset features a binary target variable, with 34.8% of cases classified as positive. On the other hand, Diabetes-Dataset-2 (Anon 2020) includes 520 instances with 16 predictive features, integrating both behavioral symptoms such as polyuria and polydipsia along with clinical indicators for the early-stage detection of diabetes.

2.3.2 CVD Datasets

CVD-Dataset-1 (Andras Janosi 1989) comprises 303 patient records with 13 features, including age, sex, chest pain type, blood pressure, and cholesterol levels. The dataset exhibits a moderate class imbalance. In contrast, CVD-Dataset-2 (Anon n.d.-b) contains 270 records with demographic diversity, offering additional validation to enhance model robustness.

2.3.3 Data Preprocessing

For both models, missing values were imputed using median/mode substitution. Categorical variables were encoded via one-hot encoding, and feature scaling was applied to normalize input ranges.

3. Results

3.1 Performance Metrics

3.1.1 Model-1 (Diabetes Prediction)

Model-1 demonstrated exceptional performance on Diabetes-Dataset-2, achieving an accuracy of 99.72%. The precision and recall were both 0.997, indicating high reliability in positive predictions and strong sensitivity for early-stage detection. The false positive rate (FPR) was remarkably low at 0.3%, which is critical for minimizing unnecessary interventions and reducing patient anxiety.

3.1.2 Model-2 (CVD Prediction)

Model-2 achieved an accuracy of 96.69% on CVD-Dataset-1. The precision was 0.9603, and the recall was 0.9662, reflecting effective identification of high-risk patients even in imbalanced datasets. These results highlight the model's robustness in handling the complexities of CVD data.

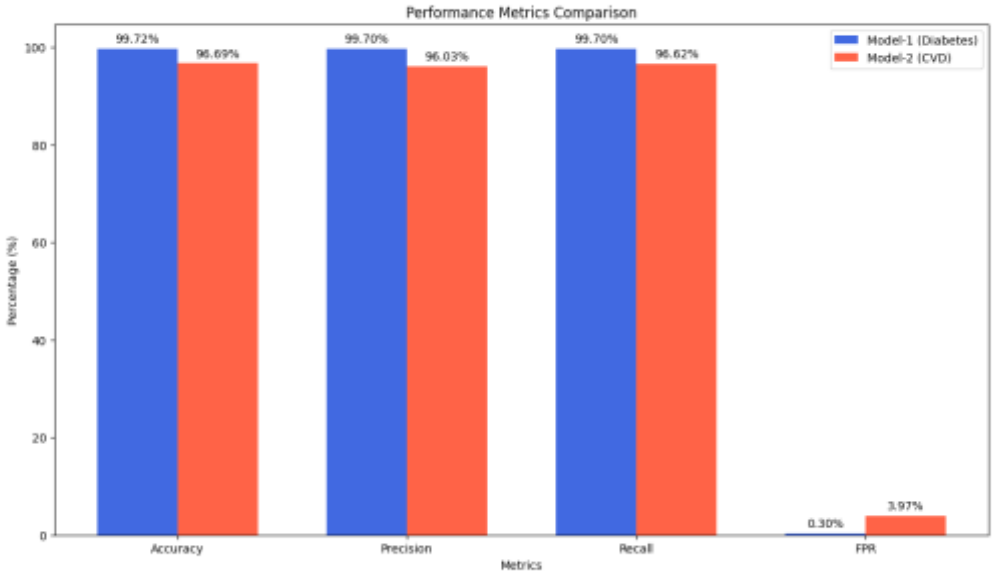

Fig. 3. Performance Comparison of Model-1 and Model-2

3.2 Computational Efficiency

Model-1 required approximately 15 minutes of training time with a memory usage of 4–6 GB. This efficiency makes it suitable for deployment in environments with limited computational resources. In contrast, Model-2 demanded significantly more resources, with training taking around 4 hours and memory usage ranging from 12–16 GB. This higher computational load is attributed to the complexity of the hybrid architecture, particularly the quantum-inspired components and the optimization algorithm.

3.3 Adaptability

Model-1 showed versatility in handling diverse data types, including both clinical measurements and behavioral symptoms. This adaptability makes it suitable for various healthcare settings where different types of data may be available. Model-2, while optimized for high-dimensional data, required careful preprocessing and feature selection to achieve optimal performance, highlighting its specialization for complex datasets.

4. Discussion

4.1 Strengths and Limitations

4.1.1 Model-1: Stacking Ensemble

The stacking ensemble approach of Model-1 provided a significant boost in accuracy and reliability for T2DM prediction. By combining multiple base learners, the model was able to capture a wide range of patterns in the data, leading to high precision and recall. The low FPR is particularly noteworthy, as it reduces the risk of misdiagnosis and unnecessary follow-up procedures. However, the model's performance is contingent on the quality and quantity of the training data. Smaller datasets or those with significant noise may limit its effectiveness.

4.1.2 Model-2: Hybrid Architecture

Model-2's integration of LSTM and QNN components allowed it to handle the sequential and high-dimensional aspects of CVD data effectively. The quantum-inspired elements enhanced the model's ability to explore feature correlations that might be missed by classical algorithms. The SIAO optimization further improved parameter tuning, contributing to its robust performance on imbalanced datasets. Nevertheless, the computational demands of this model present a barrier to deployment in resource-constrained settings. Additionally, the complexity of the architecture may require specialized expertise for maintenance and interpretation.

4.2 Comparative Insights

The choice between these models depends on the specific context of application. Model-1 offers a pragmatic solution for T2DM screening in primary care settings where computational resources are limited. Its efficiency and accuracy make it ideal for large-scale population screening programs. Model-2, while less efficient, provides deeper insights into complex CVD datasets, making it valuable in specialized cardiac units with access to advanced computational infrastructure. The trade-offs between accuracy, computational efficiency, and adaptability must be carefully considered based on the healthcare environment and data characteristics.

5. Conclusion and Future Work

This study underscores the potential of ensemble and hybrid ML models in early disease prediction. Model-1's stacking ensemble architecture offers a pragmatic solution for T2DM detection with minimal computational overhead, whereas Model-2's integration of LSTM and QNN provides a powerful tool for CVD analysis despite its resource intensity. Future research could explore hybrid ensemble-quantum frameworks to synergize the strengths of both approaches. Additionally, federated learning paradigms may address data scarcity and privacy concerns, enabling broader applicability of these models in global healthcare systems.

References

Abdollahi, Jafar, and Babak Nouri-Moghaddam. 2022. "Hybrid Stacked Ensemble Combined with Genetic Algorithms for Diabetes Prediction." Iran Journal of Computer Science 5(3):205–20. doi: 10.1007/s42044-022-00100-1.

Abidi, Mustufa Haider, Usama Umer, Syed Hammad Mian, and Abdulrahman Al-Ahmari. 2023. "Big Data-Based Smart Health Monitoring System: Using Deep Ensemble Learning." IEEE Access 11:114880–903. doi: 10.1109/ACCESS.2023.3325323.

Aman, and Rajender Singh Chhillar. 2020. "Disease Predictive Models for Healthcare by Using Data Mining Techniques: State of the Art." International Journal of Engineering Trends and Technology - IJETT 68(10):52-57,. doi: 10.14445/22315381/IJETT-V68I10P209.

Aman, and Rajender Singh Chhillar. 2021. "Analyzing Predictive Algorithms in Data Mining for Cardiovascular Disease Using WEKA Tool." International Journal of Advanced Computer Science and Applications 12(8, Art. no. 8):31. doi: 10.14569/IJACSA.2021.0120817.

Aman, and Rajender Singh Chhillar. 2023. "Optimized Stacking Ensemble for Early-Stage Diabetes Mellitus Prediction." International Journal of Electrical and Computer Engineering (IJECE) 13:7048–55. doi: 10.11591/ijece.v13i6.pp7048-7055.

Andras Janosi, William Steinbrunn. 1989. "Heart Disease."

Anon. 2020. "Early Stage Diabetes Risk Prediction."

Anon. n.d.-a. "Pima Indians Diabetes Database." Retrieved November 15, 2022 (https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database).

Anon. n.d.-b. "Statlog (Heart)."

Darolia, Aman, Rajender Singh Chhillar, Musaed Alhussein, Surjeet Dalal, Khursheed Aurangzeb, and Umesh Kumar Lilhore. 2024. "Enhanced Cardiovascular Disease Prediction through Self-Improved Aquila Optimized Feature Selection in Quantum Neural Network & LSTM Model." Frontiers in Medicine 11. doi: 10.3389/fmed.2024.1414637.

Janiesch, Christian, Patrick Zschech, and Kai Heinrich. 2021. "Machine Learning and Deep Learning." Electronic Markets 31(3):685–95. doi: 10.1007/s12525-021-00475-2.

Khan, Moien Abdul Basith, Muhammad Jawad Hashim, Jeffrey Kwan King, Romona Devi Govender, Halla Mustafa, and Juma Al Kaabi. 2020. "Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends." Journal of Epidemiology and Global Health 10(1):107–11. doi: 10.2991/jegh.k.191028.001