INTERNATIONAL JOURNAL OF Telecommunications
AIR DEFENSE COLLEGE

# A Diabetes Mellitus Prediction Model Based on Supervised Machine Learning Techniques

**Citation:** El Sherbiny, M.; Abdel Fattah, M.; Rabie, A.; Taki Eldin, A., Moustafa, H. *Inter. Jour. of Telecommunications,* ***IJT'2025****, Vol. 05, Issue 01, pp. 1-11, 2025.*

**Editor-in-Chief:** Youssef Fayed.

Received: 09/02/2025.

Accepted: date 06/03/2025.
Published: date 06/03/2025.

**Publisher's Note: The International Journal of Telecommunications, IJT, stays neutral regarding jurisdictional claims in published maps and institutional affiliations.**



**El Sherbiny, Moataz Mohamed*[1], Abdel Fattah, Mohamed Gamal[2], Rabie, Asmaa Hamdy[3], Taki Eldin, Ali Elsherbiny[4], and Moustafa, Hossam El-din[5]**

* Correspondence: Electronics and Communications Department, Faculty of Engineering, Mansoura University, Dakahlia, Egypt. moatazelsherbiny@mans.edu.eg
[2] Electronics and Communications Department, Faculty of Engineering, Mansoura University, Dakahlia, Egypt. eng.mo.gamal@mans.edu.eg
[3] Computers and Control Systems Engineering Department, faculty of engineering Mansoura University, Mansoura, Egypt. asmaahamdy@mans.edu.eg.
[4] Head of the Cyber Security Department, Faculty of Artificial Intelligence, Delta University for Science and Technology, Dakahlia, Egypt. a_takieldeen@deltauniv.edu.eg
[5]Electronics and Communications Department, Faculty of Engineering, Mansoura University, Dakahlia, Egypt. hossam_moustafa@mans.edu.eg

**Abstract:** It is no doubt that diabetes is considered one of the most common chronic diseases. Diabetes patients have high risk of diseases like renal failure, heart stroke, nerve and eye damage that can lead to blindness. Detection and prediction of diabetes mellitus is not a very easy process. Nevertheless, the cost of tests is high. It could be a revolutionary if one could know the risk of being diabetic with no need to visit busy hospitals. This could only be done through artificial intelligence. In this paper, a classification model was proposed for diabetes mellitus classification and prediction, so that early diagnosis as well as treatment could prolong patients' lives and minimize risk factors. The classification of datasets in medical healthcare is hindered by the problem of having suitable datasets. Proper processing was performed through null values imputation, normalization and encoding. Supervised algorithms were applied to ensure the effectiveness of the proposed model such as Random Forest (RF), Extreme Gradient Boosting (XGB) and Neural Network (NN). Results were compared using five performance metrics; accuracy, precision, f1-score, recall and run time. Training and testing are performed on two different datasets to ensure the robustness of proposed model. Proper preprocessing was conducted to handle the issues of non-existing values and class imbalance that may lead to a misleading biased model. Results demonstrated that RF has overtaken both remaining techniques by achieving 80.5% accuracy compared to 79.65% for XGB and 76.36% for NN on the Pima Indian Diabetes Dataset. While the questionnaire dataset results indicated RF superiority among remaining models by achieving an accuracy of 97.11% compared to 93.38% and 93.26 for NN and XGB respectively.

**Keywords:** Diabetes Mellitus, Preprocessing, Supervised Machine Learning, Performance Metrics.

## 1. Introduction

Diabetes poses major risks to various body organs since it reduces the ability of extracting energy from consumed food. It is considered as a disorder that elevates the insulin level in blood beyond the normal threshold value of 126 milligrams per deciliter (mg/dL)[1]. Diabetes types delves into three main categories as follows:

Type-1 Diabetes, Type-2 Diabetes and Gestational Diabetes. They are different in threats, causes, symptoms as well as treatment. Table 1 summarizes comparison between them. However, general symptoms such as dehydration, starvation, repeated urination, drowsiness, distorted vision, slow-wound healing time could be like other chronic diseases. Type-2 Diabetes Mellitus is a condition marked by insulin resistance, where the cells of the body no longer respond effectively to insulin [2]. In Type 2, It has primarily affected adults. According to statistics, between 90 and 95 percent of people will develop type 2 diabetes. Type 2 diabetes is commonly controlled with diet, exercise, and weight control. Still, drugs or injections could be regarded as a treatment to minimize blood sugar levels [3]. The World Health Organization (WHO) reveals that the demise percentage will be 90 percent between 2017 to 2030. Moreover, it is predicted that type-2 disease will rise to 438 million in 2030 [4][5]. Referring to the International Diabetes Federation (IDF), the prevalence of diabetes has been elevating rapidly, caused by lifestyle changes, aging populations, and urbanization. Global Diabetes Statistics in 2024 there are total cases of nearly 540 million adults between 20 and 80 years have diabetes worldwide. This number is expected to elevate by 2045 to 785 million. Moreover, undiagnosed Cases are approximately 250 million people living with undiagnosed diabetes making it around 1 in 2 people with diabetes [6]. Diabetes contributes to 6.5 million deaths annually leading to 1 every 5 seconds mortality rate. Middle Eastern countries are estimated to report some of the highest prevalence rates. Also, Africa is predicted to have a 129% increase in diabetes cases. In Egypt, there are about 20.9% (including diagnoses and undiagnosed cases) of having diabetes, which is significantly higher than the global average 10.5%[7].

**Table 1**. "Comparison between different types of diabetes".

| Point of Comparison | Type-1 Diabetes | Type-2 Diabetes | Gestational Diabetes |
|---|---|---|---|
| **Definition** | Autoimmune disease attacking insulin producing cells | Chronic condition where body does not produce enough insulin | Disease developed during pregnancy |
| **Cause** | Autoimmune reaction | Insulin resistance | Hormonal changes |
| **Risk factors** | Family history and genetic predisposition | Obesity, sedentary lifestyle, and ethnicity | History of gestational diabetes and advanced maternal age |

The paper's key contributions are given below:

- The main goal is to design an accurate machine learning model that has undergone training using the same number of instances for each class independent of dataset original distribution and size.
- Employed a proper preprocessing approach to handle the non-existing values, data rescaling and class imbalance.
- Model performance was improved via balancing the dataset using the synthetic minority oversampling technique (SMOT).
- RF achieved an overall accuracy of 80.51%, as well as AUC, precision, f1-score, recall and run time of 76%, 79.03%, 68.53%, 60.5% and 240.6 msec respectively on PIDD and accuracy of 97.11%, as well as AUC, precision, f1-score, recall and run time of 97%, 96.87%, 97.63%, 98.41% and 173.65 msec on questionnaire dataset.

The remaining parts of the manuscript are organized step by step: In the second section, review on previous studies. Then, dataset description and stating the methodology in the third section, and finally, results and discussion represented in the fourth section. Section five presents the conclusion of this paper.

## 2. Literature Review

Most of previous studies have been conducted on the same dataset named Pima Indian Diabetes Dataset (PIDD). It contains many missing values and to outliers which could be misleading to the training process of machine learning models. Several worthy contributions to this connection are listed below.

Khalel et al. [8] proposed a prediction model using supervised machine learning algorithm on the Pima Indian Diabetes (PIDD) dataset. Authors applied Naïve Bayes (NB) and K-nearest Neighbor (KNN) that yielded 79% and 69% respectively in terms of accuracy.

Febrian et al. [9] compared two supervised machine learning techniques which are KNN and NB algorithms to predict diabetes based on specific health attributes in the PIDD. The results were evaluated using confusion Matrix. Results indicated that Naive Bayes algorithm outperformed KNN, with an average value of 76.07 % accuracy, 73.37% precision, and 71.37% recall in NB. Moreover, results showed an accuracy of 73.33%, precision of 70.25%, and recall of 69.37% in KNN.

Kangra et al. [10] utilized various supervised machine learning techniques to aid in diabetes prediction: NB, KMM, SVM and LR. The experiment was piloted on the PIDD. The results of classifiers in terms of accuracy are as follows: NB scored 72.6%, KNN 66.1%, DT 71.8%, RF 64.9%.

Quan et al.[11] proposed a diabetes prediction model with the aid of two supervised machine learning algorithms. Authors study utilized PIDD. Results scored an average accuracy of 72.59% and 75.19% for Decision Tree and Random Forest algorithms respectively.

Amani et al. [12] applied both machine learning and deep learning on the PIDD dataset. The two machine learning techniques are SVM and RF. Authors stated that SVM are not able to solve non-linear function. Therefore, they applied mapping functions. Results indicated that the SVM and RF algorithms achieved an accuracy of 73.94% and 79.26% respectively.

Muhammed et al.[13] constructed a diabetes detection model for their research using various supervised machine learning algorithms including: SVM, K-Nearest Neighbor (KNN), Logistic Regression (LR), Naive Bayes (NB), DT, and RF. Authors cleaned noisy data through outliers and normalization. Result shows that both KNN and SVM models outperformed the remaining proposed models, with 77% accuracy in the experiment on PIDD. DT and RF in the last place with an accuracy of 71%.

Sajratul et al. [14] assured in research work that feature selection can maintain adequate accuracy while computational cost required decreased significantly. Implementation of feature selection subset from the entire PIDD dataset with pregnancies number, level of blood glucose, body mass index, age and diabetes pedigree function yielded maximum accuracy. Logistic Regression and Random Forest proposed models achieved 77.08% and 75% of accuracy respectively when the previous attributes are fed into them.

Huma et al. [15] proposed sampling methods to extract features such as Linear Sampling, Shuffled Sampling, and Automatic Sampling. These sampled features are fed into Naïve Bayes model achieving an accuracy of 76.33% on the PID dataset. Results can be improved by solving the null values problem, which is not presented in their study.

Harleen et al. [16] performed their research depending on SVM. Dimensionality reduction was performed on the PIDD dataset. The outliers were removed using statistical methods to improve the model accuracy. Results indicated that SVM achieved high accuracy in the training stage. However, accuracy went down to 71.3% in testing indicating probability of overfitting.

## 3. Material and Methodology

This section presents the suggested framework as well as methods of this research. It contains three main building blocks. Starting with the brief description of used datasets, moving to preprocessing of data and machine learning models. Lastly, evaluation metrics as illustrated in Figure 1.

### *3.1 Dataset Description*

Labeled data is a crucial input to supervised machine learning and deep learning classification problems [17]. A relevant collection of data aids to better machine learning classification. There are two datasets implemented in this paper which were gathered from public hosts and by agreements with medical centers and doctors. They are publicly available online hosted by UCI Machine Learning. The first dataset named Pima Indians Diabetes Dataset (PIDD) which is considered as one of the most well-known datasets in binary classification of diabetes using machine learning. It was created by the National Institute of Diabetes and Digestive and Kidney Disease. The dataset includes 768 females. The dataset implements 7 medical predictive variables in addition to one target value labeled Outcome. Predictor attributes based on certain diagnostic measurements included in the dataset. The second dataset is submitted using a questionnaire for diabetes prediction case study. It has been collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. It contains 520 samples and 17 predictive features. Detailed description of features in both datasets is illustrated in Tables 2 and 3.

**Table 2**. "Description of attributes present in PIDD".

| Attribute | Description | Missing values Percentage | Missing values count | Range |
|---|---|---|---|---|
| Pregnancies | Number of times a patient has been pregnant | - | - | 0-17 |
| Glucose | Concentration of plasma glucose at two hours in an oral glucose tolerance test (GTIT) | 23.4375% | 180 | 0-199 |
| BP | Diastolic Blood pressure (mm Hg) | 28.77% | 221 | 0-122 |
| ST | Skin fold thickness in Triceps (mm) | 38% | 292 | 0-99 |
| Insulin | Serum Insulin for two hours ($\mu$/ml) | 64.84% | 498 | 0-846 |
| BMI | Body mass index (kg/m) | 10.42% | 80 | 0-67.1 |
| DPF | Diabetes Pedigree Function | - | - | 0.078 – 2.42 |
| Age | Age in years | - | - | 21 -81 |
| Outcome | Binary target indicating diabetic or not | - | - | 0 - 1 |

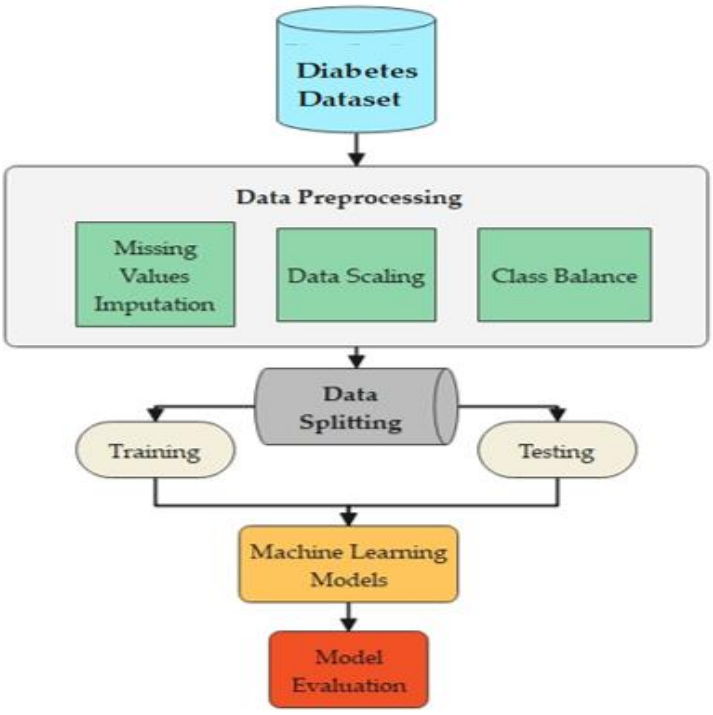



**Figure1.** Block diagram of the proposed work.

### *3.2 Data Preprocessing*

An extremely important step when developing the prediction model for this study is preprocessing [18]. Given that the collected data involves missing nominal variables along with outliers, the improper data affects the liability of proposed model [19]. The PIDD has several missing values as shown in Table 2. Patients often ignore many tests for many reasons such as cost and time, which results in missing values. Thus, diagnostic variables cannot be determined, hence an appropriate imputation approach must be used.

**Table 3**. "Description of attributes present in questionnaire dataset".

| Attribute | Description | Range (Distribution) |
|---|---|---|
| Age | Age of person in years | 16-90 |
| Gender | Sex of patient | Male (63%) or Female (37%) |
| Polyuria | Excess urination | Yes (50% or No (50%) |
| Polydipsia | Excess thirst | True (45%) or False (55%) |
| Sudden Weight Loss | Unintentional and rapid weight loss | True (42%) or False (58%) |
| Weakness | Reduced energy | True (59%) or False (41%) |
| Polyphagia | Excessive hunger or increased appetite | True (46%) or False (54%) |
| Genital Thrush | Fungal infection | True (22%) or False (78%) |
| Visual Blurring | Difficulty in seeing clearly | True (45%) or False (55%) |
| Itching | Persistent skin pruritus | True (49%) or False (51%) |
| Irritability | Emotional sensitivity or mood swings | True (24%) or False (76%) |
| Delayed Healing | Slow wound healing | True (46%) or False (54%) |
| Partial Paresis | Weakness or paralysis of muscle group | True (43%) or False (57%) |
| Muscle Stiffness | Reduced flexibility in muscles | True (38%) or False (62%) |
| Alopecia | Hair loss and hormone imbalance | True (34%) or False (66%) |
| Obesity | Excess body fat | True (17%) or False (83%) |
| Class | Binary target indicating diabetic or not | Positive (62%) or Negative (38%) |

#### 3.2.1 Missing Values Imputation

Dealing with incomplete medical records can be performed through different methods[20]. Replacing missing feature values with zero has no effective biasing in prediction but this assumption is medically impossible. Neglecting incomplete record by simply removing them can affect small-scale datasets. Other mathematical approaches such as replacing non-existing values with a constant, mean, median or most frequent. Mean Imputation was applied in this study to replace missing values with the mean of the non-missing values for that feature. Since the features are continuous and the percentage of missing data is small. This technique is known for its computational simplicity, easy of implementation and speed. Moreover, it preserves the size of dataset from the loss of data due to deletion.

#### 3.2.2 Scaling

Before fitting any model, Rescaling is required to improve its speed and accuracy. Since scaling assists models to comprehend the problem, it is typically essential to scale numerical descriptive values. This is because numerous important techniques demand it. Standard scaling, which sets the attribute to 0 mean and 1 standard deviation, was used in this work even though there are other ways to scale data. Normalization resizes each data value in a certain range using the (1) while distribution is shifted during standardization using (2);

$$X_i = \frac{X - Min.Value}{X - Max.Value} \tag{1}$$

$$y = \frac{x - \mu}{\sigma} \quad (2)$$

Where $X_i$ is scaled data, X is data to be scaled, $\mu$ is the mean while $\sigma$ is the standard deviation.

### 3.2.3 Class balance

Class imbalance occurs when the number of samples in one class of a dataset significantly outnumbers the samples in other classes. This imbalance can create challenges for machine learning models[21], as they tend to favor the majority class, leading to biased model predictions and poor misleading performance especially of the minority class. There are several techniques to handle class imbalances such as oversampling minority class, under sampling majority class or data augmentation which is typically used in image dataset [22]. PIDD includes 500 diabetic patients while the number of non-diabetic individuals is 268 as illustrated in Figure 2. A simple solution might seem to be duplicating existing minority class samples, but this does not add any new information to the dataset where exact duplicates can lead to overfitting, where the model memorizes the minority samples instead of learning generalizable patterns. Therefore, Oversampling Technique (SMOT) was applied in this study.


**Figure 2.** Shows class distribution as percentages in (a) PIDD dataset (b) Questionnaire dataset.

### *3.3 Machine Learning*

For training and testing purposes, the dataset is divided into two sections before training machine learning. The model learns to identify relationships between the data during training in order to produce accurate predictions on new unknown data during the testing stage to assess the trained model's performance. Common spliiting ratios are 80%-70% or 20%-30% for training and testing respectively.

Machine Learning (ML) make use of algorithms and mathematical algorithms in order to identify patterns in data and make predictions. ML systems enhance their performance over time through being exposed to more data. Supervised learning trains on labeled data used in classification as in our case. Given the categories of the input data, classification is a predictive modelling methodology that predicts a class label. For the classification of diabetes mellitus, this study involves two machine learning classifiers, such as Random Forest (RF), Extreme Gradient Boosting (XGB) and Neural Network (NN).

RF develops multiple decision trees in the training stage and provides output class of those individual trees. This model provides a simple adjustment that utlizie ecorrelated tree through bagging process. Bagging develops using bootstrapped samples multiple DTs . During bootstrapping, specific number of attributes are neglected from the entire coloumns [23].

An XGB is a tree-based sequential DT algorithm applied to relatively small or medium size tabular datasets [24]. It is considered to be among the most effective techniques for classification and prediction. By combining comparatively weaker and simpler models. Scalability is considered the most important feature in XGB [25] , where it implement learning through distributed computing and memory usage is well structured .

On the basis of multiple layers and neurones, NNs are built for classification and a host of other applications. The two main characteristics of neural networks are their capacity to learn how to do their tasks after being adequately trained and to generalise and generate a satisfactory response to unseen data [26]. The NN is initially given the best subset of characteristics as inputs. Every neurone computes a weighted sum for every feature subset.The output value of the neurone is then calculated by applying a transfer function to this weighted sum.

### *3.4. Model Evaluation Metrics Parameters*

A confusion matrix is a table used to describe the classification algorithm performance. Visualize and summarize the performance of classification algorithms [27]. It summarizes correct and incorrect predictions broken into categories. The confusion matrix comprises four main parameters that are used to create the classifiers measuring metrics [22]. These four outcomes are described below: True Positive (TP): means that the actual and predicted values are the same.True Negative (TN): This represents the number of predictions that the classifier correctly predicted that the negative class would be negative. False Positive (FP): negative class predicts a positive category. False Negative (FN): positive cases were misclassified to other classes. Accuracy is the indicator of cases thar are correctly identified from entire cases. Precesion is the ratio of correctly predicted postive out comes to all postive values. Recall is proprtion of correctly predcited cases amon all predicted values. While the average of precision and recall is called f1-score. The formulas for accuracy, precision, recall, specificity and f1 score are presented in equations (3)–(6), as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

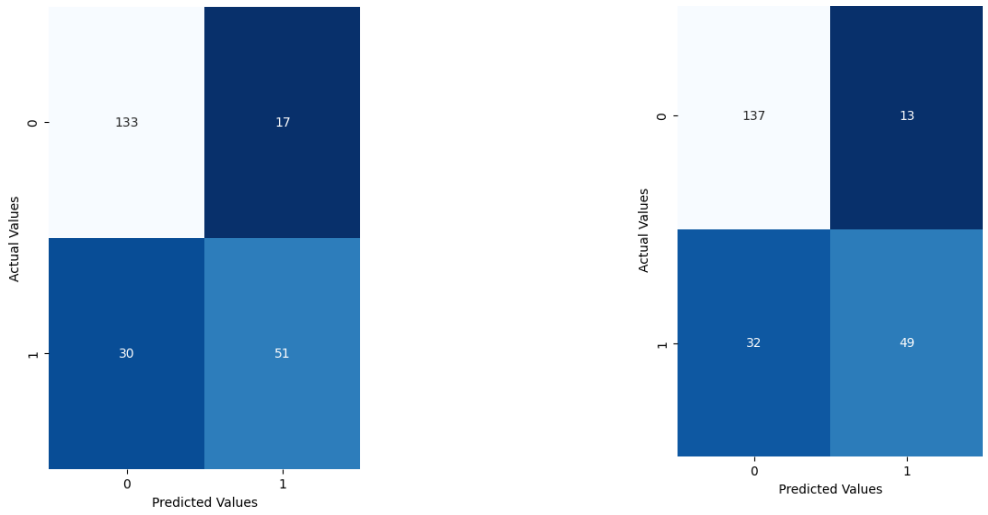
$$F1 - score = \frac{2 * Precsion * Recall}{Precision + Recall} \tag{6}$$

## 4. Experimental results and discussion

This study implemented simple imputation approach using mean value to replace non-existing ones. Dataset is delved into two partations; 80% for training, 20% for validation and testing purposes after shuffling data in random states to prevent overfitting problems. Training and testing have been applied on kaggle platform. Random Forest, Extreme gradient boosting and Neural networks are used for the classification of diabetic and non-diabetic individuals in this paper. The criterion function used to measure split equality in Random forest is gini for classification purposes. While Minimum number of samples required to split an internal node is 2.
The learning rate in XGB is set to 0.3, which defines the step size shrinkage to prevent overfitting. Moreover, the maximum depth of tree is 6. The activation function used to map input to output in neural network is sigmoid. Batch size is set to 16 while the number of epochs is 100. Confusion matrices of XGB and RF are applied on the two dataset and their results are illustrated in Figure 3 and 4. Training and validation results of neural network is represented in Figure 5. Accuracy and loss curves of neural network is presented in Figure 5 and 6. Results are then compared with relevant previous studies stated in literature review summarized in Table 3.

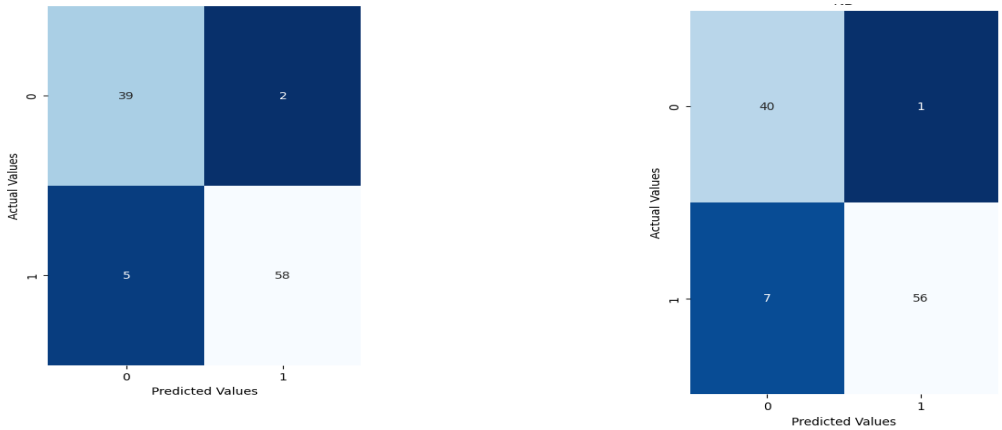
**Table 3**. "A comparison of the outcomes between the proposed system and methodologies found in the literature".

| Authors | Dataset | Methodology | Accuracy |
|---|---|---|---|
| Khalel et al. | Pima Indian Diabetes Dataset (PIDD) | KNN | 69% |
| | | NB | 79% |
| Febrian et al. | | KNN | 69.37% |
| | | NB | 76.07% |
| Kangra et al. | | NB | 72.6% |
| | | DT | 66.1% |
| | | RF | 71.8% |
| Quan et al. | | KNN | 64.9% |
| | | DT | 72.59% |
| | | RF | 75.19% |
| Amani et al. | | SVM | 73.94% |
| | | RF | 79.26% |
| Muhamed et al. | | KNN and SVM | 77% |
| | | LR and NB | 74% |
| | | DT and RF | 71% |
| Sajratul et al. | | LR | 77.08% |
| | | RF | 75% |
| Huma et al. | | NB | 76.33% |
| Harleen et al. | | SVM | 71.3% |
| Proposed Model | Questionnaire Dataset | RF | 97.11% |
| | | XGB | 93.27% |
| | | NN | 93.38% |
| | Pima Indian Diabetes Dataset (PIDD) | RF | 80.5% |
| | | XGB | 79.65% |
| | | NN | 76.45% |



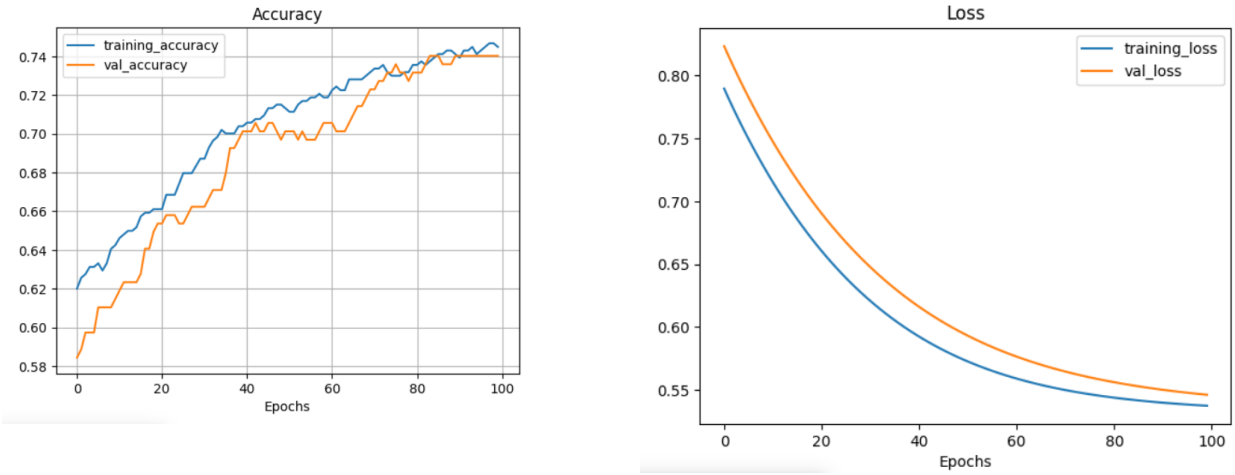

(a) (b)

**Figure 3.** Shows PIDD confusion matrices for (a) XGB (b) RF.



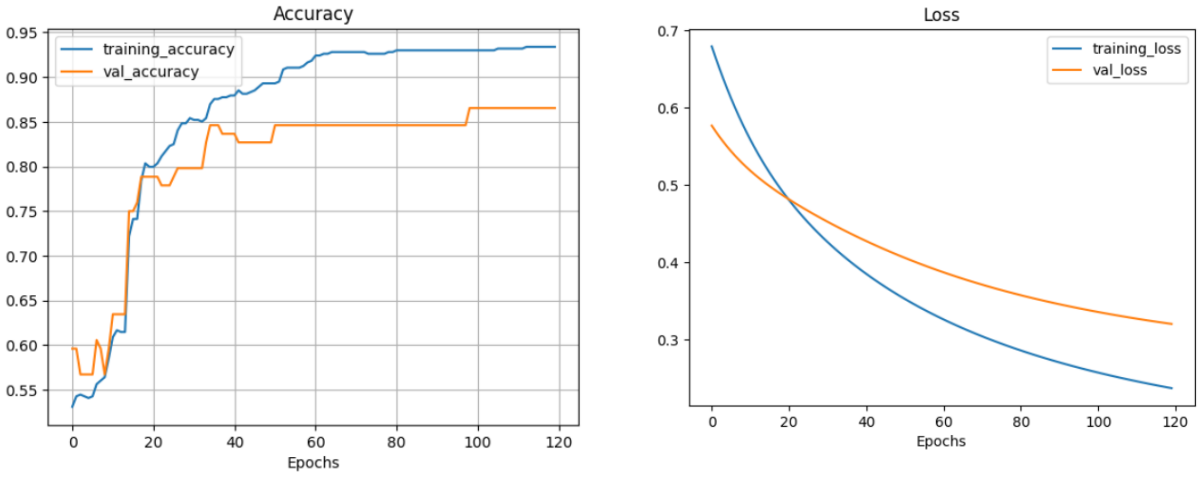

(a) (b)

**Figure 4.** Shows questionnaire dataset confusion matrices for (a) XGB (b) RF.




(a) (b)

**Figure 5.** Shows NN curves versus epochs for PIDD (a) Accuracy (b) Loss.




(a) (b)

**Figure 6.** Shows NN curves versus epochs for questionnaire dataset (a) Accuracy (b) Loss.

Random Forest yielded the highest accuracy among the proposed models because it is an ensemble technique integrates the predictions from multiple decision trees. A distinct subset of the data is used to train each tree, and the results are combined by majority voting for classification purpose. This lowers the chance of overfitting compared to a single decision tree. Moreover, this reduces the correlation between trees and ensures that

the model explores different patterns in the data. Furthermore, RF is less sensitive to noise and outliers. Random Forest tends to generalize well to unseen data because of its combination of randomness, ensemble learning, and averaging techniques.

## 5. Conclusions

The prediction of diabetes mellitus is considered a challenging medical research topic. This research involved development of a ML-based pipeline for classification T2DM based on two different datasets suffering from class imbalance and missing values. Consequently, our goal was met through applying RF, XGB and NN. Neural networks yielded an accuracy of 76.45% and 93.38% respectively for the first dataset and second dataset. To find the best model, RF achieved an overall accuracy of 80.51%, as well as AUC, precision, f1-score, recall and run time of 76%, 79.03%, 68.53%, 60.5% and 240.6 msec respectively on PIDD and accuracy of 97.11% as well as AUC, precision, f1-score, recall and run time of 97%,96.87%, 97.63%, 98.41% in 173.65 msec respectively on questionnaire dataset. While XGB yielded 79.65% accuracy as well as AUC, precision, f1-score, recall 76%,75%,68.45%,62.96% respectively in much less time of 80.5 msec on PIDD. Moreover, XGB results were 93.26%, 94%, 96.67%, 94.31%, 92.06% for accuracy, AUC, precision, f1-score and recall respectively in 49.28 msec. The proposed model yielded results which are superior to those of other studies in literature. Validation and testing were performed. In future work, applying different missing value imputation techniques close to real-life situations in addition to different class imbalance techniques. Furthermore, more machine learning and deep learning techniques will be applied on hybrid datasets.

## 6. References

[1] S. Du, V. Sullivan, M. Fang, L. Appel, E. Selvin, and C. Rebholz, "Ultra-processed food consumption and risk of diabetes: results from a population-based prospective cohort," *Diabetologia*, vol. 67, pp. 2225–2235, Feb. 2024, doi: 10.1007/s00125-024-06221-5.

[2] M. Benito, P. Marqués, C. Guillen, J. Burillo, B. Jiménez, and C. González-Blanco, "Insulin Resistance and Diabetes Mellitus in Alzheimer's Disease," *Cells*, vol. 10, Feb. 2021, doi: 10.3390/cells10051236.

[3] W. Jia *et al.*, "Standards of medical care for type 2 diabetes in China-0027," *Diabetes Metab Res Rev*, vol. 35, Feb. 2019, doi: 10.1002/dmrr.3158.

[4] H. Sun *et al.*, "IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Res Clin Pract*, vol. 183, p. 109119, 2022, doi: https://doi.org/10.1016/j.diabres.2021.109119.

[5] P. Saeedi *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Res Clin Pract*, vol. 157, p. 107843, 2019, doi: https://doi.org/10.1016/j.diabres.2019.107843.

[6] R. Siegel, A. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA Cancer J Clin*, vol. 74, Feb. 2024, doi: 10.3322/caac.21820.

[7] L. Guo and X. Xiao, "Guideline for the Management of Diabetes Mellitus in the Elderly in China (2024 Edition)," *AGING MEDICINE*, vol. 7, pp. 5–51, Feb. 2024, doi: 10.1002/agm2.12294.

[8] F. Alaa Khaleel and A. M. Al-Bakry, "Diagnosis of diabetes using machine learning algorithms," *Mater Today Proc*, vol. 80, pp. 3200–3203, 2023, doi: https://doi.org/10.1016/j.matpr.2021.07.196.

[9] M. E. Febrian, F. X. Ferdinan, G. P. Sendani, K. M. Suryanigrum, and R. Yunanda, "Diabetes prediction using supervised machine learning," *Procedia Comput Sci*, vol. 216, pp. 21–30, 2023, doi: https://doi.org/10.1016/j.procs.2022.12.107.

[10] K. Kangra and J. Singh, "Comparative analysis of predictive machine learning algorithms for diabetes mellitus," *Bulletin of Electrical Engineering and Informatics*, vol. 12, pp. 1728–1737, Feb. 2023, doi: 10.11591/eei.v12i3.4412.

[11] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting Diabetes Mellitus With Machine Learning Techniques," *Front Genet*, vol. 9, Feb. 2018, doi: 10.3389/fgene.2018.00515.

[12] A. Yahyaoui, J. Rasheed, A. Jamil, and M. Yesiltepe, "A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques," Feb. 2019. doi: 10.1109/UBMYK48245.2019.8965556.

[13] M. A. Sarwar, N. Kamal, W. Hamid, and M. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," Feb. 2018, pp. 1–6. doi: 10.23919/IConAC.2018.8748992.

[14] S. Rubaiat, M. M. Rahman, and M. K. Hasan, "Important Feature Selection & Accuracy Comparisons of Different Machine Learning Models for Early Diabetes Detection," Feb. 2018, pp. 1–6. doi: 10.1109/CIET.2018.8660831.

[15] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J Diabetes Metab Disord*, vol. 19, Feb. 2020, doi: 10.1007/s40200-020-00520-5.

[16] A. Badholia, "PREDICTIVE MODELLING AND ANALYTICS FOR DIABETES USING A MACHINE LEARNING APPROACH," *INFORMATION TECHNOLOGY IN INDUSTRY*, vol. 9, pp. 215–223, Feb. 2021, doi: 10.17762/itii.v9i1.121.

[17] R. Krishnamoorthi *et al.*, "A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques," *J Healthc Eng*, vol. 2022, pp. 1–10, Feb. 2022, doi: 10.1155/2022/1684017.

[18] C. Olisah, L. Smith, and M. Smith, "Diabetes Mellitus Prediction and Diagnosis from a Data Preprocessing and Machine Learning Perspective," *Comput Methods Programs Biomed*, vol. 220, p. 106773, Feb. 2022, doi: 10.1016/j.cmpb.2022.106773.

[19] R. Hegazii, E. Halim, and H. Mostafa, "A PROPOSED TECHNIQUE FOR BREAST CANCER PREDICTION AND CLASSIFICATION BASED ON MACHINE LEARNING," *European Chemical Bulletin*, vol. 12, pp. 7648–7656, Feb. 2023, doi: 10.48047/ecb/2023.12.8.619.

[20] M. Saar-Tsechansky and F. Provost, "Handling Missing Values when Applying Classification Models," *Journal of Machine Learning Research*, vol. 8, pp. 1625–1657, Jul. 2007.

[21] B. Mirzaei, F. Rahmati, and H. Nezamabadi-pour, "A score-based preprocessing technique for class imbalance problems," *Pattern Analysis and Applications*, vol. 25, Feb. 2022, doi: 10.1007/s10044-022-01084-1.

[22] M. Badawi *et al.*, "Skin Cancer Classification and Segmentation Using Deep Learning," *International Journal of Telecommunications*, vol. 04, pp. 1–23, Mar. 2024, doi: 10.21608/ijt.2024.280957.1045.

[23] M. Shams, Z. Tarek, and A. Elshewey, "A novel RFE-GRU model for diabetes classification using PIMA Indian dataset," *Sci Rep*, vol. 15, p. 982, Feb. 2025, doi: 10.1038/s41598-024-82420-9.

[24] S. Rahu, A. Ghulam, and F. Ali, "XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set," *Sci Rep*, vol. 12, Feb. 2022, doi: 10.1038/s41598-022-09484-3.

[25] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939785.

[26] N. A. Almansour *et al.*, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Comput Biol Med*, vol. 109, pp. 101–111, Jun. 2019, doi: 10.1016/j.compbiomed.2019.04.017.

[27] A. Dutta *et al.*, "Early Prediction of Diabetes Using an Ensemble of Machine Learning Models," *Int J Environ Res Public Health*, vol. 19, Feb. 2022, doi: 10.3390/ijerph191912378.