Diabetes Mellitus Disease Prediction and Type Classification Involving Predictive Modeling Using Machine Learning Techniques and Classifiers

B. Shamreen Ahamed, Meenakshi S. Arya, S. K. B. Sangeetha, and Nancy V. Auxilia Osvin

1SRM Institute of Science and Technology, Vadapalani Campus, Vadapalani, Chennai, India
2MIT World Peace University, Pune, Maharashtra, India

Correspondence should be addressed to B. Shamreen Ahamed; shamu1502@gmail.com

Received 14 April 2022; Revised 16 November 2022; Accepted 6 December 2022; Published 30 December 2022

1.Introduction

DM is one of the most commonly found diseases in the world today. The people affected by the disease are of varied age groups starting from newborn to the elderly. In reference to the Federation of International Diabetes, approximately 451 million people around the world have been affected by Diabetes in the year 2017 [1]. However, the people going to be affected has been predicted to be increasing and might become double the number by the year 2045 as per research. To avoid and overcome the problem, the Prediction of Diabetes is essential [2]. By predicting Diabetes in the earlier stages, the number of people being affected can be reduced, and also timely medication will eliminate the disease earlier than expected [3].

Diabetes is an ailment caused when the sugar level in the body is more than normal. When left untreated, it becomes severe thus resulting in life long insulin support for the body. The disease further leads the patient into other complications related to heart diseases, liver diseases, kidney failure, eye disorder, etc. Diabetes Mellitus is categorized into different types based on certain constraints. The primary types are GDM, Prediabetes, T1DM, and T2DM. These categories can be further subcategorized into LADA, MODY, Neonatal Diabetes, Type-3 Diabetes, Double Diabetes, Wolfram Syndrome, and Alstrom Syndrome [4].
Prediabetes is also called impaired glucose tolerant. It is a condition where the glucose level is high, however not as high as T2DM. If it is not taken care of or treated in the earlier stage, it will further lead to T2DM Diabetes in the future. It can also lead to a condition called metabolic syndrome which is a combination of three or more of the following: low levels of HDL, High BP, High triglycerides, large waist size, and high blood sugar levels. The prediabetes tests include A1C Test, FPG test, and OGTT Test [5].

Type-1 DM (T1DM) is also known as Juvenile-Diabetes (JDDM). It is a dependent on insulin where the insulin release cell is damaged by the immune system; removing the production of insulin in the body. It commonly occurs in adolescence age and its complications include skin problems, cardiovascular disease, poor blood flow, gum disease, nerve damage, pregnancy problems, retinopathy, and kidney damage [6]. The test for Type-1 are blood test (auto antibodies) and urine test (ketones). The major risk factors are age below 20 years and a history of diabetes in the family.

Type-2 Diabetes (T2DM) occurs in older individuals and is milder than T1DM. It occurs when the body does not produce enough insulin or resists insulin production by the pancreas [7]. It is also called as adult-onset Diabetes. The common causes are obesity, inactive lifestyle, nervous and immune system weakening, etc. The other complication is: eyes, nerves, kidneys, heart disease, and stroke. The diagnosis can be done based on A1C Test, FPG Test (Fast Plasma Glucose), and RPG Test (Random Plasma Glucose). The other tests include Oral Glucose Tolerance test (OGTT) and Glucose Challenge test [8].

Gestational Diabetes (GDM) is one that occurs during pregnancy and usually occurs during the second half of pregnancy. When a body does not produce enough insulin or stops insulin usage, the blood sugar level will rise that further leads to gestational pregnancy [9]. Some of the complications are being overweight, physically inactive, family history, polycystic ovary syndrome, etc. The tests include Glucose Challenge Test and Glucose Tolerance Test. Some of the risk factors are Cesarean birth, Hypoglycemia, Preeclampsia, and Type-2 diabetes after delivery. It usually disappears postdelivery of the individual but the risk to develop T2DM exists if not taken care of [10].

Machine Learning is a subclass of Artificial Intelligence. It aims at creating computer systems that is used to discover patterns in data training to perform classification and prediction for new data. Machine learning is used to combine tools from statistics, data mining, and optimization for model generation. It focuses on finding an accurate representation of the knowledge automatically extracted from the data [11]. There are many algorithms in Machine Learning that can be used for prediction. Some of them include Support Vector Machine (SVM), Random Forest (RF), XGBoost (XGB), Light Gradient Boosting Machine (LGBM), Decision Tree (DT), Gradient Boosting Machine (GBM), Naive Bayes (NB), Logistic Regression (LoR), and Linear Regression (LiR) [12].

The paper is organized in the following way: Section 2 explains the work related that uses ML algorithms for determining Diabetes Mellitus (DM) disease. Section 3 characterizes the overview of several ML algorithms that can be used for the proposed architecture. Section 4 provides a discussion about the architecture, the dataset used, features, preprocessing, etc. Section 5 states the implementation and Section 6 denotes the study conclusion.

2. Related Works

The works related to ML algorithms for Diabetes prediction are commonly used in the medical industry. ML techniques have been used by many researchers to predict DM in order to obtain the best and most accurate results.

Zou et al. [13], have used ML algorithms and techniques for predicting DM disease. The classifiers used are DT, RF, and Neural Network. The dataset used includes PIMA and Hospital physical examination data from Luzhou, China. The Pima dataset consists of 9 attributes while the examination dataset consists of 14 attributes. The tool used is WEKA. The accuracy reported by the classifiers used gives the highest accuracy of 80.8% for hospital data and 77% for the pima dataset by using the Random Forest Classifier. However, the accuracy obtained can be further improved with other Classifiers and Techniques.

Zarkogianni et al. [14], have used the concept of hybrid wavelet neural networks (HWNNs) and self organizing maps (SOMs) constitute. The dataset is collected from 560 patients who are affected by both cardiovascular disease (CVD) and Diabetes (DM) is chosen. The highest AUC curve gives 71.48%. The proposed method is superior to Binomial Linear Regression (BLR) by applying techniques to produce reliable CVD risk scores. Out of 560 patients, 41 patients had DM also had nonfatal CVD. Out of 41, 4 experienced stroke and the others experienced coronary heart disease. The shortcomings of the paper involves the need to improve the accuracy percentage and can also focus on one particular dataset than a hybrid model.

Alić et al. [15], have classified diabetes and cardiovascular disease (CVD) using Artificial Neural Network (ANN) and Bayesian Network (BN). The ANN used is a multilayer neural-network with Levenberg–Marquardt learning algorithm. The BN is Naive Bayes which provides the highest accuracy of DM and CVD as 99.51% and 97.92%. The accuracy using ANN for Diabetes disease is 72.7% and 99% and for CVD is 80% and 95.91%. The accuracy using BN for Diabetes disease is 71% and 99.51% and for CVD is 78% and 97.92%. The ANN uses the sigmoid transfer function and BN uses probability theory.

Sneha and Gangil [16], focuses on the detection of Diabetes in the early stages using optimal feature selection. The algorithms used are DT and RF with a specificity of 98.20% and 98%. Naive Bayes states an accuracy of 82.30%. The research carried out by the authors also generalizes the features to increase the accuracy of classification. A total of 5 algorithms are compared: SVM, RF, NB, DT, and KNN. It uses a rapid-miner data mining tool. The analysis of feature in the dataset is carried out. The highest accuracy is given by DecisionTree and RandomForest as mentioned above. The accuracy of SVM is 77.73% and 73.48% in the existing method and 77% for SVM and 82.30% for NB in the
proposed method. The future scope of the research is to improve the metrics value.

Tafa et al. [17] have used the algorithms SVM and Naive Bayes and proposed an integrated model for Diabetes prediction. Three different datasets have been used on the model. The data has been collected from Kosovo. The dataset consists of eight attributes. The data of 402 patients were taken where 80 patients were affected by type 2 diabetes. Some attributes such as diet and physical activity have not been used commonly in other studies which is the uniqueness. The data for training and testing has been divided equally. The proposed model provides an accuracy of 97.6% by using combined algorithms. However, SVM provides an accuracy of 95.52% and Naive Bayes provides an accuracy of 94.52% when implemented separately. The future scope can involve running the model with other ML algorithms for analysis and testing matrices.

Mercaldo et al. [18] have used six ML classifiers that include Multilayer Perceptron, J48, JRip, Hoeffding Tree, Random Forest, and BayesNet. The dataset used is PIMA. The main algorithms used are BestFirst and GreedyStepwise. They are used to represent the attributes in order to increase the performance classification. Four attributes are taken namely diabetes pedigree function, body mass index, age, and plasma glucose concentration. A 10 fold-cross validation is used for the dataset. The result obtained are precision value 0.757, recall value 0.762, and F-measure value 0.759 by using the Hoeffding Tree algorithm. The algorithm used on the model can be varied and the parameters can also be modified for future work and accuracy improvisation.

Kandhasamy and Balamurali [19] have used multiple classifiers J48, SVM, RF, and K-Nearest Neighbors (KNNs). The dataset is taken from the UCI repository. The matrices compared are specificity, sensitivity, and accuracy. The classification was performed on the dataset by preprocessing and without preprocessing using 5-fold cross validation. The results show that the decision tree J48 classifier has the highest accuracy of 73.82% without preprocessing and the classifiers KNN (k = 1) and Random Forest produced the highest accuracy rate of 100% after preprocessing process.

Annamalai and Nedunchelian [20], have used the OWDANN Algorithm for Diabetes Mellitus Prediction. The proposed system consists of 2 phases, disease prediction, and severity level estimation. The preprocessing is carried out on the PIMA dataset. Features are extracted from preprocessing dataset and classification is done using OWDANN. The severity level estimation phase uses diabetes positive dataset for preprocessing and predicted using GDHC. The accuracy obtained is 98.97%, sensitivity of 94.98%, and specificity of 95.62%.

Davitt et al. [21], have used metabolic disorder characterized by elevated blood glucose concentration to either: insulin resistance, less insulin secretion, or both. The etiologic Diabetes classification are T1DM, T2DM and GDM. The Diabetes classification includes genetic effects of beta-cell function, genetic defects in insulin action, drug or chemical induced, disease of the exocrine pancreas, endocrinopathies, post-transplant, and genetic syndrome.

Ahmad and Arya [22], have used RR-interval-signals known as heart rate viability (HRV) signals can be used for noninvasive Diabetes detection. An explanation on the methodology of how the classification of diabetics and normal HRV Signals using deep learning architectures is explained. An employment of LSTM, CNN, and its combinations for extracting complex temporal dynamic features of the input HRV data is carried out. The various features are passed onto the SVM. An improvement in the performance of 0.03% and 0.06% in CNN and CNN-LSTM Architecture has been achieved.

3. Dataset

In this section, the description of the dataset is done. The dataset taken is the PIMA Indian dataset that is taken form UCI Repository and a survey dataset that was collected and curated. The details about the dataset is given.

3.1. PIMA Dataset. In the research work related to Diabetes Mellitus, PIMA Dataset has been commonly used and studied on by many researchers. The dataset is available in the UCI Repository [1]. The dataset consists of 9 attributes: pregnancy, glucose, blood pressure, insulin, skin thickness, BMI, diabetes pedigree function, age, and outcome. The total number of instances are 768.

A sample of the dataset is given below in Figure 1.

3.2. Survey. The second dataset that is used is a Clinical Survey dataset collected from a Diagnostic Center, Srinagar. It consists of 734 instances with attributes: age, fasting, post_pran, waist, BMI, systolic, diastolic, Hba1c, gender, history, and class.

A sample of the dataset is given below in Figure 2.

The goal of both the dataset used is to identify and utilize the factors that are more predominant in the occurrence of the disease in a person. The PIMA dataset consists of attributes that are required for the Prediction of Diabetes and the Clinical Survey dataset that is taken to identify and classify diabetes as either Prediabetes or Diabetes.

4. Theoretical Concepts and Algorithms

The theoretical concepts of the Machine Learning Classifiers used is explained as follows.

4.1. Machine Learning Algorithms

4.1.1. Gradient Boosting. A Gradient boosting classifier is a combination of many learners (weak) formed into a predictive-model typically as Decision Trees. The number of trees is based on a number of values in the dataset used. It is mainly used when the bias error in the model needs to be decreased. A gradient-descent-technique is chosen to obtain values of the coefficients [23].

In order to obtain the value of the coefficient, the loss function used needs to be calculated. It is calculated using \((y_1 - y_1')^2\), where \(y_1\) is the actually calculated value and...
\( y_{1}' \) is the finally value that is predicted by the model. So \( y_{1}' \) is replaced with \( G_n(X) \) which represents the target value \[24\]. It is mathematically given as follows:

\[
G_{n+1}(X) = G_n(X) + \gamma_n H_1(x, e_n),
\]

\[
L_1 = (y_1 - y_{1}')^2,
\]

\[
L_1 = (Y - G_n(x))^2.
\]

4.1.2. **Light_Gradient_Boosting_Machine** (LGBM). The evaluation of LGBM performance is represented to be high-performance and is considered as “gradient boosting framework” based on Decision_Tree algorithm. It is an advanced version of the Gradient Boosting Framework. It is majorly used for ranking and classifying. It divides the tree leafwise with the best-fit value. It can be calculated using many improvement techniques for data and can be given by variance evaluation after diving the values \[25\]. It can be given by the following equation:

\[
Y_1 = \text{Base Tree}(X) - lr1 \ast \text{Tree 1}(X) - lr1 \ast \text{Tree 2}(X).
\]

4.1.3. **XGBoost** (XGB). A supervised regression model named XGBoost is used for identifying the validity of the objective function and base learners. The concept of Ensemble learning is used to combine independent weak learning models for prediction. XGBoost is one of the ensemble learning methods. It is given as follows:

\[
\text{obj}(\theta) = \sum_{i=1}^{n} l(y_i - \hat{y}_i) + \sum_{j=1}^{J} \delta(f_j),
\]
Where \( f_j \) denotes predictive value from the \( j \)th tree [27]. The MSE (mean squared error) is given as follows:

\[
\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i^j)^2,
\]

(4)

### 4.1.4. Decision Tree (DT)

The DT Entropy is generated as follows: A node \( k \) is taken and \( J \) class labels are identified. The value of \( J \) ranges from 1 to \( J \). It is given mathematically as follows:

\[
\text{Entropy}(k) = - \sum_{j=1}^{J} p(j | k) \log_2 (j | k).
\]

(5)

### 4.1.5. Random_Forest (RF)

The RF consolidates the output values or outcomes of a number of Decision Tree together in order to obtain one result. The DT considered are taken as a base_row sampling technique and column sampling technique. The quantity of base learners is improved depending on the inputs and the variance is reduced to increase the accuracy [29]. It is taken into account as one of the important bagging methodologies.

\[
\text{Random Forest} = \text{DT (base learner)} + \text{begging (Row sampling with replacement)}
\]

+ feature bagging (column sampling)

+ aggregation (mean/median, majority vote).

(6)

### 4.1.6. Naive_Bayes (NB)

NB is one of the classification methods that uses conditional probability values to divide the data using the algorithm. It is also used to detect the behavior of the various patients involved. It is majorly used to implement large dataset. It is a collaborative classification model involving Logistic Regression for patients data classification into different groups. It is good for predictions involving real time, multiclass, recommendation system, text based classification, and sentiment analysis [30].

The Bayesian Formula for calculating the Naive Bayes Algorithm is as follows:

\[
\Pr(A | B) = \Pr(B | A) \Pr(A) \Pr(B) \text{Predictor_Prior_Probability,}
\]

where \( \Pr(A | B) = \text{Posterior\_Probability} \), \( \Pr(B | A) = \text{Likelihood\_Probability} \), \( \Pr(A) = \text{Class\_Prior\_Probability} \), \( \Pr(B) = \text{Likelihood\_Probability} \) [31].

Many Machine Learning methods and techniques can be tested and used along with classifiers for Diabetes disease prediction. However, for the dataset used, the best suited classifiers are considered Gradient Boosting Classifiers (GBM, LGBM, and XGB) from Table 6 and Decision Tree based on the Simulation mechanism earlier used. However, other classifiers such as Random Forest, Naive Bayes, and Support Vector Machine are also considered for final accuracy percentage analysis [32].

### 4.2. Correlation Matrix

A correlation matrix in Machine Learning is used to summarize the attributes used the dataset and to identify the attributes with the utmost importance for identification and consideration during predictive analysis.
4.3.1. One Hot Encoding. The one hot encoding technique is used in the paper to transform categorical values into numeric data. The categorical variables present in the dataset are initially encoded and considered as ordinal, followed by representing integer value as binary value as either 0 or 1. The binary variables are also called as “dummy variables” in Machine Learning [33].

The dataset after one hot encoding appears as shown in Figure 8.

The one hot encoding procedure is followed by data implementation on the predictive model built using the Machine Learning Classifiers. The predictive model consists of the dataset and the parameters suitable according to the algorithms used. Each algorithm in Machine Learning consists of some specific parameters necessary for effective utilization. The parameters are fine-tuned using values that produce the highest accuracy percentage based on a particular model developed [34].

Some of the commonly used parameters are learning_rate, max_depth, n_estimators, min_samples_split, min_samples_leaf, max_features, subsample, random_state, etc.

4.3.2. Data Augmentation. Data Augmentation is a technique that is used to increase the amount of data that are not uniform or decrease the quantity of data to remove excess. The amount of data can be increased by adding duplicate values. This technique is called Over-Sampling. The quantity of data can be reduced and this technique is called Down-Sampling.

In the paper, the concept of Over-Sampling is being used to increase the balance of the class values. The values before and after sampling is shown in Figures 9 and 10.

5. Architecture

The architecture as shown in Figure 11, consists of the working flow of the procedure. Initially, the data is chosen from the various databases available and the most suitable dataset is chosen. The PIMA dataset is taken for the study from UCI Repository. The dataset consists of 768 entries and 9 attributes. The data chosen is preprocessed, followed by procedures of feature selection and extraction is carried out. The data is further preprocessed using EDA until all of the defects are rectified. The dataset is then cleaned and set for
training and testing procedures. The dataset is divided into training_data (TrData) and testing_data (TeData).

The various Machine Learning classifiers are then compared and the best suited classifier for the dataset is chosen. The parameters are fine tuned on the predictive model developed and the performance matrices are calculated. After the prediction procedures are carried out, another dataset is taken to predict the type classification of Diabetes. The dataset for classifying the type of Diabetes as prediabetes and normal is taken from a survey conducted in a laboratory. Finally, the algorithm producing the highest accuracy percentage is selected.

6. Results and Discussion

The prediction for Diabetes Mellitus is done using the model built using the dataset PIMA dataset initially and then the highest accuracy-producing algorithms are chosen and further incorporated for Type Classification.

The accuracy obtained for the various classifiers are given in Table 1. The abbreviation for the classifiers are LR, XGB, GB, DT, ET, RF, and LGBM.

Figure 12 denotes the bar graph of the accuracy percentage obtained while using Classifiers DT, LR, RF, XGB, GBM, ET, and LGBM.

Table 1, denotes that LGBM and RF produce the highest accuracy. XGB also produces high accuracy. Therefore
another predictive mechanism for the classifiers RF, LGBM, and GBM are conducted for the dataset PIMA.

The algorithms producing the highest accuracy for the PIMA Indian dataset are Random Forest, Light Gradient Boosting Machine, and Gradient Boosting Machine. The

Table 1: Comparison of all classifiers.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>PIMA Indian Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td></td>
</tr>
<tr>
<td>XGBoost</td>
<td></td>
</tr>
<tr>
<td>Gradient Boosting Machine</td>
<td></td>
</tr>
<tr>
<td>Decision Tree</td>
<td></td>
</tr>
<tr>
<td>Extra Trees</td>
<td></td>
</tr>
<tr>
<td>Random Forest</td>
<td></td>
</tr>
<tr>
<td>Light Gradient Boosting Machine</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>75.20%</td>
</tr>
<tr>
<td>83.30%</td>
<td></td>
</tr>
<tr>
<td>94.1%</td>
<td></td>
</tr>
<tr>
<td>94.40%</td>
<td></td>
</tr>
<tr>
<td>94.60%</td>
<td></td>
</tr>
<tr>
<td>94.80%</td>
<td></td>
</tr>
<tr>
<td>95.20%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 11: Architecture.

Figure 12: Accuracy percentage.

Figure 13: Gradient boosting classifier.

Accuracy of the GBM on test set : 0.929

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.98</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td>1.0</td>
<td>0.83</td>
<td>0.96</td>
<td>0.89</td>
</tr>
</tbody>
</table>

accuracy

|     |         |          | 0.93    | 154     |

| macro avg | 0.91     | 0.94     | 0.92    | 154     |
| weighted avg | 0.94   | 0.93     | 0.93    | 154     |

Figure 14: Random forest classifier.

Accuracy of the RANDOM FOREST on test set : 0.909

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.94</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>1.0</td>
<td>0.84</td>
<td>0.87</td>
<td>0.85</td>
</tr>
</tbody>
</table>

accuracy

|     |         |          | 0.91    | 154     |

| macro avg | 0.89     | 0.90     | 0.89    | 154     |
| weighted avg | 0.91   | 0.91     | 0.91    | 154     |
Accuracy of the LGBM on test set: 0.922

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>f1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.96</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>1.0</td>
<td>0.84</td>
<td>0.91</td>
<td>0.88</td>
</tr>
</tbody>
</table>

accuracy
macro avg | 0.92    | 154
weighted avg | 0.93    | 154

Figure 15: Light gradient boosting classifier.

Table 2: Results after data augmentation.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GBM</td>
<td>92.9</td>
<td>0.98</td>
<td>0.83</td>
<td>0.92</td>
</tr>
<tr>
<td>RF</td>
<td>90.9</td>
<td>0.94</td>
<td>0.84</td>
<td>0.93</td>
</tr>
<tr>
<td>LGBM</td>
<td>92.2</td>
<td>0.96</td>
<td>0.84</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Figure 16: Gradient boosting machine graph.

Figure 17: Random forest graph.

Figure 18: Light gradient boosting machine graph.
The technique of Data Augmentation is implemented in the 3 algorithms mentioned and the accuracy is obtained. The Data Augmentation and Sampling results for GBM, RF, and LGBM are given below in Figures 13–15. The accuracy and other performance matrices are cumulatively given below in Table 2, Figures 16–18.

The Type Classification of Diabetes disease is further classified using another dataset collected from the survey as shown in Figure 19. The output obtained for type classification is given in Figure 20.

The label column in the above table denotes if the value is prediabetic, diabetic, or normal. The value 0 indicates Normal, 1 indicates Prediabetes and 2 indicates Diabetes. When the values obtained from the expected output and predicted output are the same, then the accuracy obtained from the above-given calculation indicates 100% accuracy using the model built and trained for type classification procedures.

The research aims at improving the Prediction of Diabetes disease among people. The Classification of Diabetes is necessary to estimate the level of severity and to consider precautions in future for the health awareness in today’s healthcare industry.

7. Conclusion and Future Scope

From the above-given study, it can be concluded that the LGBM algorithm provides higher accuracy at a maximum when compared with RF and GB classifiers. Therefore, the LGBM algorithm is well suited for the PIMA dataset used in the study.

The LGBM algorithm varies from RF and GB in the following ways: the parameters used in LGBM is different than GB and RF. The parameter tuning varies with each algorithm and the model is built based on the classifier used. Therefore in this paper, a predictive model is built using the LGBM algorithm and the accuracy is obtained as shown in Table 1 for the dataset used. In addition to prediction procedures, the Type Classification of the type of Diabetes is also predicted and calculated.

The Diabetes Mellitus disease prediction can further be improvised by enhancing the dataset using other advanced methodologies like Transformer based learning. The attributes used can also be used in different combinations for identification. The classifiers used can be fine-tuned more to predict the disease with higher accuracy and the probability of occurrence of the disease can be calculated. This will further improve the accuracy percentage and deliver a more profound model to predict Diabetes Mellitus Disease among affected people.

Data Availability

The dataset used is taken from UCI Repository and the links are given as follows: (i) https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. (ii) Survey data.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

Authors’ Contributions

All authors equally contributed to the study.

References


