Alfarama Journal of Basic & Applied Sciences

https://ajbas.journals.ekb.eg
ajbas@sci.psu.edu.eg

Faculty of Science Port Said University

http://sci.psu.edu.eg/en/

October 2025, Volume 6, Issue IV

DOI: 10.21608/ajbas.2025.391259.1259

ISSN 2682-275X

Submitted: 09/06/2025
Accepted: 16/07/2025

Pages: 412 - 430

DIABETES DETECTION VIA MACHINE LEARNING: TACKLING CHALLENGES AND ENVISIONING FUTURE INNOVATIONS

Tasneem N. Shindy[1,*], Mostafa Herajy[1], Wael A. Awad[2]

[1]Department of Mathematics and Computer Science, Faculty of Science, Port Said University, Port Said, Egypt
[2]Department of Computer Science, Faculty of Computers and Artificial Intelligence, Damietta University, Damietta, Egypt

***Corresponding author: tasneemshindy96@sci.psu.edu.eg**

ABSTRACT

Diabetes is one of the most serious diseases globally, affecting millions of individuals world- wide. Scientists are working to reduce the prevalence and incidence of this condition. Therefore, extensive research in this field has sought to pinpoint the most effective techniques for predicting diabetes. Examples of previously used approaches for predicting diabetes include data mining (DM), deep learning (DL), and machine learning (ML). Researchers employ these techniques to forecast diabetes at early stages and mitigate its impact. Many ML algorithms have been uti- lized, such as Support Vector Machine (SVM), ordering points to identify the clustering structure (Optics), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), XGBoost, and Logistic Regression (LR). While some studies confirm the effectiveness of these methods, recent findings underscore the superior effi- ciency of neural networks and deep learning. In this paper, we compare seven ML algorithms (including an enhanced deep learning model) using confusion matrix analysis and accuracy per- formance for accurate diabetes prediction using three different datasets. Our findings indicate that the enhanced deep learning model demonstrates high performance of 84\%, 93\% and 100\% on three datasets, PID, Taipei, and German, respectively, outperforming all other evaluated machine learning algorithms in this paper

Keywords: Diabetes Prediction, Machine Learning Algorithms, Neural Networks.

1. INTRODUCTION

Diabetes is a serious condition impacting around 537 million people globally, primarily in low- and middle-income countries, and is directly linked to 1.5 million deaths annually [1]. According to the IDF Diabetes Atlas 2023, an estimated 589 million adults aged 20–79 years, approximately one in nine adults worldwide, live with diabetes. By 2050, the number will reach 853 million, which is one in eight persons around the world [2]. Diabetes is a condition in which the human body cannot produce the necessary amount of insulin to regulate and monitor sugar levels [3]. Among other complications, this disease can result in heart disease, nerve injury, kidney disease, and damage to blood vessels. [4]. According to medical experts, diabetes is a disease that occurs when the blood glucose or blood sugar level in the human body is abnormally elevated [5]. In normal, glucose levels typically range between [70-99] mg per deciliter. When the blood glucose level rises above 126 milligrams per deciliter, it is diagnosed, while a level between 100 and 125 indicates pre-diabetes [6]. The number of people with diabetes around the world is at an all-time high, which shows how important it is to take action right away. Each year, diabetes leads to a significant number of deaths. The prevalence of diabetes is projected to increase annually. Early detection requires substantial machine learning support, as the condition is incurable and can lead to severe complications within our healthcare system [7]. Moreover, the objective of predicting diseases at an early stage has become critical. However, clinicians often encounter difficulties in achieving precise diagnostic outcomes when relying solely on symptom-based evaluations.

Machine learning is a standard and expanding methodology that employs recursive learning to offer powerful methods for classification and analysis [8]. Machine learning techniques enable researchers to train and test classification models. In the appropriate context of training and testing, machine learning has provided significant assistance in predicting diseases [9]. In recent years, data mining and machine learning have evolved into effective and supportive techniques in the medical field. Data mining is utilized in the pre-processing phase to analyze and extract information from healthcare data, while machine learning generates predictions from processed data using various methods [10]. Furthermore, Many ML algorithms have been utilized, such as SVM & Optics & RF & DT & KNN & GNB & LR & NN models [11]. However, there is no clear indication in the litrature of which of these approaches produces the best prediction accuracy.

Many machine learning models have been previously proposed to handle diabetes prediction. For example, Nesreen et al. [4] forecast an individual's likelihood of developing diabetes using artificial neural networks. The objective was to minimize the error function during the training of the neural network model. Throughout the training of the ANN model, the average error function of the neural network was 0.01, achieving an accuracy rate of 87.3\% in predicting a person's diabetes status. Abdulhadi et al. [12] conducted research and developed a tool that can assist medical professionals in identifying diabetes early and improving patient quality of life through the use of supervised learning techniques. The article discusses various modeling training methods, where the Random Forest algorithm was found to have the highest accuracy of 82\%. Jobeda et al. [32] utilized seven machine learning algorithms (DT, KNN, RF, NB, AB, LR, and SVM) to predict diabetes using the PIDD dataset. They evaluated the results using various indicators and found that all models achieved an accuracy level of over 70\%. Additionally, they implemented the NN model, which had the highest accuracy among all PIDD models at 88.6\%. A study by Bader et al. [13] improved the Artificial Back propagation Scaled Conjugate Gradient Neural Network (ABP-SCGNN) model. They used Mean Squared Error (MSE) and

accuracy as evaluation metrics. To train the ANN models, they employed varying numbers of neurons in the hidden layer, ranging from five to fifty. According to their experimental findings, the 20-neuron ABP-SCGNN model achieved 93\% accuracy. This study, conducted by Sivaranjani et al. [15] examines the use of (RF) and (SVM) in identifying specific characteristics. The researchers utilized Principle Component Analysis (PCA) to reduce the dimensionality of their analysis. The results showed that the prediction accuracy of RF was 83\%, while SVM had an accuracy of 81.4\%. Rady et al. [16] utilized eight algorithms on the PIDD dataset. The algorithms employed include logistic regression, support vector machines with linear and nonlinear kernels, random forest, decision tree, adaptive boosting classifier, K-nearest neighbors, and naïve Bayes. The highest performance, achieving 98\% accuracy, was attained by the random forest. Tasin et al. [17], used a range of techniques for diabetes prediction, including DT & SVM & RF & LR & KNN utilized eight algorithms on the PIDD dataset. The algorithms employed include logistic regression, support vector machines with linear and nonlinear kernels, random forest, decision tree, adaptive boosting classifier, K-nearest neighbors, and naïve Bayes. The highest performance, achieving 98\% accuracy, was attained by the random forest. Mamatha et al.[18] explored various techniques for classification and clustering, including Gaussian Naive Bayes, OPTICS, and BIRCH. They also utilized several performance metrics, with OPTICS being the most effective. Kangra et al. [6], used (SVM), (DT), (NB), (RF), (LR), and (KNN) machine learning algorithms on two datasets. The results demonstrated that SVM achieved an accuracy of 74\% for the first dataset, while KNN and RF outperformed with 98.7\% accuracy for the second dataset. Hassan et al. [20] analyzed the dataset using five machine-learning techniques LR & XGBoost & RF& CatBoost & NN. the CatBoost achieved accuracy at 73%.

Nevertheless, there are many restrictions and limitations in the previously conducted research, including incomplete or inaccurate data, generalizability, and unbalanced datasets. One factor that may impact the quality and dependability of the model is incomplete or inaccurate data, where important parameters, like insulin levels, are missing from some datasets; for instance, imputed or anticipated values are used, which adds more uncertainty to the projections. Moreover, the use of a single dataset or data from a particular location or community is a common barrier that limits the capacity to extrapolate results to larger or more diverse populations. Models developed using sparse or homogeneous data might not function effectively in different populations or actual clinical situations. Diabetes datasets frequently contain a disproportionate number of non-diabetic cases compared to diabetic instances. This results in models that are less successful at detecting actual positive cases of diabetes since they are biased toward the majority class.

In this paper, we examine and compare seven ML approaches utilized to predict diabetes in early stages. We employed three datasets: the first is the Pima Indians Diabetes Dataset (PIDD)[29]. The second dataset, the Taipei dataset[30], and the third dataset is a German dataset[31]

This paper is structured as follows: Section 2 outlines the methodology we apply in our research and also provides a brief overview of the machine learning algorithms employed in the predictions. Section 3 presents the main results of our paper by comparing the different approaches based on different comparison parameters and their accuracy. Section 4 presents the conclusion and culture work.

2. RESEARCH METHOD

In this section, we provide a brief overview of the steps taken in this paper to evaluate the seven machine learning techniques commonly used for diabetes prediction. We begin by introducing each technique and discussing its key features.

2.1. Logistic Regression (LR)

LR is powerful statistical tool has numerous applications in various domains. Despite its limitations, it remains one of the most prevalent and effective techniques for modeling relationships between variables [21]. By learning its theoretical foundations, types, and limitations, researchers can easily utilize LR to analyze complicating data sets. Hence implementing LR model to diabetes datasets with important details, researchers may forecast the probability of an individual being diabetic (binary outcome: 0 for non-diabetic, 1 for diabetic) depending on those features.[22].

2.2. K-Nearest Neighbors (KNN)

KNN is a simple and effective classification algorithm in ML. It classifies new data points based on the similar data points share the same class. In diabetes prediction, KNN classifies patients by analyzing the K-closest training samples and assigning the most frequent class label (diabetic or not). [23]. KNN's performance depends on choicing of K and the distance metric, and it can be computationally intensive for large datasets. Research show KNN achieves competitive accuracy in diabetes prediction. [24].

2.3. Optics algorithm

OPTICS is known as ordering points to identify the clustering structure's cluster analysis. It can serve as a standalone tool to gain insights into the distribution of a dataset. For instance, it can guide subsequent analysis and data processing or act as a pretreatment step for other algorithms that work on the identified clusters. It is one of the unsupervised machine learning algorithms. Furthermore, it addresses clustering problems by grouping points that are similar to each other, and to achieve this, it requires certain parameters [18]. The first parameter is the Core Distance, defined as the minimum radius required to classify a specific point as a core point. The second parameter is the reachability distance, representing the distance between point *o* and point *p*. Notably, the algorithm generates a comprehensive cluster hierarchy, necessitating that the user determines the final clustering, which incurs a runtime complexity of $O(n^2 log(n^2))$ [13].

2.4. XGBoost Algorithm

Data scientists can use a variety of techniques to create precise prediction models. In the past few years, XGBoost (Extreme Gradient Boosting) has become a very popular technique. The ensemble learning concept, which combines several models to enhance overall predictive performance, is the foundation of the machine learning algorithm XGBoost. It is especially well-suited for problems involving supervised learning, like regression and classification.

This is accomplished by a method called gradient boosting, which iteratively trains a series of weak learners (simple decision trees) in order to reduce the mistakes caused by the earlier models. XGBoost

can progressively raise the model's forecast accuracy by doing this. It is applicable to many different fields [20].

2.5. Support Vector Machines(SVM)

SVM is a developed prediction method. The basic principle of SVM is to find the hyperplane that maximizes the margin between different classes (classes here diabetic and non-diabetic) of input data (features) in dataset, where can effectively separate classes and detect complex patterns in the data. Although, challenges like feature selection, data imbalance and parameter tuning, still play an important rolein improving model quality. overall SVM is strong classification to be approach for diabetes early detection and treatment, contributing to improved patient outcomes[25].

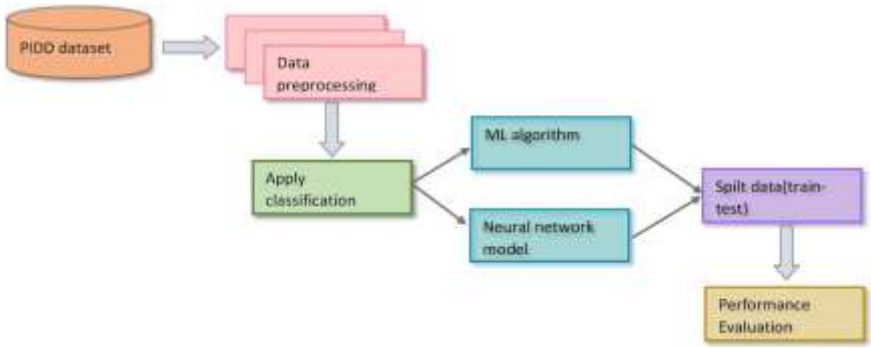

Figure 1: Proposed data-processing and Diabetes Prediction model using the seven ML models

2.6. Random Forest Algorithm (RF)

RF is one of the supervised ML algorithms. It consists of several decision trees and takes the average of their output to improve the accuracy, in other words, the greater the number of decision trees increases the more accuracy is achieved but it makes the algorithm slower. RF is used for both regression and classification. Furthermore, it solves the problem of overfitting [6]. Finally, generate votes for the prediction to select the most majority votes. Recently, many of the research in diabetes prediction adopts the random forest machine learning algorithms. For instance, V. Jackins et al. in [26] used random forest and Bayesian algorithms to discover cancer, coronary heart disease, as well as diabetes disease.

2.7. Neural Network and Deep Learning

Neural networks are artificial models that simulate the structure of the human brain [27]. They composed of many layers: an input layer(features), hidden layers(one or more), and an output layer(class). There is a weight assigned to each connection that changes as learning progresses. The architec- ture of these networks is organized in two primary stages. First, feed-forward neural networks ac- cept raw data through the input layer, process it via hidden layers with appropriate activation func- tions, and generate outputs based on computed activations and network weights. Second, the back- propagation algorithm is employed to iteratively adjust these weights, minimizing prediction errors over time. [28].

In the result section, we discuss the result of the prediction and determine the best method to predict diabetics with more efficiency. Figure 1 shows the proposed model from data pre-processing to Diabetes Prediction, as applied in our paper.

2.8. Dataset

For the purpose of this study, we have utilized three different datasets: the Pima Indians Diabetes Dataset (PIDD) [29], the Taipei dataset [30], and the German dataset [31] to better evaluate the seven machine learning approaches. The PIDD contains eight numeric-valued features and a total of 768 samples from both females and males. The eight features included are: pregnancy, age, BMI, blood pressure, insulin level, skin thickness, glucose, and outcome. The Taipei dataset is derived from a municipal medical center and consists of 15,000 women aged 20–80 years (collected 2018–2022). The dataset has eight different characteristics of the subjects, including the number of pregnancies, glucose level, diastolic blood pressure, sebum thickness, insulin level, age, BMI, and outcome. According to the German dataset, the structure and features closely match the well-known Pima Indians Diabetes Dataset, but the German dataset is larger and was collected at Frankfurt Hospital in Germany. The dataset is available on platforms such as Kaggle; it contains information on 2,000 patients and eight features: pregnancy, age, BMI, blood pressure, insulin level, skin thickness, glucose, and outcome.

2.9. Data Preprocessing and Experiment Design

In our study, we employ three diabetes datasets—PID, German, and Taipei—to investigate early detection strategies. We begin by pre-processing the data to enhance prediction accuracy through a series of steps that include filling missing values, feature selection, and data normalization. Missing values are replaced with the mean of the corresponding feature, and univariate feature selection is performed to identify the correlation between each feature and the output variable. As a result, the least correlated features are removed, leaving Age, Glucose, Preg, BMI, and Insulin as the most significant predictors, as shown in Fig. 2 for the PID, German, and Taipei datasets. Each dataset is split into two sets 67% for train set and 33% for test set.

2.10. Model Evaluations

Performance metrics play an important role for assessing how effectively ML models performed. They determine how accurate a model predicts outcomes or classifies data, and testing on unseen data helps assess its generalization capabilities. Commonly used evaluation metrics include Accuracy (AC) defined by Equation (1), Recall (R) by Equation (2), Precision (P) by Equation (3) , F1-score by Equation (4). Accuracy, one of the most widely used metrics, shows what proportion of the model's predictions were accurate. Meanwhile, the confusion matrix (CM) provides a detailed summary of the model's performance by comparing actual outcomes with predicted ones, allowing for an in-depth analysis of true_positives, true_negatives, false_positives, and false_negatives.

$$AC = \frac{True\ prediction}{Total\ prediction} = \frac{Tp + TN}{Tp + TN + Fp + FN} \quad (1)$$

$$R = \frac{Tp}{Tp + FN} \quad (2)$$

$$\mathrm{p} = \frac{Tp}{Tp + Fp} \quad (3)$$

$$f1 - score = 2 \times \frac{R \times p}{R + p} \quad (4)$$

Likewise, the Receiver operating curve (ROC) is a performance metric that illustrates the likelihood of a classification model's predictions being accurate.

3. RESULTS AND DISCUSSION

In this section, we provide a detailed evaluation of the performance of the seven machine learning techniques applied to the three datasets.

3.1. Result of ML Algorithms

To compare the performance of the seven ML models on the diabetes datasets, we evaluate each model using confusion matrices and ROC curves. Figures [3] through [10] present these results for all three datasets, showcasing the ROC curves and CM for each examined model. Moreover, the performance metrics: AC, R, P and f1-score, for each model applied to the three datasets (PIDD, Taipei, and German) are summarized in Table 1. These metrics were calculated using Equations (1–4). In the following, we discuss the performance of each model based on these outcomes.

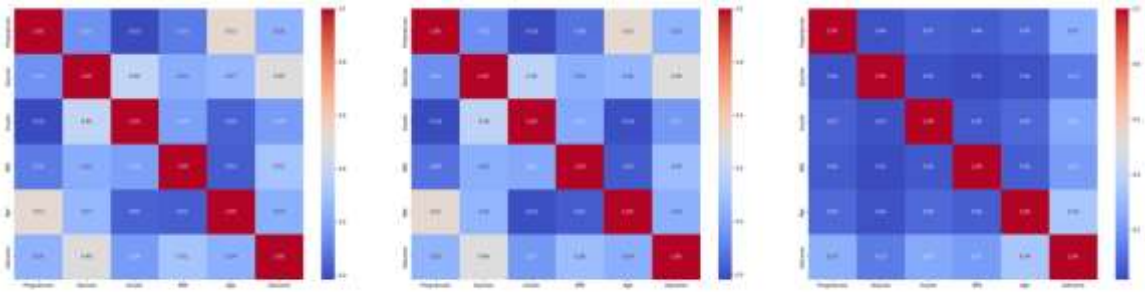

Figure 2: After pre-processing correlation between inputs and outputs for three datasets: (A) Correlation of PID dataset, (B) Correlation of German datase, (C) Correlation of Taipei dataset.

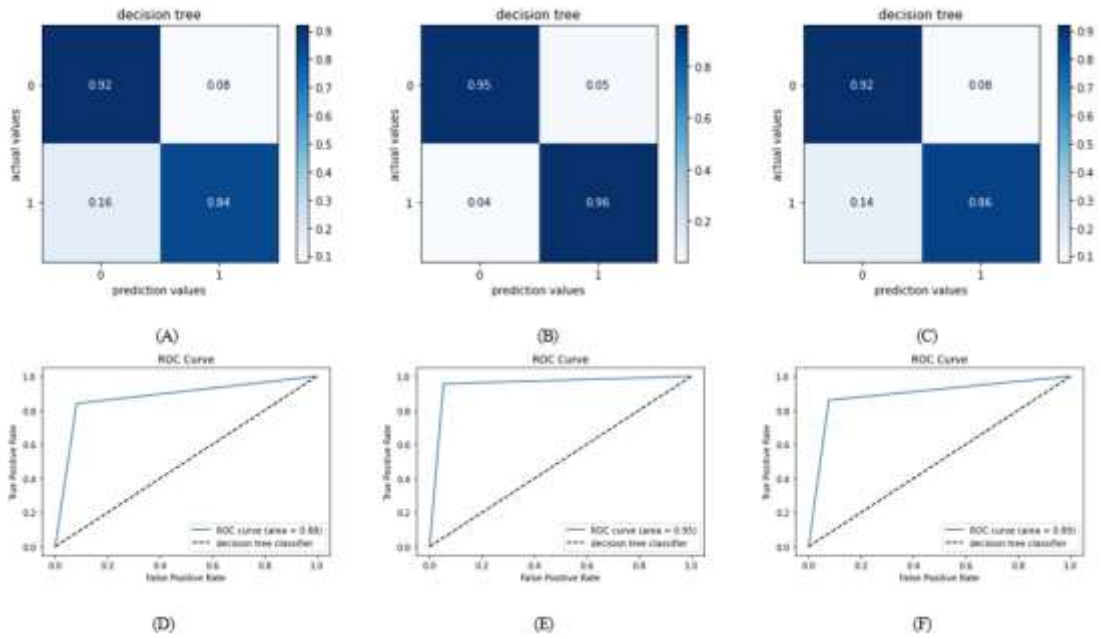

Figure 3: ROC curves and CM for the three datasets of the DT algorithm. The ROC curves, which plot the true positive rate against the false positive rate, are detailed as follows: Panel (A) shows the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Panel (D) presents the ROC curve for the PID dataset with an AUC value of 0.88, indicating that the DT curve is closest to the top-left corner. Similarly, Panel (E) illustrates the ROC curve for the German dataset with an AUC of 0.95, and Panel (F) shows the ROC curve for the Taipei dataset with an AUC of 0.89, both demonstrating that the DT algorithm consistently achieves optimal performance.

According to the PID dataset in Table [1], the RF algorithm demonstrates a high accuracy of 0.81, compared to other algorithms, and 0.81 in terms of Precision, Recall, and F1-score. Conversely, the KNN algorithm exhibits lower performance, with 0.77 in Precision, 0.74 in F1-score, and 0.74 in Accuracy compared to the other algorithms.

For the Taipei dataset in Table [1], the GB algorithm achieved impressive results, with a Precision of 0.94, a Recall of 0.93, an F1-score of 0.94, and an accuracy of 0.94, outperforming the other algorithms. The RF algorithm also performed well, reaching an accuracy of 0.93. In contrast, the GNB algorithm lagged behind, with a precision of 0.74, a Recall of 0.71, an F1-score of 0.72, and an Accuracy of 0.76. Moreover, for the German dataset in Table [1], the RF and GB algorithm recorded a high score of 0.99 across Recall, F1-score, an accuracy and Precision of 0.99, 0.98 for RF and GB respectively compared to other algorithms, while the LR algorithm shows lower performance with 0.74 in a precision, 0.78 in Accuracy, and 0.73 in both Recell and F1-score.

Based on these accuracy comparisons, GB and RF are the most effective models for predicting diabetes: on the German dataset, both models achieved 0.99 in an accuracy , while on the Taipei dataset, GB attained an accuracy of 0.94 compared to RF's 0.93. Additionally, on the PID dataset, RF outper- formed the other models with an accuracy of 0.81.

Based on the ROC curve analysis in Figures [3] through [10], in the PID dataset, GB outperformed the other models with the highest AUC value of 0.94, followed by RF with an AUC of 0.92. In the German

dataset, XGBoost achieved the highest AUC value of 0.97, outperforming the remaining models. Meanwhile, in the Taipei dataset, both XGBoost and GB led with an AUC of 0.93, followed closely by RF with an AUC of 0.92.

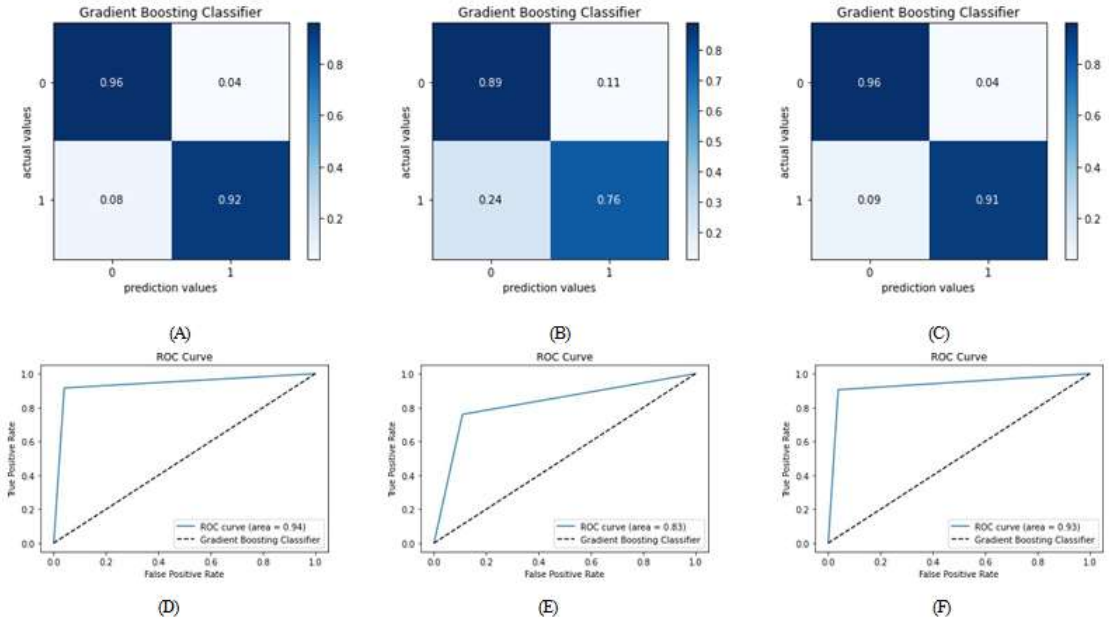

Figure 4: ROC curves and CM for the three datasets using the GB algorithm. The ROC curves, which plot the true positive rate against the false positive rate, are detailed as follows: Panel (A) shows the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Panel (D) presents the ROC curve for the PID dataset with an AUC of 0.94, indicating that the GB curve is closest to the top-left corner. Similarly, Panel (E) displays the ROC curve for the German dataset with an AUC of 0.83, and Panel (F) illustrates the ROC curve for the Taipei dataset with an AUC of 0.93, both reflecting optimal performance near the top-left corner.

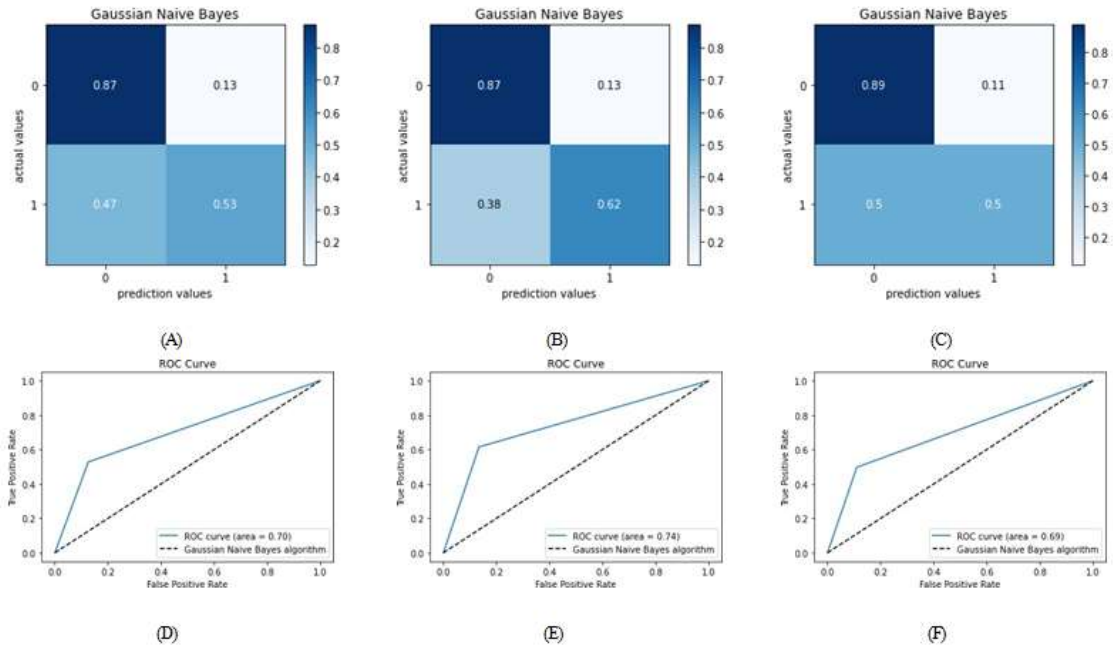

Figure 5: ROC curves and CM for the three datasets using the GB algorithm. The ROC curves, which plot the true positive rate against the false positive rate, are detailed as follows: Panel (A) shows the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Panel (D) presents the ROC curve for the PID dataset with an AUC of 0.94, indicating that the GB curve is closest to the top-left corner. Similarly, Panel (E) displays the ROC curve for the German dataset with an AUC of 0.83, and Panel (F) illustrates the ROC curve for the Taipei dataset with an AUC of 0.93, both reflecting optimal performance near the top-left corner.

3.2. Result of the Neural Network Model

After comparing various machine learning algorithms, we first implemented Jobeda's neural network model from [32].This model consists of Four-layers: first(input layer) with five features, second ,third (hidden layers) containing 26 and 5 neurons respectively, and fourth(output layer). Using learning rate of 0.01 with ReLU activation function on three datasets (PID, German, and Taipei), Jobeda's model achieved accuracies of 0.73, 0.88, and 0.92 on the respective datasets.

Subsequently, we developed an improved Four-layer model with a similar architecture, where the input layer contains five features, followed by two hidden layers (30,50) and a final output layer. In our design, the hidden layers utilize ReLU and Sigmoid activation functions. We experimented with different numbers of neurons in the hidden layers while maintaining a learning rate of 0.01. To optimize our model, we adjusted various hyperparameters such as the number of epochs and the number of neurons in the hidden layers evaluating performance across epoch values of 600, 800, and 1000 using the PID, German, and Taipei datasets. After extensive training, we determined that a configuration with two hidden layers containing 30 and 50 neurons respectively, a learning rate of 0.01, and 800 training epochs achieved the best results: accuracies of 0.84 on the PID dataset, 100 on the German dataset, and 0.93 on the Taipei dataset.

Figure 11 presents the ROC_curves and CM for the neural network model on the three datasets. Panel A shows the CM for the PID dataset, Panel B for the German dataset, and Panel C for the Taipei dataset. Panels D, E, and F display the ROC curves for the PID, German, and Taipei datasets, with AUC of 0.74, 0.94, and 0.91 respectively.

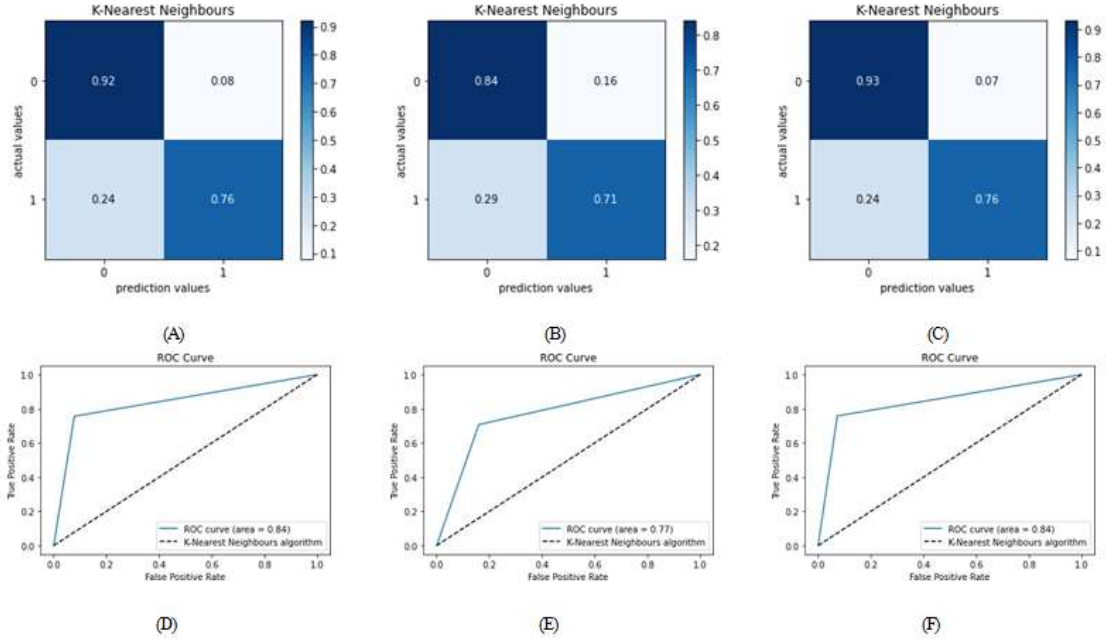

Figure 6: ROC curves and CM for three datasets of the KNN algorithm. The ROC curves plot the true positive rate against the false positive rate. Specifically, Panel (A) displays the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Additionally, Panel (D) displays the ROC curve for the PID dataset, with an AUC of 0.84, indicating that the curve is located near the top-left corner, which signifies strong performance. Similarly, Panel (E) illustrates the ROC curve for the German dataset, with an AUC of 0.77, while Panel (F) shows the ROC curve for the Taipei dataset, with an AUC of 0.84, the curve being closest to the top-left corner among the three

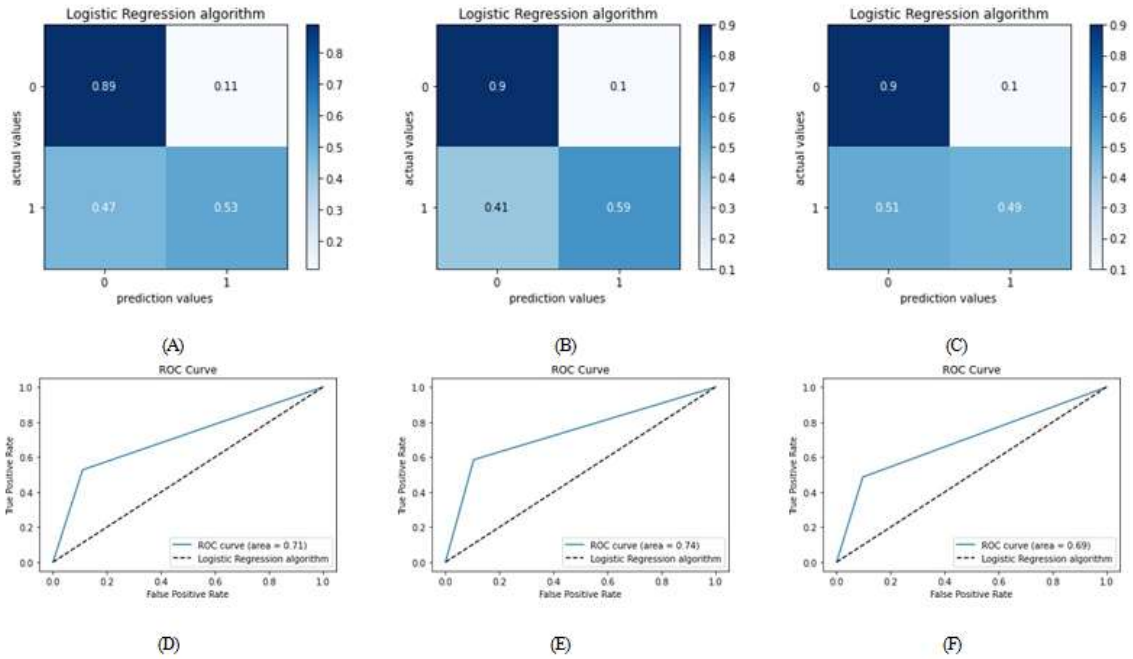

Figure7: ROC curves and CM for three datasets of the KNN algorithm. The ROC curves plot the true positive rate against the false positive rate. Specifically, Panel (A) displays the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Additionally, Panel (D) displays the ROC curve for the PID dataset, with an AUC of 0.84, indicating that the curve is located near the top-left corner, which signifies strong performance. Similarly, Panel (E) illustrates the ROC curve for the German dataset, with an AUC of 0.77, while Panel (F) shows the ROC curve for the Taipei dataset, with an AUC of 0.84, the curve being closest to the top-left corner among the three.

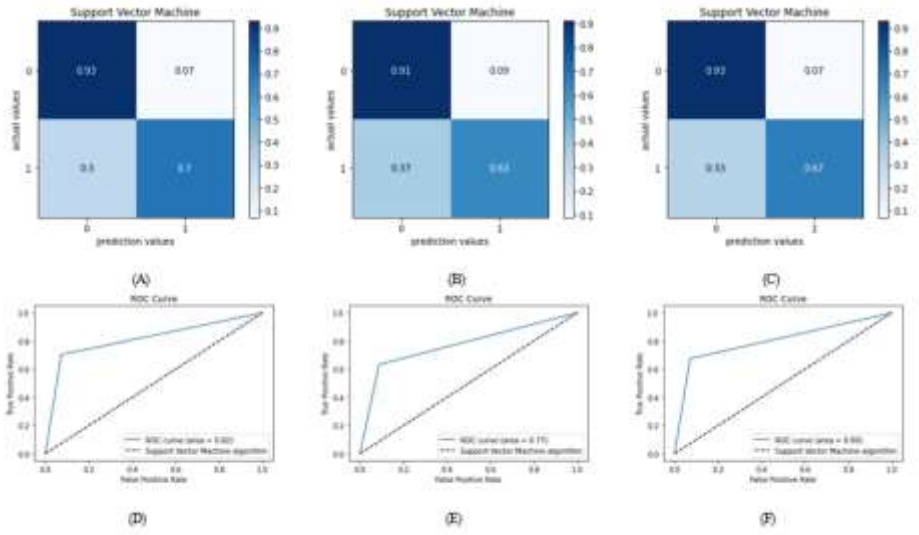

Figure 8: ROC curves and CM for three datasets of the SVM algorithm, with the ROC curves plotting the true positive rate against the false positive rate. Specifically, Panel (A) shows the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Panel (D) presents the ROC curve for the PID dataset with an AUC of 0.82, indicating that the SVM curve is closest to the top-left corner. Panel (E) displays the ROC curve for the German dataset with an AUC of 0.77, showing that the curve is near the top-left corner, while Panel (F) depicts the ROC curve for the Taipei dataset with an AUC of 0.80, again demonstrating optimal performance close to the top-left corner.

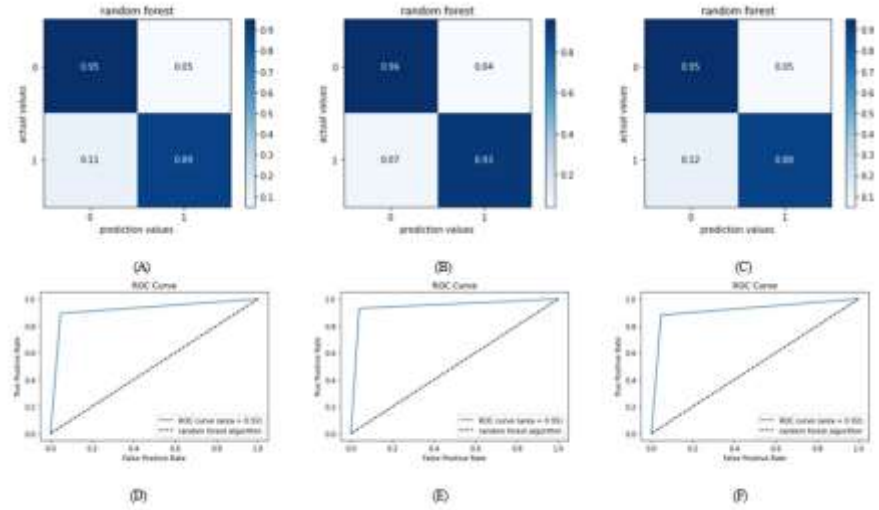

Figure 9: ROC curves and CM for three datasets of the RF model, where the ROC curves plot the true positive rate against the false positive rate. Specifically, Panel (A) displays the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Panel (D) presents the ROC curve for the PID dataset with an AUC value of 0.92, indicating that the RF curve is closest to the top-left corner. Similarly, Panel (E) shows the ROC curve for the German dataset with an AUC value of

0.95, and Panel (F) illustrates the ROC curve for the Taipei dataset with an AUC value of 0.92, both confirming optimal performance as the curves are nearest to the top-left corner.

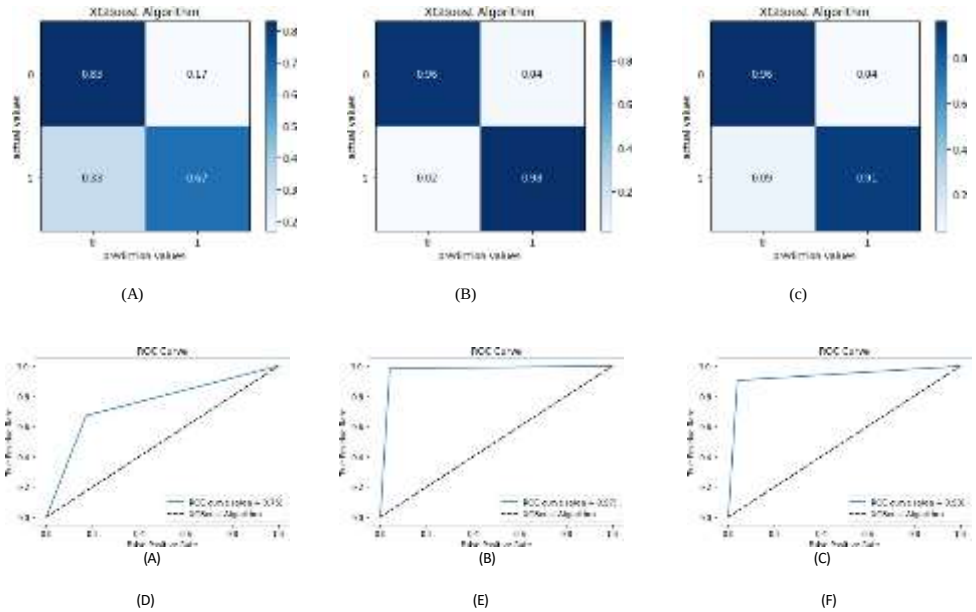

Figure 10: ROC curves and CM for the three datasets using the XGBoost algorithm. The ROC curves, which plot the true positive rate against the false positive rate, are detailed as follows: Panel (A) shows the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Panel (D) presents the ROC curve for the PID dataset with an AUC of 0.75, indicating that the GB curve is closest to the top-left corner. Similarly, Panel (E) displays the ROC curve for the German dataset with an AUC of 0.97, and Panel (F) illustrates the ROC curve for the Taipei dataset with an AUC of 0.93, both reflecting optimal performance near the top-left corner

Table [2] compares the performance of Baseline model, where Jobeda's model recorded accuracies of 73%, 88%, and 92% on the PID, German, and Taipei datasets respectively, Kumarmangal's model acheived 72%, 85%, and 91% on the PID, German, and Taipei datasets respectively, Alsulami's model have accuracies of 72%, 98%, and 92% on the PID, German, and Taipei datasets respectively, our enhanced deep learning model achieved significantly higher accuracies of 84%, 100%, and 93% on the corresponding datasets..

Based on the accuracy comparison as shown in Table [1] and [2], improved neural network model demonstrates high performance of 84%, 93% and 100% on three datasets (PID, Taipei, German), respectively. Furthermore, it's the most suitable model for predicting diabetes due to its superior performance compared to other algorithms.

4. CONCLUSION AND FUTURE WORK

Diabetes is a serious disease in the world, and it makes scientists and practitioners worldwide care about it. ML and DL techniques have been used recently to predict this disease in the early stages. In this paper, we utilized three datasets and performed pre-processing data, such as missing values and feature

selection. Additionally, we discussed and compared the performance of seven ML techniques using the well-established accuracy measures. Then we improve the deep learning by adjusting key hyperparameters. Our findings show that the deep learning model achieves accuracies of 84\%, 93\% and 100\% on three datasets (PID, Taipei, and German), respectively. Outperformed all the other algorithms on the three datasets. In our work, we encountered challenges of the small dataset as well as PID dataset not being updated since 2014. We hope in the future to tackle these challenges. Future work will involve refining the deep learning approach by considering different architectures for different techniques like RNN, CNN, and LSTM. Additionally, we aim to utilize ANOVA and p-value techniques to compare model performance and hyperparameter optimization to further improve predictive performance.

Table 1: Performance metrics of different machine learning algorithm of the PIDD,Taipei and Germany datasetes of the seven machine learning algorithms.

	PID Dataset				Taipei dataset				Germany Dataset			
Algorithm	Precision	Recell	F1-score	Accuracy	Precision	Recell	F1-score	Accuracy	Precision	Recell	F1-score	Accuracy
DT	77	75	78	78	91	90	90	91	98	98	98	98
GNB	77	75	76	78	74	71	72	76	75	75	75	79
LR	78	79	79	79	75	71	72	77	74	73	73	78
KNN	77	73	74	77	86	84	85	87	85	86	86	87
SVM	80	76	77	80	84	82	83	85	81	80	81	83
GBC	78	78	78	80	94	93	94	94	98	99	99	99
RF	81	81	81	81	93	92	92	93	99	99	99	99
XGBoost	75	75	77	77	94	93	94	94	96	97	97	96

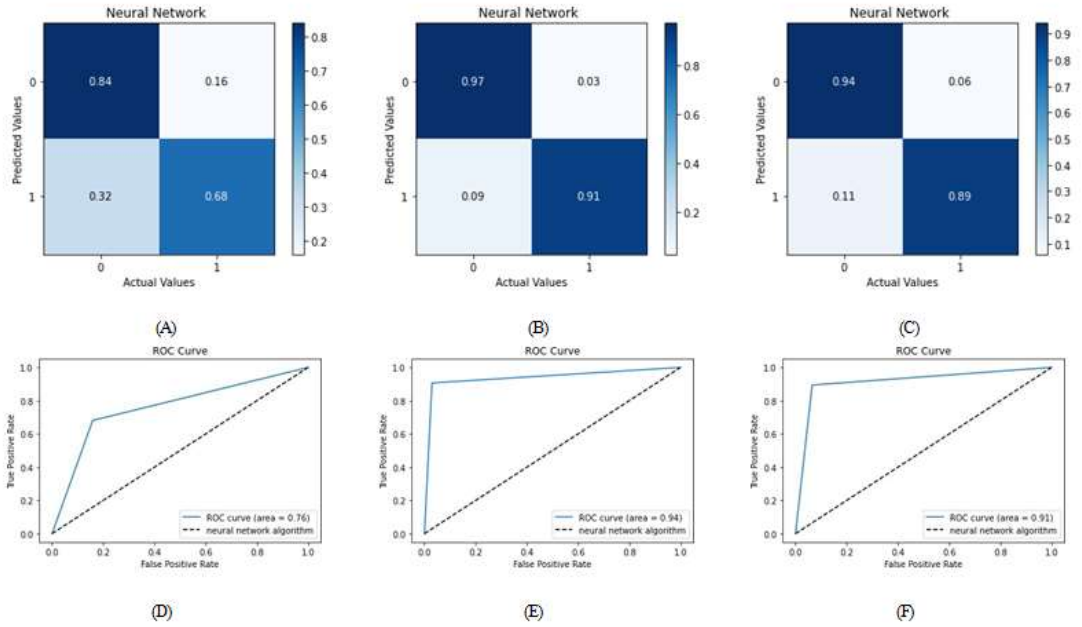

Figure 11: ROC curves and CM for three datasets of the NN model. The ROC curves, which plot the true positive rate against the false positive rate, are detailed as follows: Panel (A) displays the confusion matrix for the PID dataset, Panel (B) for the German dataset, and Panel (C) for the Taipei dataset. Panel (D) presents the ROC curve for the PID dataset with an AUC value of 0.76, indicating that the NN curve is near the top-left corner. Panel (E) shows the ROC curve for the German dataset with an AUC of 0.94,

meaning that the NN curve is closest to the top-left corner. Finally, Panel (F) depicts the ROC curve for the Taipei dataset with an AUC of 0.91, also demonstrating that the NN curve is closest to the top-left corner.

Table 2: Performance analysis of Baseline model on three datasets

Model	hidden layer	PID dataset	Taipei dataset	German dataset
Jobeda's model	(26,5)	73	92	88
Kumarmangal's model	(16,4)	72	91	85
Alsulami's model	(128,265,265)	72	92	98
Our Model	(30,50)	84	93	100

5. REFERENCES

[1] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:8869–8879, 2017.

[2] International.Diabetes.Federation.(IDF).Diabetes.Atlas,.11th.Edition,2023, https://diabetesatlas.org/data-by-location/global/..URL.https://diabetesatlas.org/data-by-location/global/

[3] Dharmarathne Gangani, N. Jayasinghe Thilini, Bogahawaththa Madhusha, D.P.P. Meddage, and Rathnayake Upaka. A novel machine learning approach for diagnosing diabetes with a self-explainable interface. *Healthcare Analytics*, 5:100301, 2024. ISSN 2772-4425.

[4] Samer El_Jerjawi Nesreen and S. Abu-Naser Samy. Diabetes prediction using artifi- cial neural network. *International Journal of Advanced Science and Technology*, 121:54– 64, 2018.

[5] Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin, Ying Ju, and Hua Tang. Predicting di- abetes mellitus with machine learning techniques. *Frontiers in Genetics*, 9, 2018. ISSN 1664-8021.

[6] Kirti Kangra and Jaswinder Singh. Comparative analysis of predictive machine learning algo-rithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, 12:1728–1737, 06 2023.

[7] Sisodia Deepti and Singh Sisodia Dilip. Prediction of diabetes using classification algo- rithms. *Procedia Computer Science*, 132:1578–1585, 2018. ISSN 1877-0509.

[8] A. S. Hovan George, Shahul Aakifa, A. Shaji George, T. Baskar, and A. Shahul Hameed. A survey study on big data analytics to predict diabetes diseases using supervised classification methods. *Partners Universal International Innovation Journal*, 1(1):1–8, 2023.

[9] Divya Kaur Bhullar, Natassha Shievanie Selvaraj, Fung Teng Choong, Chen Wan Jing, Kang Xiaoxi, Dini Handayani, Norhidayah Hamzah, Muharman Lubis, and Teddy Mantoro. De-

veloping a predictive supervised machine learning models for diabetes. In *2021 IEEE 7th International Conference on Computing, Engineering and Design (ICCED)*, pages 1–6, 2021.

[10] Chaitanya Suryadevara. Diabetes risk assessment using machine learning: A comparative study of classification algorithms. *International Journal of Applied Engineering Research and Development*, 8:1–10, 02 2022.

[11] Roshi Saxena, Sanjay Kumar Sharma, Manali Gupta, and G C Sampada. A comprehensive review of various diabetic prediction models: A literature survey. *Journal of Healthcare Engi- neering*, 2022:15, 2022.

[12] Nour Abdulhadi and Amjed Al-Mousa. Diabetes detection using machine learning classification methods. In *2021 International Conference on Information Technology (ICIT)*, pages 350–354, 2021.

[13] Muhammad Mazhar Bukhari, Bader Fahad Alkhamees, Saddam Hussain, Abdu Gumaei, Adel Assiri, and Syed Sajid Ullah. An improved artificial neural network model for effective diabetes prediction. *Complexity*, page 10, 2021. ISSN 1076-2787.

[14] C Kranthirekha, S Pothalaiah, Sharmila Vallem, Veerlapati Siri, U Madhuri, Sunkireddy Akshitha, and Kranthi Chennaboina. Diabetics prediction using machine learning techniques. *Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edi- tion)*, Vol: 42:485–494, 11 2023.

[15] S Sivaranjani, S Ananya, J Aravinth, and R Karthika. Diabetes prediction using machine learn- ing algorithms with feature selection and dimensionality reduction. In *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 141–146, 2021.

[16] Mohamed Rady, Kareem Moussa, Mahmoud Mostafa, Abdelrahman Elbasry, Zeyad Ezzat, and Walaa Medhat. Diabetes prediction using machine learning: A comparative study. In *2021 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 279–282, 2021.

[17] Isfafuzzaman Tasin, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. Diabetes prediction using machine learning and explainable ai techniques. *Healthcare Technology Letters*, 10(1-2): 1–10, 2023.

[18] Mamatha Bai B G, B. Nalini, and Jharna Majumdar. Analysis and detection of diabetes using data mining techniques—a big data application in health care. *AISC*, 882:443–455, 2019. ISSN 2194-

5357.

[19] S. Kumar and M. Venkatesulu. Gramian matrix data collection-based random forest classifica- tion for predictive analytics with big data. *Soft Computing*, 23:8621–8631, 2019. ISSN 1433- 7479.

[20] Hassan Shojaee-Mend, Farnia Velayati, Batool Tayefi, and Ebrahim Babaee. Prediction of di- abetes using data mining and machine learning algorithms: A cross-sectional study. *Healthc Inform Res*, 30(1):73–82, 2024.

[21] Suja A. Alex, J. Jesu Vedha Nayahi, H. Shine, and Vaisshalli Gopirekha. Deep convolutional neural network for diabetes mellitus prediction. *Neural Computing and Applications*, 34, 2022. ISSN 1433-3058.

[22] M Malini, B Gopalakrishnan, and S Naveena. Prediction and detection of diabetics melli- tus using different machine learning approaches. In *2022 International Conference on Ap- plied Artificial Intelligence and Computing (ICAAIC)*, pages 323–330, 2022.

[23] Tarannom Parhizkar. K-nearest neighbors (knn) algorithm for energy prediction models. *Jour- nal of Cleaner Production*, 07 2021.

[24] Rastogi Rashi and Bansal Mamta. Diabetes prediction model using data mining techniques. *Measurement: Sensors*, 25:100605, 2023. ISSN 2665-9174.

[25] Mostafa Safdari Shadloo. Application of support vector machines for accurate prediction of convection heat transfer coefficient of nanofluids through circular pipes. *International Journal of Numerical Methods for Heat & Fluid Flow*, 31(8):2660–2679, 2021.

[26] Chun-Yang Chou, Ding-Yang Hsu, and Chun-Hung Chou. Predicting the onset of diabetes with machine learning methods. *Journal of Personalized Medicine*, 13(3), 2023. ISSN 2075-4426.

[27] Md Alamin Talukder, Md Manowarul Islam, Md Ashraf Uddin, Mohsin Kazi, Majdi Khalid, Arnisha Akhter, and Mohammad Ali Moni. Towards reliable diabetes prediction: Innovations in data engineering and machine learning applications. *Digital health*, 2024.

[28] Kenas Farid, Saadia Nadia, Ababou Amina, Ababou Noureddine, Zabat Mahdi, and BenSiSaid Karim. Neural network-based estimation of lower limb joint kinematics: A minimally intrusive approach for gait analysis. *Medicine in Novel Technology and Devices*, 23:100318, 2024. ISSN 2590-0935.

[29] Pim India Diabetes. Pim India Diabetes Dataset, https://www.kaggle.com/datasets/mathchi/diabetes-

data-set. URL https://www.kaggle.com/datasets/mathchi/diabetes-data-set.

[30] Taipei Diabetes. Taipei Diabetes Dataset, https://drive.google.com/file/d/1eAplOYO-k7ZYHj4uHAY1tEr8VTeaxS6u/view.

[31] German diabetes. German diabetes dataset , https://www.kaggle.com/datasets/johndasilva/diabetes. URL https://www.kaggle.com/datasets/johndasilva/diabetes.

[32] Jamal Khanam Jobeda and Y. Foo Simon. A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4):432–439, 2021. ISSN 2405-9595.