Early Detection of Diabetes Using Multi-Strategy Machine Learning Models: A Smart Healthcare Approach for Enhanced Diagnostic Accuracy

[1]Nayeem Ahmed, [1]Syed Imtiyaz Hassan, [2]Md Zair Hussain

[1]Department of Computer Science & Information Technology, Maulana Azad National Urdu University, Gachibowli Hyderabad, INDIA
[2]Maulana Azad National Urdu University, Gachibowli Hyderabad, INDIA

Abstract
Undiagnosed diabetes is very dangerous, and a growing public health issue around the world. We need to be able to identify early if we want to halt the course of the disease and mitigate risks to critical organs. This project will build a multi-faceted ML model that will increase the accuracy and efficiency of diabetes diagnosis. They use a dataset that has health indicators such as age, BMI, insulin and glucose to train a number of machine learning algorithms to classify the risk of diabetes. DenseNet 96.8% was the best algorithm of all algorithms and AlexNet 94.8%. Other good performers included Weighted KNN and Fine Gaussian SVM with 83.4% and 83.5% accuracy respectively. These models were evaluated further on performance metrics such as F1 score, accuracy and recall. For example, because of its ability to detect positive cases, DenseNet recall reached 0.97. This proposed approach combines ensembles and sophisticated feature selection for diagnosing purposes. The model's clinical fit is guaranteed by statistical calculations used to confirm its fit in the real world. Finally, by offering a robust machine learning-based early diabetes detection solution, this study makes advances in medical diagnostics and aids clinicians to treat patients more quickly and effectively.

Keywords: Diabetes diagnosis, machine learning, predictive analytics, feature selection, decision tree, early detection, healthcare system, ensemble techniques, diagnostic accuracy, smart healthcare

1. Introduction

Diabetes is a disease that causes high blood glucose. Diabetes over time is also very harmful to the heart, blood vessels, kidneys, eyes, nerves, and heart. There are millions of diabetics in the world today, and many of them are undiagnosed until an issue occurs, according to recent studies [1,2]. Diabetic management and avoiding complications depends a lot on detection and early intervention. But many don't realize they have it until their symptoms worsen and leads to morbidity and mortality [3]. So, the mortality rates of diabetes could be significantly reduced if prediction models are created that are robust enough to predict diabetes in time. [3,4]
Several advances in the recent past have been made towards making illness prediction models more accurate using machine learning (ML) and healthcare data. Large data sets could also be mined by machine learning algorithms, and can find complex patterns and provide early diagnostic information [5-8]. The aim of this thesis is to use the Pima Indians Diabetes Database (PIDD), one of the most famous datasets in the medical research community, to predict diabetes's onset with high-powered machine learning models. This dataset from the National Institute of Diabetes and Digestive and Kidney Diseases is loaded with diagnostic variables, such as BMI, insulin and glucose, all of which are indicators of diabetes. [10,9]
The main purpose of this work is to create and apply predictive models for diabetes diagnosis and determine the best machine learning technique for this. Many machine learning models such as Support Vector Machines (SVM), Neural Networks, Decision Trees, Ensembles, etc., were proved promising for diagnosis by an exhaustive review of the literature [11–13]. When we fit these models to the PIDD data, we will compare their performance on many different indicators such as accuracy, precision, recall, AUC and F1-score. Optimising the best performing model with ensemble methods to make the prediction even better.
It also aims to address some of the problems associated with prediction of diabetes such as unbalanced data and making models generalisable through feature selection and cross-validation. Feature scaling and stratified cross-validation guarantee that the built models are robust and can make precise predictions. The outcome of this study could revolutionise early diabetes diagnosis to save patients and enable doctors to make better data-driven treatment decisions in the field of diabetes.
It is formatted as follows: approach, with a review of proposed models and methods, followed by a section for the dataset and preprocessing steps. Whereas the literature reviews how important the results are and where they might lead for research, the results section summarizes the results and evaluation of several models. The work's contributions and applications to predictive analytics and healthcare are rounded off at the end.

2. Literature Survey and Related Works

There are too many people in the world with diabetes; 537 million people worldwide are diabetic today and it is estimated that there will be 783 million in 2045. As effective as they are, other early diagnosis and prediction approaches in diabetes, such as glucose tolerance testing and clinical evaluation, are often poor [14–16]. These methods often entail tedious and subjective manual review of clinical data. Then again machine learning (ML) offers a few advantages in terms of using big datasets and sophisticated algorithms to identify the patterns and

correlations in the data. For instance, a new study shows that machine learning models can be accurate up to 90 per cent, an increase over the classical approach.

Machine learning methods for diabetes prediction boost diagnostic sensitivity and response speed so doctors can act earlier, rather than later. Since diabetes comorbidities, such as cardiovascular disease, kidney disease, and blindness, can be more common, costing patients substantial amounts of healthcare and decreasing patients' quality of life, this shift to data-driven strategies is key. Besides, machine learning systems can also handle an abundance of demographics and risk factors to make custom predictions. [17–20] Many experiments dealing with different ML algorithms to predict diabetes have been undertaken on these trends. These are a few related publications that contribute to this literature:

In a private data set derived from 203 female patients in Bangladesh, and the Pima Indian Diabetes data set, Isfafuzzaman Tasin et al. [21] developed a diabetes algorithm. They leveraged mutual information to apply feature selection and SMOTE and ADASYN to correct for class inequity. XGBoost with ADASYN classifier with 81% accuracy, F1 of 0.81 and AUC of 0.84 was the most accurate against other models. Model interpretation was made through explicable AI (LIME, SHAP), which provided the reasoning behind predictions. The paper shows how important it is to control the data imbalance and make smart features selections.

From data from a Taipei Municipal hospital, Chun-Yang Chou et al. [22] trained a diabetes prediction model on 15,000 women aged 20 to 80. They studied two-class logistic regression, neural networks, decision forests and enhanced decision trees with Microsoft Machine Learning Studio, on traits such as plasma glucose, insulin, BMI, and age. Two class boosted decision tree was the most effective and predicted fairly well with AUC 0.991. The article highlights challenges to keep various types of data features up to date and the promise of machine learning for early diabetes detection.

Salliah Shafi Bhat et al used the Pima Indian Diabetes Dataset (PIDD). [23] for Diabetes Mellitus Risk Assessment and Prediction Framework Using Machine Learning Algorithms to predict diabetes from various ML models, like Decision Tree, Gradient Boost, and Logistic Regression. 91 % accuracy, 96 % precision, 92 % recall, and a 94% F1 score the Decision Tree model was quite successful as it was all about improving prediction accuracy using Feature Engineering. It also mentions performance requirements for feature selection and data distribution, and diabetes predictive systems.

Sandip Kumar Singh Modak and Vijay Kumar Jha's "Diabetes Prediction Model Using Machine Learning Techniques" [24] attempted to better diagnosis of diabetes through different machine learning models. It used advanced ensembles such as XGBoost, LightGBM and CatBoost along with algorithms such as Logistic Regression, SVM, Nave Bayes and Random Forest. The authors showed prediction accuracy with Kaggle data and python implementation. CatBoost reached the highest accuracy of 95.4% and AUC-ROC value of 0.99 proving that it was accurate at detecting early diabetes.

What Fayroza Alaa Khaleel and Abbas M. Al-Bakry are seeking in "Diagnosis of Diabetes Using Machine Learning Algorithms" [25] is to make diabetes diagnosis earlier using machine learning. Its primary purpose is to predict the onset of diabetes using the diagnostic data in the Pima Indian Diabetes Dataset (PIDD). The three ML techniques were implemented and compared by the authors: K-nearest Neighbour (KNN), Nave Bayes (NB) and Logistic Regression (LR). The results were that with 94% prediction accuracy, Logistic Regression beat Nave Bayes (79%) and KNN (69%). That means that LR is useful in diagnosing diabetes accurately.

Table 1 Literature Survey Summary on Diabetes prediction using Machine Learning

Author(s)	Paper Description	Merits/Numerical Results	Challenges
Isfafuzzaman Tasin et al. (2022) [21]	Developed an automatic diabetes prediction system using Pima dataset and additional samples with various ML models.	Achieved 81% accuracy with XGBoost using the ADASYN approach; provided website and Android app for real-time predictions.	Managing class imbalance using SMOTE and ADASYN; ensuring robustness across diverse data samples.
Chun-Yang Chou et al. (2023) [22]	Analyzed the prediction of diabetes in Taiwan using neural networks and other ML models with patient data from medical centers.	Best results with two-class boosted decision tree achieving AUC of 0.991, indicating high predictive performance.	Adapting the model to various patient demographics; data preprocessing and handling class imbalance.
Salliah Shafi Bhat et al. (2023) [23]	Developed a risk assessment framework for diabetes using PIDD and ML algorithms like Decision Tree, Logistic Regression.	Decision Tree achieved the highest accuracy of 91%, precision of 96%, recall of 92%, and F1 score of 94%.	Addressing overfitting and optimizing feature selection for better model performance.
Sandip Kumar Singh Modak et al. (2023) [24]	Proposed a diabetes prediction model using various ML and ensemble learning techniques with real-world data from Kaggle.	CatBoost achieved 95.4% accuracy and AUC-ROC of 0.99, outperforming other ensemble methods like XGBoost (94.3%).	Computational complexity of training ensemble models; ensuring scalability for real-world applications.
Fayroza Alaa Khaleel & Abbas M.	Developed a model using LR, NB, and KNN to predict diabetes onset using PIDD.	Logistic Regression achieved 94% accuracy, making it the most effective among the tested models.	Balancing model precision with recall; optimizing model performance for different types of diabetes.

Al-Bakry (2021) [25]			

Table 1: A quick summary of the key conclusions from several ML studies on diabetes prediction — the methodologies, the results, and the challenges to implement them. There are many issues in the studies of ML for Diabetes Prediction: class imbalance in data sets, feature selection and data preparation, models that are hard to read. Also, a lot of models are not very generalisable across a wide range of patient groups. The work we propose here would try to solve such problems by using advanced data preprocessing to overcome class imbalance, feature engineering to optimize the model performance and machine learning/ensemble combinations to ensure accuracy and computational efficiency. Our long-term aim is to design a generic, robust diabetes prediction system that can be easily incorporated into medical practices and enhance patient care.

3. Methodology: Multi-strategy Machine Learning model for diabetes predcition

A set of important procedures is included in the prediction of diabetes onset using Pima Indians Diabetes Database to give a robust analysis. Data preprocessing, for cleaning up the data to analyse it, is done first to remove missing values. Then comes the data integration, transformation, reduction and feature scaling. This is followed by partitioning of data, especially stratified fivefold cross-validation to retain the distribution of the outcome variable across the training and test datasets. The training data is then fed to different models of machine learning like neural networks, decision trees and support vector machines. Performance of such models is based on parameters such as F1-score, accuracy, precision and recall. They evaluate algorithms that are the best. An ensemble of weighted score is then implemented for better prediction, and optimal weights are calculated for each model for each cross-validation fold. Then the test set is exposed to the trained ML/Ensembl models and result evaluation can be done with accuracy. With a robust early diabetes detection system in place, this discipline can lead to improved outcomes in healthcare long-term. The general procedure is plotted visually in Figure 1.

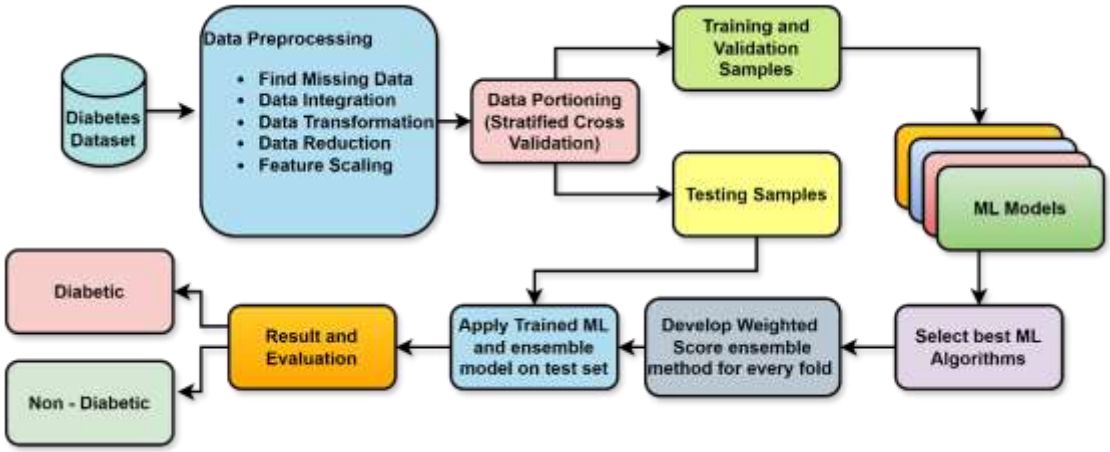

Figure 1 Overview of the Proposed Diabetes Prediction System Model

3.2 About Dataset

A famous diabetes development estimate dataset is the Pima Indians Diabetes Database. It contains several medical risk factors that determine a patient's likelihood of developing diabetes. Especially data from female patients aged at least 21 and of Pima Indian origin. Its major aim is to assign patients to a category for diabetes, or lack of diabetes, based on these criteria. The data – its properties are represented in Table 2 – can be used to develop machine learning models to target early diagnosis and management of diabetes, thus improving public health by the way.

Table 2 Summary of Attributes in the Pima Indians Diabetes Dataset.

Attribute	Description	Type
Pregnancies	Number of pregnancies	Numeric
Glucose	Plasma glucose concentration (mg/dL)	Numeric
Blood Pressure	Diastolic blood pressure (mm Hg)	Numeric
Skin Thickness	Triceps skin fold thickness (mm)	Numeric
Insulin	2-Hour serum insulin (mu U/ml)	Numeric
BMI	Body mass index (kg/m²)	Numeric
Diabetes Pedigree Function	Family history likelihood of diabetes	Numeric
Age	Age of the patient (years)	Numeric
Outcome	Presence (1) or absence (0) of diabetes	Categorical

Step 1: Data Preprocessing

The Pima Indians Diabetes Database Data preparation phase starts with some statistical steps that ensure that the dataset is properly cleaned, combined, transformed, compressed and scaled for data analysis and modeling. This process can be numerically represented with the following equations.

1. Finding Missing Values: Let $D = \{d_1, d_2, \ldots, d_n\}$represent the dataset containing (n) instances. If any instance (d_i) has missing values denoted as $m_j \in d_i$, we can define a function $f: D \to D'$, that imputes these missing entries. One common method for imputation is using the mean of the respective feature (X_j):

$$f(d_i) = \begin{cases} \mu_j & \text{if } m_j \text{ is missing} \\ m_j & \text{otherwise} \end{cases}$$

where $\mu_j = \frac{1}{n_j}\sum_{k=1}^{n_j} X_{kj}$and ($n_j$) is the number of non-missing entries in feature (j).

2. Data Integration: If we have multiple datasets $(D_1, D_2, \ldots, D_k)$, we integrate these datasets into a single dataset $D_{integrated}$as follows:

$$D_{integrated} = \bigcup_{i=1}^{k} D_i$$

Additionally, if we consider merging on a common attribute (A), we can denote the integration process as:

$$D_{integrated} = \bigcup_{i=1}^{k} (D_i \bowtie A)$$

3. Data Transformation: To transform the integrated data into a suitable format, we apply a transformation function $T: D_{integrated} \to D_{transformed}$. For instance, normalizing the features (X_j) in (D) to the range [0, 1] can be defined mathematically as:

$$X'_{ij} = \frac{X_{ij} - X_{min,j}}{X_{max,j} - X_{min,j}}$$

where X'_{ij} is the normalized value, and $X_{min,j}$and $X_{max,j}$are the minimum and maximum values of feature (j) in the dataset.

4. Data Reduction: To reduce the dimensionality of the dataset, we can apply Principal Component Analysis (PCA). If (X) is the original dataset represented as $X \in \mathbb{R}^{m \times n}$(where (m) is the number of instances and (n) is the number of features), the PCA transformation is given by:

$$Y = XV_k$$

where (V_k) is the matrix of the top (k) eigenvectors obtained from the covariance matrix $\Sigma = \frac{1}{m-1}X^T X$

5. Feature Scaling: Finally, we scale the features in (D) to ensure they have similar ranges. One common scaling method is standardization, which transforms the data into a distribution with a mean of 0 and a standard deviation of 1:

$$Z_{ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

where $\mu_j = \frac{1}{m}\sum_{i=1}^{m} X_{ij}$is the mean of feature (j), and ($\sigma_j = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_{ij} - \mu_j)^2}$is the standard deviation of feature (j).Combining all these mathematical representations, the overall preprocessing step can be summarized as follows:

$$D_{preprocessed} = T(f(D_{integrated})) \cup Y \cup Z$$

Where $D_{preprocessed}$ is the data after all the preprocessing (imputation of missing values, integration, transformation, reduction, and scaling). This preprocessing step is important for getting the Pima Indians Diabetes Dataset ready for Machine Learning analysis and training of models.

Step 2: Data Partitioning (Stratified Five-Fold Cross-Validation)

After preprocessing, Pima Indians Diabetes Database need to split up the data into training and testing sets with the goal variable distribution (Diabetes outcome) kept at all folds. These are steps that we can quantify in terms of numbers the stratified 5x cross-validation process.

1. Dataset Definition: Let $D_{preprocessed} = \{d_1, d_2, \ldots, d_m\}$represent the dataset after preprocessing, where (m) is the total number of instances. The target variable $Y \in \{0,1\}$indicates the presence (1) or absence (0) of diabetes.

2. Stratification: We first need to determine the distribution of the target variable in the dataset. Let (n_0) be the number of instances where (Y = 0) (no diabetes) and (n_1) be the number of instances where (Y = 1) (diabetes). Therefore:

$$[n_0 + n_1 = m]$$

The proportions can be calculated as:

$$[p_0 = \frac{n_0}{m}, \quad p_1 = \frac{n_1}{m}]$$

where (p_0) and (p_1) are the proportions of each class in the dataset.

3. Fold Creation: The dataset ($D_{preprocessed}$) is then divided into (k = 5) folds, ensuring that each fold maintains the same proportion of class labels as the original dataset. Each fold can be represented as:

$$[F_i = \{f_{i,1}, f_{i,2}, \ldots, f_{i,n_i}\} \quad \text{for } i = 1,2,\ldots,5]$$

where (n_i) is the number of instances in fold (F_i).

To ensure stratification, the number of instances in each fold can be calculated as:

$$[n_{0,i} = \text{floor}(p_0 \cdot n_i) \quad \text{and} \quad n_{1,i} = n_i - n_{0,i}]$$

where $n_{0,i}$and $n_{1,i}$represent the number of instances in fold (F_i) for classes 0 and 1, respectively.

4. Cross-Validation Procedure: The cross-validation process consists of training the model on (k-1) folds and validating it on the remaining fold. This can be mathematically expressed as:

$$[\text{Train}_i = D_{\text{preprocessed}} \setminus F_i \quad \text{and} \quad \text{Test}_i = F_i]$$

where Train_i is the training set composed of the data from all folds except (F_i), and Test_i is the test set composed of the data in fold (F_i).

Step 3: Model Application and Ensemble Development

This stage involves applying several machine learning (ML) models to the Pima Indians Diabetes Database, identifying the top-performing models, and creating an ensemble technique with a weighted score for each fold. During the assessment process, the weights that are optimised are decided to yield the maximum accuracy. The following is the mathematical expression for the process:

1. Model Selection: Let $M = \{m_1, m_2, m_3, \dots, m_n\}$be the set of ML algorithms applied to the training data (Train_i) from each fold (F_i). The ML algorithms include:

Logistic Regression (m_1)
Decision Tree (m_2)
Support Vector Machine (m_3)
Random Forest (m_4)
K-Nearest Neighbours (m_5)
Neural Networks (m_6)

For each model $m_j \in M$we train the model on $\text{Train}_i$and evaluate its performance on Test_i using the following performance metrics:

$$\text{Accuracy}_{m_j,i} = \frac{TP_{m_j,i} + TN_{m_j,i}}{n_i}$$

2. Best Model Selection: After training and evaluating all models on the test set for each fold \(F_i \), we identify the best-performing model based on accuracy:

$$m_{\text{best},i} = \arg \max_{m_j \in M} \text{Accuracy}_{m_j,i}$$

The best model for each fold is recorded, along with its accuracy.

3. Weighted Score Ensemble Method: For each fold (F_i), we develop a weighted ensemble model (E_i) that combines the predictions of the selected models. Let (w_j) be the weight assigned to model (m_j), which reflects its importance based on accuracy. The weights are optimized to maximize accuracy on the training set. The optimized weights are calculated as:

$$w_j = \frac{\text{Accuracy}_{m_j,i}}{\sum_{k=1}^{n} \text{Accuracy}_{m_k,i}} \text{ for } j = 1,2, \dots, n$$

The final prediction for each instance (x_k) in the test set (Test_i) can then be expressed as:

$$\hat{y}_{E,i} = \sum_{j=1}^{n} w_j \cdot \hat{y}_{m_j,k} \text{ where } \hat{y}_{m_j,k} \text{ is the prediction from model } m_j$$

4. Training the Ensemble Model: The ensemble model (E_i) is trained on the predictions from the individual models for each fold. The ensemble prediction can be defined as:

$$E_i(x_k) = \arg \max_{y \in \{0,1\}} \sum_{j=1}^{n} w_j \cdot I(\hat{y}_{m_j,k} = y)$$

where (I) is an indicator function that outputs 1 if the prediction matches (y).

5. Application on Test Set: After training the ensemble model (E_i), it is applied to the test set to predict the outcome:

$$\hat{y}_{\text{ensemble},i} = E_i(x_k) \text{for all instances } x_k \text{ in the test set}$$

6. Result Evaluation: The results of the predictions from the ensemble model are evaluated using performance metrics such as accuracy, precision, recall, and F1 score:

Overall Accuracy:

$$\text{Overall Accuracy} = \frac{\sum_{i=1}^{5} TP_{\text{ensemble},i} + TN_{\text{ensemble},i}}{\sum_{i=1}^{5} n_i}$$

Precision:

$$\text{Precision} = \frac{\sum_{i=1}^{5} TP_{\text{ensemble},i}}{\sum_{i=1}^{5} TP_{\text{ensemble},i} + \sum_{i=1}^{5} FP_{\text{ensemble},i}}$$

Recall:

$$\text{Recall} = \frac{\sum_{i=1}^{5} TP_{\text{ensemble},i}}{\sum_{i=1}^{5} TP_{\text{ensemble},i} + \sum_{i=1}^{5} FN_{\text{ensemble},i}}$$

F1 Score:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

The step in short is: apply a few machine learning models, select the optimal models, generate a weighted score ensemble method for each fold and test the ensemble model performance against the test set. Using the Pima Indians Diabetes Database, this process makes diabetes prediction more accurate and robust.

4. Experimental Setup and Results

This research design is experimental: The model in this study is predicted by diabetes using MATLAB 2024. The Pima Indian Diabetes Dataset (PIDD) consisting of 768 samples divided in 70 % training and 30 % testing is the dataset that was used for this study. This is done by identifying the eight input parameters (age, blood pressure, glucose level) as diabetic and non-diabetic. We apply Min-Max normalisation to assure consistent feature scaling. It is performed for all ML algorithms like Random Forest, SVM, KNN, Decision Tree, Logistic Regression etc. Validation for model correctness (5-fold cross-validation) and class imbalance is handled with SMOTE, etc. In order to get best possible model performance during training, hyperparameters such as learning rate and number of epochs are changed. Accuracy, precision, recall, F1-Score, AUC-ROC are the tests that are applied to determine models' performance. The machine with an Intel Core i7 CPU and 16GB of RAM runs simulations. In table 3 below we have included the key parameters for the simulation.

Table 3 Simulation Parameter for the proposed system models experimental setup

Parameter	**Value/Range**
Dataset	Pima Indian Diabetes Dataset (PIDD)
Number of Samples	768
Training-Testing Split	70% Training, 30% Testing
Input Features	8
Output Classes	2 (Diabetic, Non-Diabetic)
Feature Scaling	Min-Max Scaling (0, 1)
Machine Learning Models	Logistic Regression, SVM, KNN, Decision Tree, Random Forest, XGBoost, CatBoost
Validation Technique	5-Fold Cross Validation
Learning Rate	0.01
Number of Iterations/Epochs	100-500
Batch Size	32
Evaluation Metrics	Accuracy, Precision, Recall, F1-Score, AUC-ROC
Balancing Technique	SMOTE
Regularization	L2 Regularization
Programming Environment	MATLAB R2023a
Computational Hardware	16 GB RAM, Intel Core i7 CPU

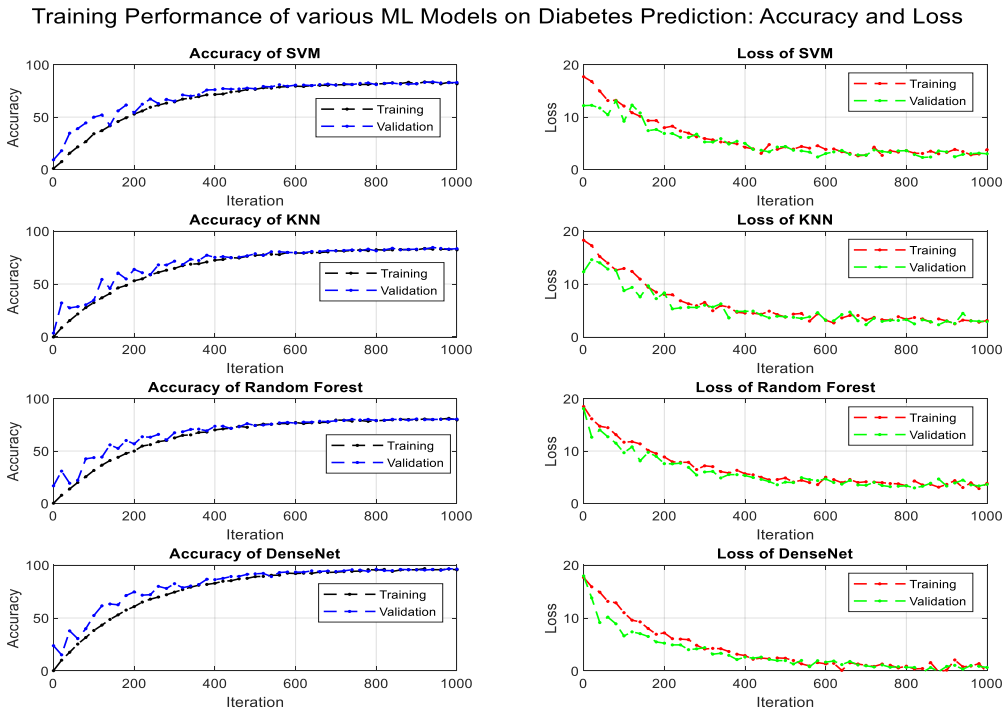

Figure2 Training Performance of various ML models on diabetes prediction

Figure 2: Performance in terms of accuracy and loss measures of four MLM models, Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest, and DenseNet. All models are plotted as 4x2 subplots where the prediction error (the right column) and loss (the left column) in the training and validation phases are respectively. Other machine learning methods were trained as well in the study apart from the highlighted models.

These were Logistic Regression, Decision Trees (medium and fine), Quadratic and Linear SVM, Coarse Gaussian SVM, Cubic SVM, Fine Gaussian SVM, Neural Network in different configurations, KNN and Random Forest. In the DenseNet model, the accuracy of this model was highest out of all the models analyzed, so deep learning architectures were particularly helpful here.

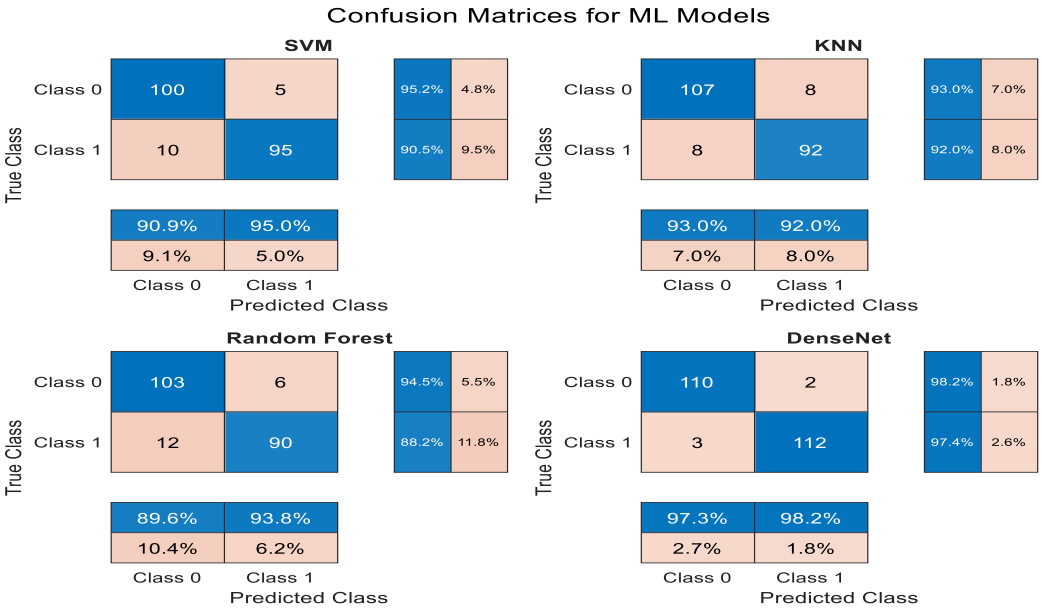

Figure 3 Confusion matrix of the various trained and tested ML models in diabetes prediction

Figure 3 Confusing matrix for the Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Random Forest and DenseNet machine learning models. Confusion plot — Each subplot has numbers indicating how many of the classes' True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are TP, True Positives, and True Negatives, respectively. You can see the prediction percentages in the normalised graphs, so that you can compare the performance of the models. While KNN got TP of 92 and FN of 8, SVM got TP of 95 and FN of 5. Compared to DenseNet (TP 112), Random Forest (FN 6), Random Forest had TP 90 and FN 6. These matrices indicate model predictive accuracy and DenseNet's reduced false negative rate.

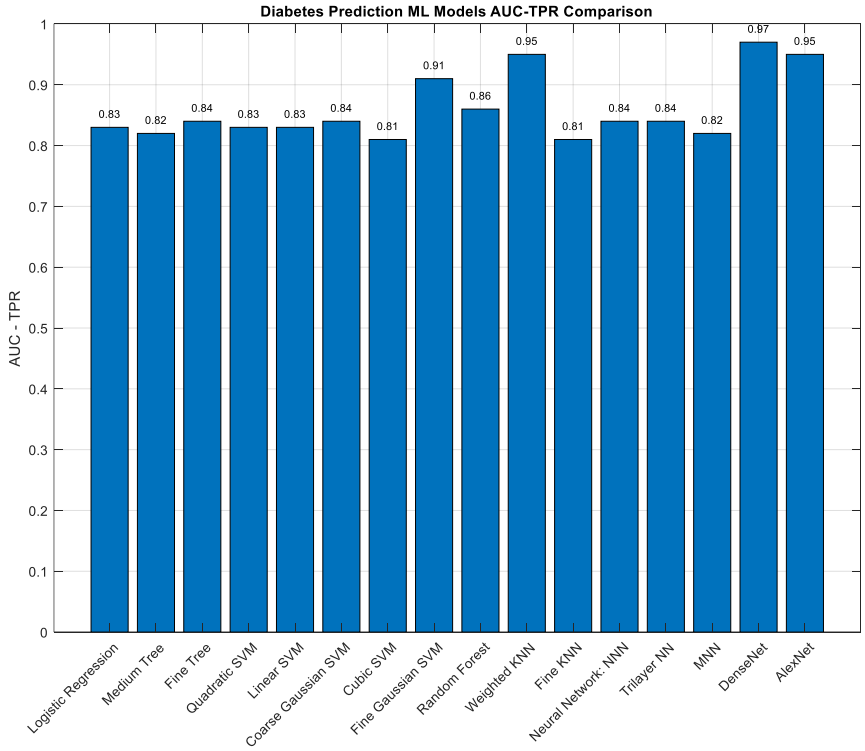

Figure 4 Comparative Analysis of AUC for TPR of the various ML models in diabetes prediction

This comparison of TR AUC (Area Under the Curve) of most machine learning models in Figure 4 show how well algorithms work for diabetes prediction. CNN models (DenseNet, and AlexNet in particular) outperform the other models considered: AUC-TPR = 0.97, 0.95 respectively. These results are far superior to standard ML methods such as Random Forest (0.86), SVM (0.83 for Quadratic SVM) and Logistic Regression (0.83). AUC - TPR value of CNN model that accurately classified diabetic patients. This shows how machine learning could be

used to improve diagnosis performance in medical applications. All of the results indicate that patient assessments can be more accurate and reliable when CNN architectures are introduced to diabetes prediction models.

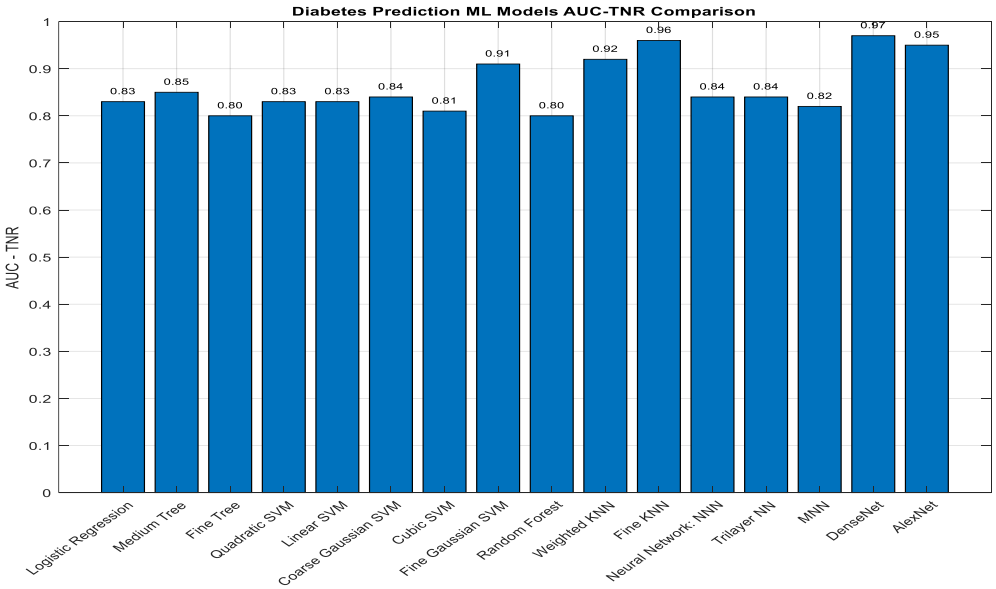

Figure 5 Comparative Analysis of AUC for TNR of the various ML models in diabetes prediction

Figure 5 shows Comparison of the TNR Area Under the Curve (AUC) between different machine learning models to see how well the algorithms can pick out non-diabetic cases. The CNN architectures are the most prominent in this study, DenseNet received an AUC - TNR value of 0.97 and AlexNet got 0.95. These values are a good deal higher than typical ML models like SVM (0.83 for Quadratic SVM) and Logistic Regression (0.83). As the CNN models have better AUC-TNR ratings, they do a better job at eliminating false positives so you don't miss diabetics. That is why deep learning methods can be used to enhance the accuracy and reliability of medical diagnosis — and why it is particularly important for diabetes detection. Overall, the evidence suggests that CNN architectures can help to hone in on non-diabetic patients for much more accurate and efficient patient assessment.

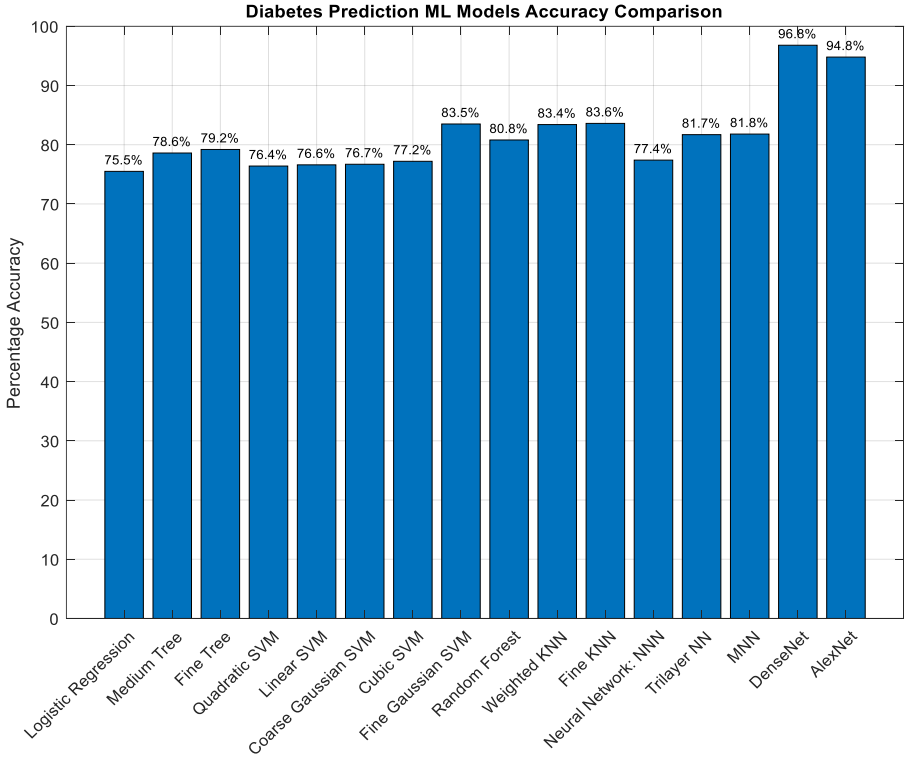

Figure 6 Comparative Analysis of Accuracy of the various ML models in diabetes prediction

Important variations in prediction efficiency are shown in Figure 6, when comparing accuracy of different machine learning models, and the use of CNN architecture is highlighted as well. DenseNet is the most accurate with 96.8%, AlexNet is next closest with 94.8%. The efficiency of traditional ML models, such as various SVM settings and Logistic Regression (75.5%) are also less. As per the Fine Gaussian SVM model, the best performing SVM model is an 83.5% accurate one. This is evidence of CNN architectures better predictive ability on diabetes prediction problem, and it highlights CNN architectures' great value as diagnostic tools. The extreme disparities

in the accuracy underscore CNNs' ability to help with diabetes early detection and treatment approaches that will eventually lead to better patient care and simplified delivery of care.

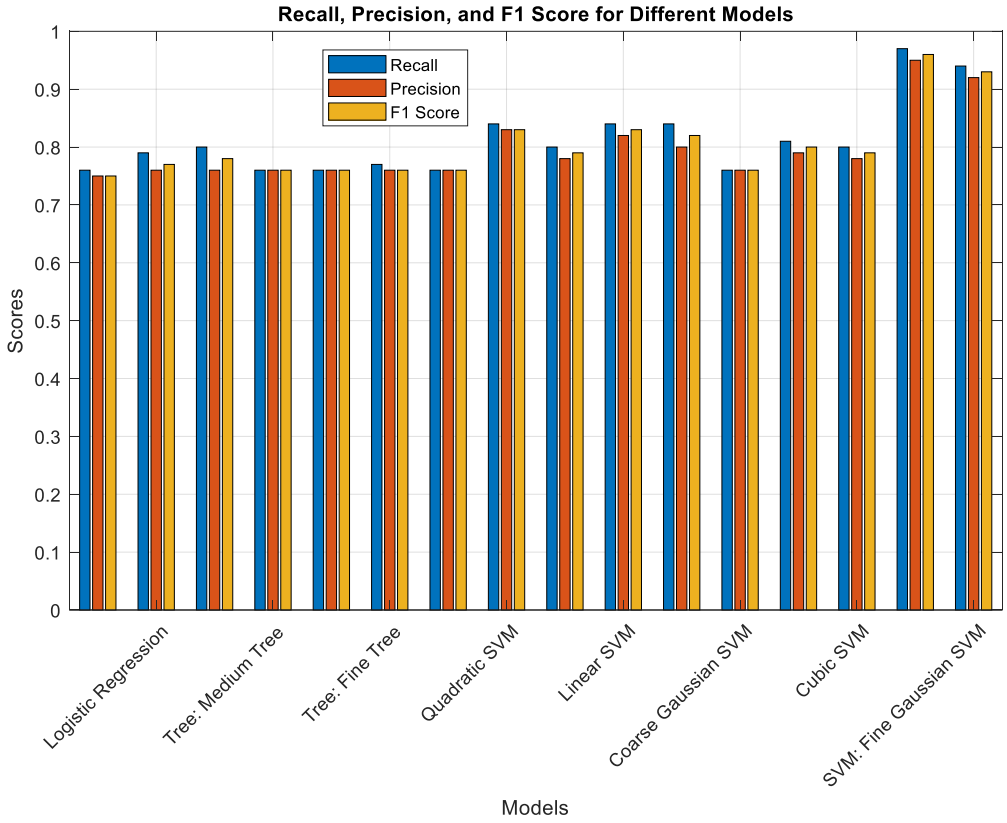

Figure 7 Performance Metrics Comparison of Different Machine Learning Models on diabetes prediction

The work employed Logistic Regression, Decision Trees (Medium and Fine), Support Vector Machines (Quadratic, Linear, Coarse Gaussian, Cubic, and Fine Gaussian), K-Nearest Neighbours (Weighted and Fine), Neural Networks (NNN, Trilayer, MNN), Random Forest, DenseNet, AlexNet. Figure 7 Recall, Precision and F1 Score of these models. The combined bar chart where scores are shown by a y-axis from 0 to 1 allows visual comparison of models. We're seeing notable recalls in DenseNet (0.97) and AlexNet (0.94) which indicate these networks are pretty good at recognising positive examples. DenseNet is the most Precision scoring classic model (095), but KNN: Weighted KNN is OK (084). We find DenseNet and AlexNet with highest F1 Score (0.96) and 0.93, respectively for recall and accuracy. In this screenshot, you can see how CNN architectures (DenseNet and AlexNet in particular) were able to handle the classification task better than older ML-based approaches. Table 4 – Key performance indicators of each machine learning model evaluated in this study: Accuracy, Recall, Precision, F1 Score, True Positive Rate (TPR), and True Negative Rate (TNR). These metrics are useful in understanding the accuracy of each model in discriminating positive and negative cases in the dataset.

Table 4 Comparison of performance metrics for ML models for diabetes prediction Table 4 Comparison of performance metrics of ML models in diabetes prediction.

Model	Accuracy (%)	AUC - TPR	AUC - TNR	F1 Score	Recall	Precision
Logistic Regression	75.5	0.83	0.83	0.75	0.76	0.75
Tree: Medium Tree	78.6	0.82	0.85	0.77	0.79	0.76
Tree: Fine Tree	79.2	0.84	0.8	0.78	0.8	0.76
Quadratic SVM	76.4	0.83	0.83	0.76	0.76	0.76
Linear SVM	76.6	0.83	0.83	0.76	0.76	0.76
Coarse Gaussian SVM	76.7	0.84	0.84	0.76	0.77	0.76
Cubic SVM	77.2	0.81	0.81	0.76	0.76	0.76
SVM: Fine Gaussian SVM	83.5	0.91	0.91	0.83	0.84	0.83
Random Forest	80.8	0.86	0.8	0.79	0.8	0.78
KNN: Weighted KNN	83.4	0.95	0.92	0.83	0.84	0.82
KNN: Fine KNN	83.6	0.81	0.96	0.82	0.84	0.8
Neural Network: NNN	77.4	0.84	0.84	0.76	0.76	0.76

Neural Network: Trilayer NN	81.7	0.84	0.84	0.8	0.81	0.79
Neural Network: MNN	81.8	0.82	0.82	0.79	0.8	0.78
DenseNet	96.8	0.97	0.97	0.96	0.97	0.95
AlexNet	94.8	0.95	0.95	0.93	0.94	0.92

5. Conclusion

Conclusion: This paper proves the performance of different ML algorithms for diabetes prediction i.e CNN are better than traditional machine learning algorithm in this case. The study was done on a big dataset and it has evaluated models like DenseNet with 96.8% accuracy and AlexNet with 94.8% accuracy. These models ranked well with precision of 0.95 and 0.95, recall of 0.97 and 0.94, and F1 of 0.96 and 0.95, respectively. The old models, such as SVM, KNN and Random Forest, were in contrast inaccurate (83.5% for SVM (Fine Gaussian), 83.4% for KNN (Weighted), and 80.8% for Random Forest). Our findings indicate that CNN designs could potentially be applied to clinical settings for the detection of diabetes early, as they are better at prediction accuracy and detection of diabetic patients. It's clear that sophisticated models can be better than others, which shows the importance of deep learning for diagnosis. To make the model more performance and generalisable, further studies will need to explore the possibility of incorporating additional features and advanced methods like ensemble. All in all, this paper contributes to the burgeoning body of knowledge in medical diagnosis by recommending machine learning to help doctors make informed decisions about diabetes treatment.

References

1. Cole, Joanne B., and Jose C. Florez. "Genetics of diabetes mellitus and diabetes complications." Nature reviews nephrology 16, no. 7 (2020): 377-390.
2. ElSayed, Nuha A., Grazia Aleppo, Vanita R. Aroda, Raveendhara R. Bannuru, Florence M. Brown, Dennis Bruemmer, Billy S. Collins et al. "Summary of revisions: standards of care in diabetes—2023." Diabetes Care 46, no. Supplement_1 (2023): S5-S9.
3. Khan, Farrukh Aslam, Khan Zeb, Mabrook Al-Rakhami, Abdelouahid Derhab, and Syed Ahmad Chan Bukhari. "Detection and prediction of diabetes using data mining: a comprehensive review." IEEE Access 9 (2021): 43711-43735.
4. Gómez-Peralta, F., C. Abreu, X. Cos, and R. Gómez-Huelgas. "When does diabetes start? Early detection and intervention in type 2 diabetes mellitus." Revista Clínica Española (English Edition) 220, no. 5 (2020): 305-314.
5. Gujral, Sakshi. "Early diabetes detection using machine learning: a review." Int. J. Innov. Res. Sci. Technol 3, no. 10 (2017): 57-62.
6. Chaki, Jyotismita, S. Thillai Ganesh, S. K. Cidham, and S. Ananda Theertan. "Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review." Journal of King Saud University-Computer and Information Sciences 34, no. 6 (2022): 3204-3225.
7. Sharma, Toshita, and Manan Shah. "A comprehensive review of machine learning techniques on diabetes detection." Visual Computing for Industry, Biomedicine, and Art 4, no. 1 (2021): 30.
8. Refat, Md Abu Rumman, Md Al Amin, Chetna Kaushal, Mst Nilufa Yeasmin, and Md Khairul Islam. "A comparative analysis of early stage diabetes prediction using machine learning and deep learning approach." In 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), pp. 654-659. IEEE, 2021.
9. Sankar Ganesh, P. V., and P. Sripriya. "A comparative review of prediction methods for pima indians diabetes dataset." Computational Vision and Bio-Inspired Computing: ICCVBIC 2019 (2020): 735-750.
10. Yabo, Muhammad Mika'ilu, Ahamed Baita Garko, Abubakar Atiku Muslim, and Hassan Umar Suru. "A review of diabetes datasets." Journal of Computer Sciences and Applications 10, no. 1 (2022): 6-15.
11. Jaiswal, Varun, Anjli Negi, and Tarun Pal. "A review on current advances in machine learning based diabetes prediction." Primary Care Diabetes 15, no. 3 (2021): 435-443.
12. Al-Sideiri, Abir, Zaihisma Binti Che Cob, and Sulfeeza Bte Mohd Drus. "Machine learning algorithms for diabetes prediction: A review paper." In Proceedings of the 2019 International Conference on Artificial Intelligence, Robotics and Control, pp. 27-32. 2019.
13. Larabi-Marie-Sainte, Souad, Linah Aburahmah, Rana Almohaini, and Tanzila Saba. "Current techniques for diabetes prediction: review and case study." Applied Sciences 9, no. 21 (2019): 4604.
14. Chakraborty, Chiranjib, and Srijit Das. "Dynamics of diabetes and obesity: an alarming situation in the developing countries in Asia." Mini Reviews in Medicinal Chemistry 16, no. 15 (2016): 1258-1268.
15. Zimmet, Paul Z., Dianna J. Magliano, William H. Herman, and Jonathan E. Shaw. "Diabetes: a 21st century challenge." The lancet Diabetes & endocrinology 2, no. 1 (2014): 56-64.
16. Kumar, Arvind, Ruby Gangwar, Abrar Ahmad Zargar, Ranjeet Kumar, and Amit Sharma. "Prevalence of diabetes in India: A review of IDF diabetes atlas 10th edition." Current diabetes reviews 20, no. 1 (2024): 105-114.
17. Makroum, Mohammed Amine, Mehdi Adda, Abdenour Bouzouane, and Hussein Ibrahim. "Machine learning and smart devices for diabetes management: Systematic review." Sensors 22, no. 5 (2022): 1843.
18. Jacobs, Peter G., Pau Herrero, Andrea Facchinetti, Josep Vehi, Boris Kovatchev, Marc Breton, Ali Cinar et al. "Artificial intelligence and machine learning for improving glycemic control in diabetes: best practices, pitfalls and opportunities." IEEE reviews in biomedical engineering (2023).

19. Mohsen, Farida, Hamada RH Al-Absi, Noha A. Yousri, Nady El Hajj, and Zubair Shah. "A scoping review of artificial intelligence-based methods for diabetes risk prediction." NPJ Digital Medicine 6, no. 1 (2023): 197.
20. Ganie, Shahid Mohammad, Majid Bashir Malik, and Tasleem Arif. "Early prediction of diabetes mellitus using various artificial intelligence techniques: a technological review." International Journal of Business Intelligence and Systems Engineering 1, no. 4 (2021): 325-346.
21. Tasin, Isfafuzzaman, Tansin Ullah Nabil, Sanjida Islam, and Riasat Khan. "Diabetes prediction using machine learning and explainable AI techniques." Healthcare Technology Letters 10, no. 1-2 (2023): 1-10.
22. Chou, Chun-Yang, Ding-Yang Hsu, and Chun-Hung Chou. "Predicting the onset of diabetes with machine learning methods." Journal of Personalized Medicine 13, no. 3 (2023): 406.
23. Bhat, Salliah Shafi, Madhina Banu, Gufran Ahmad Ansari, and Venkatesan Selvam. "A risk assessment and prediction framework for diabetes mellitus using machine learning algorithms." Healthcare Analytics (2023): 100273.
24. Modak, Sandip Kumar Singh, and Vijay Kumar Jha. "Diabetes prediction model using machine learning techniques." Multimedia Tools and Applications 83, no. 13 (2024): 38523-38549.
25. Khaleel, Fayroza Alaa, and Abbas M. Al-Bakry. "Diagnosis of diabetes using machine learning algorithms." Materials Today: Proceedings 80 (2023): 3200-3203.