Hindawi
Applied Computational Intelligence and Soft Computing
Volume 2022, Article ID 7899364, 11 pages
https://doi.org/10.1155/2022/7899364

Research Article

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

B. Shamreen Ahamed,[1] **Meenakshi S. Arya**,[2] **S. K. B. Sangeetha**,[1] **and Nancy V. Auxilia Osvin**[1]

[1]*SRM Institute of Science and Technology, Vadapalani Campus, Vadapalani, Chennai, India*
[2]*MIT 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

Academic Editor: Babek Erdebilli (B. D. Rouyendegh)

The Diabetes-Mellitus (DM) disease is considered a persistent ailment that is triggered by excessive sugar levels in the blood of a person. It gives rise to severe health complications when left untreated and can also give rise to related diseases such as cardiac attack, nervous damage, foot problems, liver and kidney damage, and eye problems. These problems are caused by a series of factors interrelated to one another such as age, gender, family history, BMI, and Blood Glucose. Various Machine-Learning (ML) algorithms are being used in order to predict and detect the disease to avoid further complications of health. The Diabetes prediction process can be further improvised by identifying the type a person is being affected by and the probability of the occurrence of the related diseases. In order to perform the mentioned task, two types of the dataset are used in the study, namely, PIMA and a clinical survey dataset. Various ML algorithms such as Random Forest, Light Gradient Boosting Machine, Gradient Boosting Machine, Support Vector Machine, Decision Tree, and XGBoost are being used. The performance metrics used are accuracy, precision, recall, specificity, and sensitivity. Techniques such as Data Augmentation and Sampling are used. In comparison with the research conducted previously, the paper focuses on improvisation of the accuracy with a percentage of 95.20 using the LGBM Classifier, and Diabetes is also classified as Prediabetes or Diabetes using many Classification mechanisms.

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 who 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, RandomForest, 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 diabetic 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 []. 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.

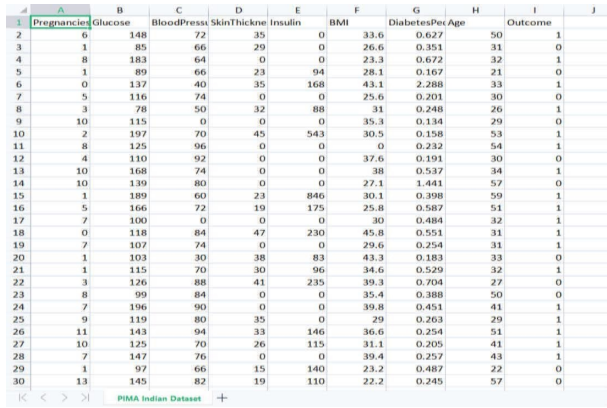
4. Theoretical Concepts and Algorithms

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 $(y1 - y1')^2$, where $y1$ is the actually calculated value and



	A	B	C	D	E	F	G	H	I	J
1	Pregnancies	Glucose	BloodPressu	SkinThickne	Insulin	BMI	DiabetesPec	Age	Outcome	
2	6	148	72	35	0	33.6	0.627	50	1	
3	1	85	66	29	0	26.6	0.351	31	0	
4	8	183	64	0	0	23.3	0.672	32	1	
5	1	89	66	23	94	28.1	0.167	21	0	
6	0	137	40	35	168	43.1	2.288	33	1	
7	5	116	74	0	0	25.6	0.201	30	0	
8	3	78	50	32	88	31	0.248	26	1	
9	10	115	0	0	0	35.3	0.134	29	0	
10	2	197	70	45	543	30.5	0.158	53	1	
11	8	125	96	0	0	0	0.232	54	1	
12	4	110	92	0	0	37.6	0.191	30	0	
13	10	168	74	0	0	38	0.537	34	1	
14	10	139	80	0	0	27.1	1.441	57	0	
15	1	189	60	23	846	30.1	0.398	59	1	
16	5	166	72	19	175	25.8	0.587	51	1	
17	7	100	0	0	0	30	0.484	32	1	
18	0	118	84	47	230	45.8	0.551	31	1	
19	7	107	74	0	0	29.6	0.254	31	1	
20	1	103	30	38	83	43.3	0.183	33	0	
21	1	115	70	30	96	34.6	0.529	32	1	
22	3	126	88	41	235	39.3	0.704	27	0	
23	8	99	84	0	0	35.4	0.388	50	0	
24	7	196	90	0	0	39.8	0.451	41	1	
25	9	119	80	35	0	29	0.263	29	1	
26	11	143	94	33	146	36.6	0.254	51	1	
27	10	125	70	26	115	31.1	0.205	41	1	
28	7	147	76	0	0	39.4	0.257	43	1	
29	1	97	66	15	140	23.2	0.487	22	0	
30	13	145	82	19	110	22.2	0.245	57	0	

PIMA Indian Dataset

Figure 1: PIMA Indian diabetes.

	A	B	C	D	E	F	G	H	I	J	K
1	Age	Fasting	Post_pran	Waist	BMI	Systolic	Diastolic	Hba1c	Gender	History	Class
2	59	147	207	36.5	30.1	135	85	7.2	F	1	yes
3	48	125	140	34.1	25.5	120	80	6.4	M	1	no
4	70	156	210	35.7	30.5	137	87	6.9	F	1	yes
5	25	115	141	34.4	22.6	125	80	5.7	M	1	no
6	50	120	140	34.2	23.1	130	85	5.8	M	1	no
7	25	132	190	34.7	27.7	128	83	5.9	M	1	yes
8	36	115	135	33.5	22.1	120	85	5.6	F	1	no
9	66	160	230	36.3	26.2	135	90	7.7	F	1	yes
10	60	146	203	36.5	32.1	125	85	7.1	F	1	yes
11	29	112	142	35.1	30.1	121	87	5.9	M	0	no
12	56	157	211	34.3	28.1	125	90	6.6	F	1	yes
13	28	112	140	32.2	26.8	100	70	5.2	M	1	no
14	53	116	145	33.9	23.7	130	85	5.8	F	1	no
15	50	167	219	34.2	26.2	140	92	7.1	M	1	yes
16	61	123	186	34.6	27.8	145	95	5.9	F	1	yes
17	34	125	144	34.8	26.5	140	85	5.8	F	1	no
18	47	180	240	36.8	30.3	155	96	7.8	M	1	yes
19	31	88	129	32.1	20.1	100	77	4.1	F	1	no
20	68	159	209	36.1	27.5	140	90	6.9	F	1	yes
21	38	110	140	35.5	25.4	125	82	5.7	F	0	no
22	41	107	131	34.4	24.8	120	80	5.6	M	1	no
23	35	101	123	33.1	21.1	120	75	5.4	M	1	no
24	25	95	125	32.1	24.3	121	83	5.6	F	0	no
25	25	126	150	32.6	26.1	123	87	5.9	M	0	no
26	61	170	240	38.1	32.2	155	95	7.3	F	1	yes
27	31	127	159	34.3	29.7	130	83	6.2	M	1	yes
28	30	129	179	34.5	27.2	130	86	5.9	F	1	yes
29	42	125	156	37.2	31.9	137	85	6.2	F	1	yes
30	30	98	133	32.7	21.9	115	80	5.2	F	1	no

Figure 2: Survey dataset.

$y1'$ is the finally value that is predicted by the model. So $y1'$ is replaced with $G_n(X)$ which represents the target value [24]. It is mathematically given as follows:

$$\begin{aligned} G_{n+1}(X) &= G_n(X) + \gamma_n H1(x, e_n), \\ L1 &= (y1 - y1')^2, \\ L1 &= (Y - G_n(x))^2. \end{aligned} \tag{1}$$

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:

$$Y1 = \text{Base Tree}(X1) - lr1 * \text{Tree 1}(X1) - lr1 * \text{Tree 2}(X1). \tag{2}$$

The value determines the way in which the DT algorithm can be used to split the data and implement the values. The equation represents the number of trees that can be used in the model depending on the count of instances in the dataset used. When compared with GB, LGBM is comparatively faster and the parameters used are different that can further increase or decrease the efficiency [26].

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:

$$\text{obj}(\theta) = \sum_{1}^{n} l(y_i - \hat{y}_i) + \sum_{j=1}^{j} \delta(f_j), \tag{3}$$

Where f_j denotes predictive value from the jth tree [27]. The MSE (mean squared error) is given as follows:

$$\text{MSE} = \frac{\sum_{i=1}^{n} \left(y_i - y_i^p\right)^2}{n}, \tag{4}$$

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:

$$\text{Entropy}(k) = -\sum_{j=1}^{J} p(j \mid k) \log_2 (j \mid k). \tag{5}$$

LGBM can be used in 2 methods, namely, GBDT (GradientBoostingDecisionTree) and GOSS (Gradientbasedone-sidedsampling). Treewise method is used to provide the best fit, and other boosting algorithms use the depthwise method to divide. It provides better results when compared with the other existing boosting algorithms [28].

4.1.5. Random_Forest (RF). The RF consolidates the outvalues 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.

$$\begin{aligned} &\text{Random Forest} \\ &= \text{D}T\,(\text{base learner}) + \text{begging}\,(\text{Row sampling with replacement}) \\ &\quad + \text{feature bagging}\,(\text{column sampling}) \\ &\quad + \text{aggregetion}\,(\text{mean/median, majority vote}). \end{aligned} \tag{6}$$

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:

$$\Pr(A \mid B) = \frac{\Pr(B \mid A)\Pr(A)}{\Pr(B)}, \tag{7}$$

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

Many MachineLearning 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 . . .. 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 summerize the attributes used the dataset and to identify the attributes with the atmost importance for identification and consideration during predictive analysis. The patterns used for identification are used for Decision making of the process. The matrix is represented as cells and each cell is used to calculate the relation and correlation between 2 attributes. The visualization of the result for the PIMA Dataset is given in Figure 3.

4.3. Data Preprocessing. In the study, Data Preprocessing is done which is a technique used in Machine Learning that is used to organize and clean the data for further processing and analysis. The transformation and encoding of data is carried out with processes involving: data integration, data transformation, data reduction, and data cleaning. The importance of data preprocessing is for accuracy and precisions in values of data an for easier interpretation of data features by the algorithms.

The importance of features is given by Feature extraction and Feature Selection. The necessary features for the process is selected from all the features available in the dataset. This is given in Figure 4.

The data processing reduces duplicate values and outliers of the data with inconsistent data points. The data quality is reassured after the steps of data profiling, data cleaning, and data monitoring. The paper involves data preprocessing steps as shown in Figures 5 and 6.

Following the data preprocessing, the class values are introduced with a new set of attributes containing the previously used attributes. The complete dataset after the above-given process appears as follows in Figure 7.

The dataset as shown in Figure 7 consists of categorical values and in order to perform the operation using column value categorization, the concept of one hot encoding is carried out.

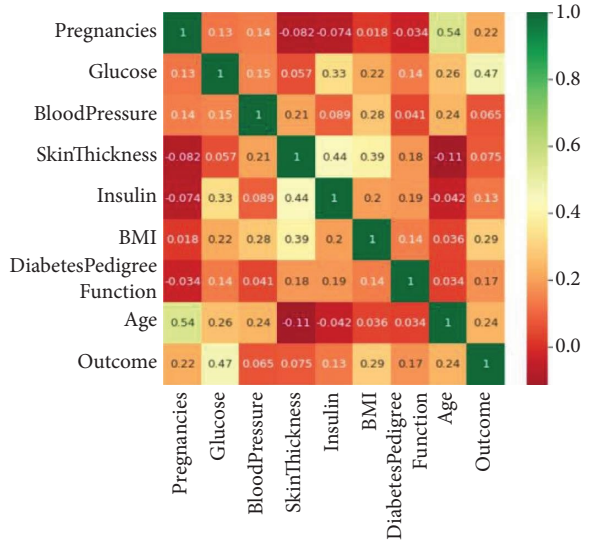


Figure 3: Correlation matrix.

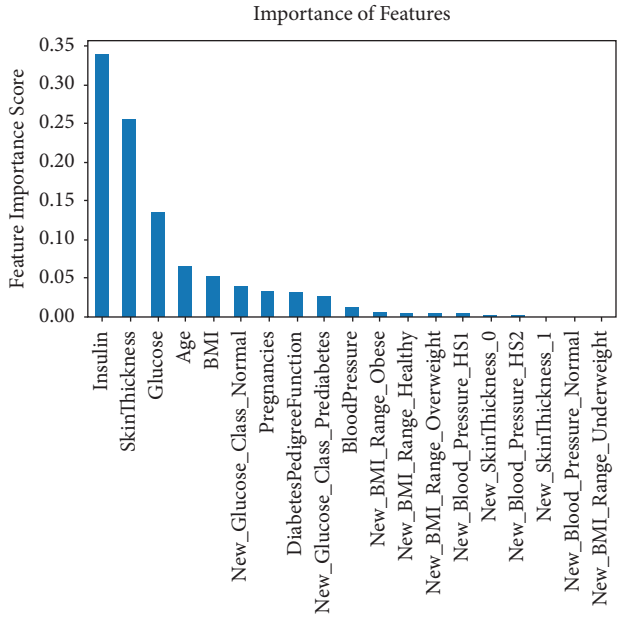


Figure 4: Importance of features.

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 MachineLearning 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

```
[ ]  df.isnull().sum()

Pregnancies                 0
Glucose                     5
BloodPressure              35
SkinThickness             227
Insulin                   374
BMI                        11
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 5: Before data preprocessing.

```
[ ]  df.isnull().sum()

Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Figure 6: After data preprocessing.

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

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	New_Glucose_Class	New_BMI_Range	New_Blood_Pressure	BMI	DiabetesPedigreeFunction	Age	Outcome	New_SkinThickness
0	6.0	148.0	72.0	35.0	169.5	Prediabetes	Obese	Normal	33.6	0.627	50.0	1.0	0
1	1.0	85.0	66.0	29.0	102.5	Normal	Overweight	Normal	26.6	0.351	31.0	0.0	0
2	8.0	183.0	64.0	32.0	169.5	Prediabetes	Healty	Normal	23.3	0.672	32.0	1.0	0
3	1.0	89.0	66.0	23.0	94.0	Normal	Overweight	Normal	28.1	0.167	21.0	0.0	0
4	0.0	137.0	40.0	35.0	168.0	Normal	Obese	Normal	43.1	1.949	33.0	1.0	0
5	5.0	116.0	74.0	27.0	102.5	Normal	Overweight	Normal	25.6	0.201	30.0	0.0	0
6	3.0	78.0	50.0	32.0	88.0	Normal	Obese	Normal	31.0	0.248	26.0	1.0	0
7	10.0	115.0	70.0	27.0	102.5	Normal	Obese	Normal	35.3	0.134	29.0	0.0	0
8	2.0	197.0	70.0	45.0	424.5	Prediabetes	Obese	Normal	30.5	0.158	53.0	1.0	0
9	8.0	125.0	96.0	32.0	169.5	Normal	Obese	HS2	34.3	0.232	54.0	1.0	0
10	4.0	110.0	92.0	27.0	102.5	Normal	Obese	HS2	37.6	0.191	30.0	0.0	0
11	10.0	168.0	74.0	32.0	169.5	Prediabetes	Obese	Normal	38.0	0.537	34.0	1.0	0
12	10.0	139.0	80.0	27.0	102.5	Normal	Overweight	HS1	27.1	1.441	57.0	0.0	0
13	1.0	189.0	60.0	23.0	424.5	Prediabetes	Obese	Normal	30.1	0.398	59.0	1.0	0
14	5.0	166.0	72.0	19.0	175.0	Prediabetes	Overweight	Normal	25.8	0.587	51.0	1.0	0
15	7.0	100.0	74.5	32.0	169.5	Normal	Obese	Normal	30.0	0.484	32.0	1.0	0
16	0.0	118.0	84.0	47.0	230.0	Normal	Obese	HS1	45.8	0.551	31.0	1.0	0
17	7.0	107.0	74.0	32.0	169.5	Normal	Overweight	Normal	29.6	0.254	31.0	1.0	0
18	1.0	103.0	30.0	38.0	83.0	Normal	Obese	Normal	43.3	0.183	33.0	0.0	0
19	1.0	115.0	70.0	30.0	96.0	Normal	Obese	Normal	34.6	0.529	32.0	1.0	0

Figure 7: Complete dataset.

utcome	New_Glucose_Class_Normal	New_Glucose_Class_Prediabetes	New_BMI_Range_Underweight	New_BMI_Range_Healty	New_BMI_Range_Overweight	New_BMI_Range_Obese	New_Blood_Pressure_Normal	New_Blood_Pressure_HS1	New_Blood_Pressure_HS2
1.0	0	1	0	0	0	1	1	0	0
0.0	1	0	0	0	1	0	1	0	0
1.0	0	1	0	1	0	0	1	0	0
0.0	1	0	0	0	1	0	1	0	0
1.0	1	0	0	0	0	1	1	0	0

Figure 8: One hot encoding.

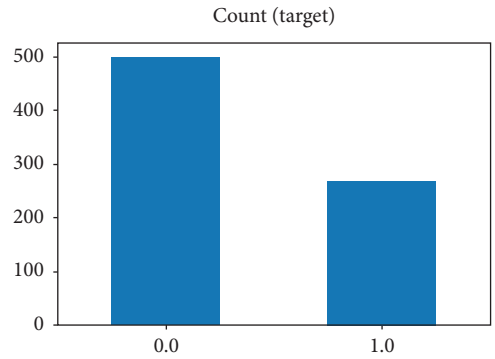


Figure 9: Before sampling.

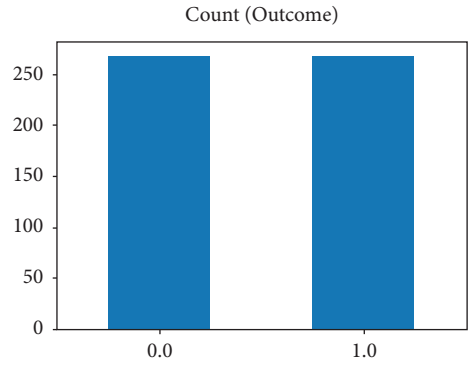


Figure 10: After sampling.

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

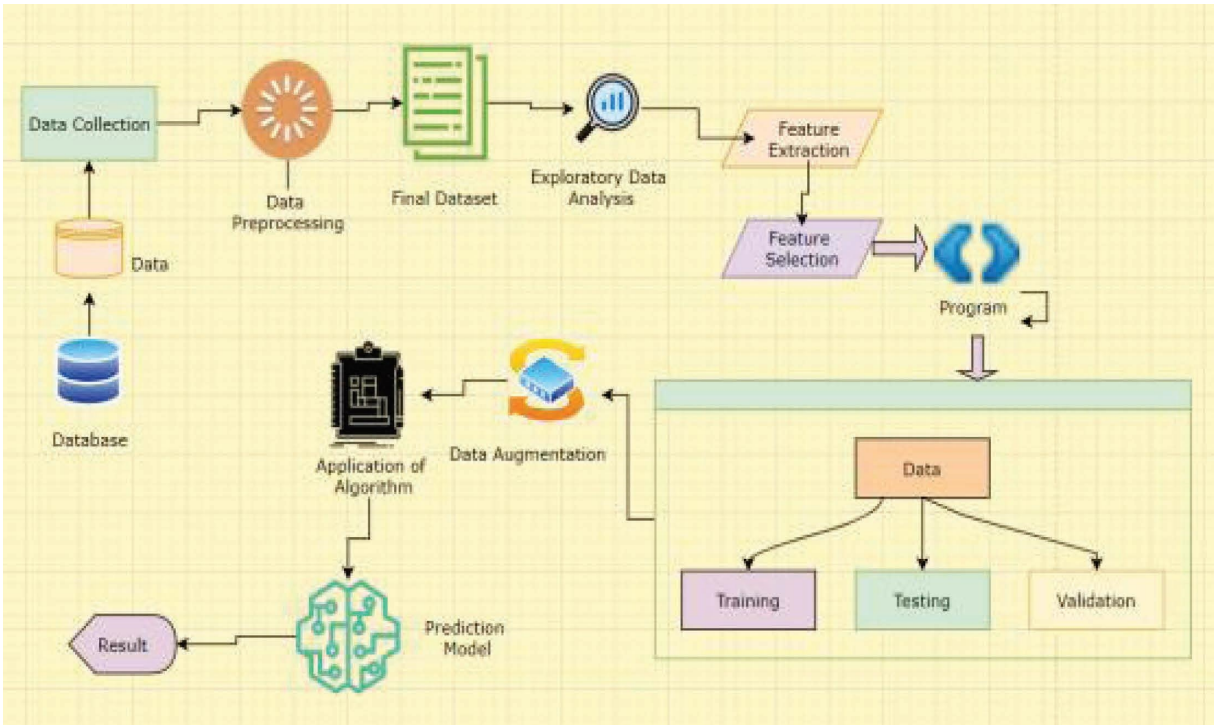


Figure 11: Architecture.

Table 1: Comparison of all classifiers.

Dataset	PIMA Indian Dataset						
Algorithm	Logistic Regression	XGBoost	Gradient Boosting Machine	Decision Tree	Extra Trees	Random Forest	Light Gradient Boosting Machine
Accuracy	75.20%	83.30%	94.1%	94.40%	94.60%	94.80%	95.20%

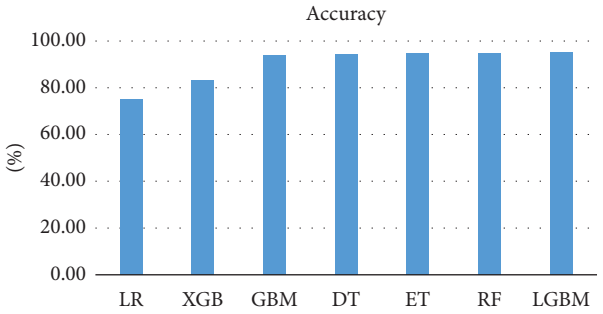


Figure 12: Accuracy percentage.

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

Accuracy of the GBM on test set : 0.929

	precision	recall	f1-score	support
0.0	0.98	0.92	0.95	107
1.0	0.83	0.96	0.89	47
accuracy			0.93	154
macro avg	0.91	0.94	0.92	154
weighted avg	0.94	0.93	0.93	154

Figure 13: Gradient boosting classifier.

Accuracy of the RANDOM FOREST on test set : 0.909

	precision	recall	f1-score	support
0.0	0.94	0.93	0.93	107
1.0	0.84	0.87	0.85	47
accuracy			0.91	154
macro avg	0.89	0.90	0.89	154
weighted avg	0.91	0.91	0.91	154

Figure 14: Random forest classifier.

Accuracy of the LGBM on test set : 0.922

	precision	recall	f1-score	support
0.0	0.96	0.93	0.94	107
1.0	0.84	0.91	0.88	47
accuracy			0.92	154
macro avg	0.90	0.92	0.91	154
weighted avg	0.93	0.92	0.92	154

Figure 15: Light gradient boosting classifier.

Table 2: Results after data augmentation.

Classifier	Accuracy	Precision		Recall		F1-score	
		0	1	0	1	0	1
GBM	92.9	0.98	0.83	0.92	0.96	0.95	0.89
RF	90.9	0.94	0.84	0.93	0.87	0.93	0.85
LGBM	92.2	0.96	0.84	0.93	0.91	0.94	0.88

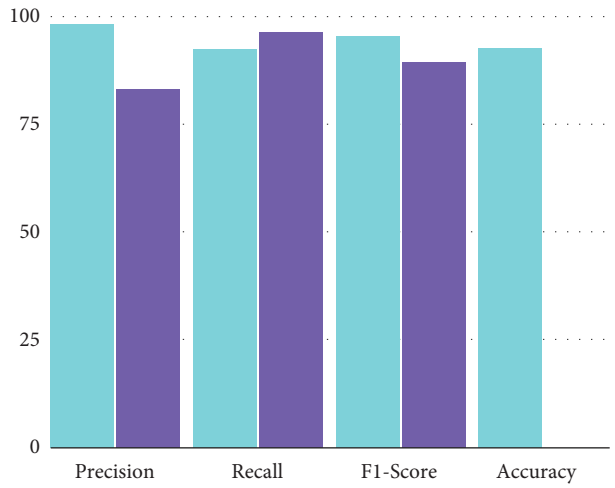


Figure 16: Gradient boosting machine graph.

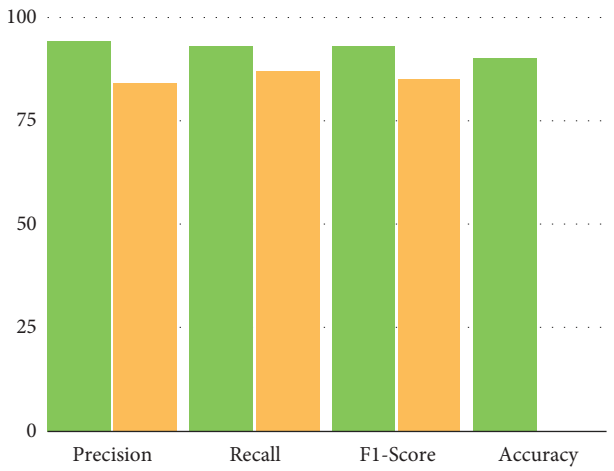


Figure 17: Random forest graph.

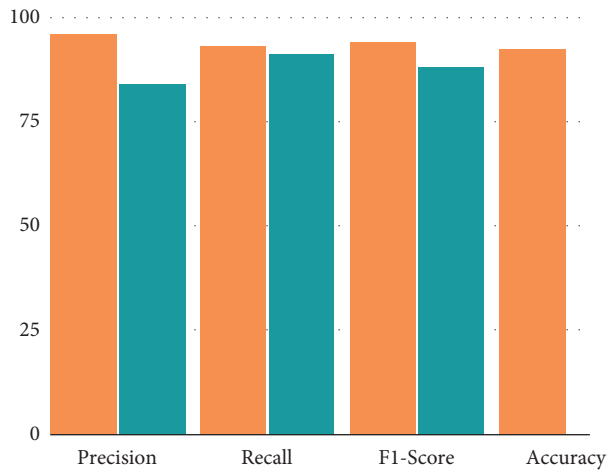


Figure 18: Light gradient boosting machine graph.

	Age	Fasting	Post_pran	Waist	BMI	Systolic	Diastolic	Hba1c	New_type	Gender_F	Gender_M	History_0	History_1	Class_no	Class_yes	label
0	0.737	0.558	0.649	0.941	0.769	0.333	0.000	0.643	Diabetes	1	0	0	1	0	1	2
1	0.158	0.047	-0.221	-0.471	-0.115	-0.667	-0.714	0.071	Pre-Diabetes	0	1	0	1	1	0	1
2	1.316	0.767	0.688	0.471	0.846	0.467	0.286	0.429	Diabetes	1	0	0	1	0	1	2
3	-1.053	-0.186	-0.208	-0.294	-0.673	-0.333	-0.714	-0.429	No	0	1	0	1	1	0	0
4	0.263	-0.070	-0.221	-0.412	-0.577	0.000	0.000	-0.357	Pre-Diabetes	0	1	0	1	1	0	1

Figure 19: Type classification.

	Expected Output	Predicted Output
559	0	0
33	2	2
275	2	2
198	1	1
192	1	1

Figure 20: Type classification output.

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

[1] X. Li, J. Zhang, and F. Safara, "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm," *Neural Processing Letters*, pp. 1–17, 2021.

[2] A. Arora, N. Shoeibi, V. Sati, A. González Briones, P. Chamoso, and E. Corchado, "Data augmentation using Gaussian mixture model on CSV files," *Distributed Computing and Artificial Intelligence*, Springer, L'Aquila, Italy, 2021.

[3] M. Shuja, S. Mittal, and M. Zaman, "Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree," *International Journal of Applied Engineering Research*, vol. 13, 2018.

[4] J. Chaki, S. Thillai Ganesh, S. 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*, vol. 34, no. 6, pp. 3204–3225, 2022.

[5] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. Garcia-Garcia, "Machine learning and deep learning predictive models for type 2 diabetes: a systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, no. 1, p. 148, 2021.

[6] N. G. Ramadhan and A. Romadhony, "Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 7, 2021.

[7] S. Islam Ayon, Department of Computer Science and Engineering Khulna University of Engineering & Technology Khulna-9203 Bangladesh, and M. Milon Islam, "Diabetes prediction: a deep learning approach," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 2, pp. 21–27, 2019.

[8] U. M. Butt, S. Letchmunan, M. Ali, F. H. Hassan, A. Baqir, and H. H. R. Sherazi, "Machine learning based diabetes classification and prediction for healthcare applications," *Journal of Healthcare Engineering*, vol. 2021, pp. 1–17, 2021, Hindawi.

[9] Y. Deng, L. Lu, L. Aponte et al., "Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients," *Npj Digit. Med.*vol. 4, no. 1, p. 109, 2021.

[10] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: applications and solutions," *ACM Computing Surveys*, vol. 52, no. 4, pp. 1–36, 2019.

[11] F. Zhuang, Z. Qi, K. Duan et al., "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[12] K. Li, J. Daniels, and C. Liu, P. Herrero and P Georgiou, Convolutional recurrent neural networks for glucose prediction," *IEEE J. Biomed. Health Inform.*vol. 24, no. 2, pp. 603–613, 2020.

[13] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, p. 515, 2018 Nov 6.

[14] K. Zarkogianni, M. Athanasiou, A. C. Thanopoulou, and K. S. Nikita, "Comparison of machine learning approaches toward assessing the risk of developing cardiovascular disease as a long-term diabetes complication," *IEEE J Biomed Health Inform*, vol. 22, no. 5, pp. 1637–1647, 2017.

[15] B. Alić, L. Gurbeta, and A. Badnjević, "Machine learning techniques for classification of diabetes and cardiovascular diseases," in *Proceedings of the 2017 6th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1–4, Bar, Montenegro, June 2017.

[16] N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, no. 1, p. 13.

[17] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," in *Proceedings of the 2015 4th Mediterranean Conference on Embedded Computing (MECO)*, pp. 378–382, Budva, Montenegro, June 2015.

[18] F. Mercaldo, V. Nardone, and A. Santone, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques," *Procedia Computer Science*, vol. 112, pp. 2519–2528, 2017.

[19] J. pradeep Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Computer Science*, vol. 47, pp. 45–51, 2015.

[20] R. Annamalai and R. Nedunchelian, "Diabetes mellitus prediction and severity level estimation using OWDANN algorithm," *Computational Intelligence and Neuroscience*, Article ID 5573179, 11 pages, 2021.

[21] C. Davitt, K. E. Flynn, R. K. Harrison, A. Pan, and A. Palatnik, "Current practices in gestational diabetes mellitus diagnosis and management in the United States: survey of maternal-fetal medicine specialists," *American Journal of Obstetrics and Gynecology*, vol. 225, no. 2, pp. 203-204, 2021 Aug.

[22] B. S. Ahamed and M. S. Arya, "Prediction of type 2 diabetes using the LGBM classifier methods and techniques," *Turkish Journal of Computer and Mathematics Education*, vol. 12, No.12.

[23] A. Jaggi, A. Sharma, N. Sharma, R. Singh, and P. S. Chakraborty, "Diabetes prediction using machine learning," 2021, https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/,.

[24] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, "Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation," *Advances in Human-Computer Interaction*, vol. 2022, pp. 1–14, 2022.

[25] M. A. Makroum, M. Adda, A. Bouzouane, and H. Ibrahim, "Machine learning and smart devices for diabetes management: systematic review," *Sensors*, vol. 22, no. 5, p. 1843, 2022.

[26] A. Nomura, M. Noguchi, M. Kometani, K. Furukawa, and T. Yoneda, "Artificial intelligence in current diabetes management and prediction," *Current Diabetes Reports*, vol. 21, no. 12, p. 61, 2021.

[27] Y. Spoorthy and T. Sunitha, "Diabetes prediction in women using machine learning techniques," *INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) NCCDS – 2021*, vol. 09, no. 12, 2021.

[28] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019.

[29] B. S. Ahamed, M. S. Arya, and A. O. Nancy V, "Prediction of type-2 diabetes mellitus disease using machine learning classifiers and techniques," *Frontiers of Computer Science*, vol. 4, 2022, https://www.frontiersin.org/articles/10.3389/fcomp.2022.835242.

[30] J. N. Bagrecha, "Diabetes disease prediction using neural network," *International Journal for Research in Applied Science and Engineering Technology*, vol. 7, no. 4, pp. 3888–3893, 2019.

[31] A. Ahmed, S. Aziz, A. Abd-alrazaq, F. Farooq, and J. Sheikh, "Overview of artificial intelligence–driven wearable devices for diabetes: scoping review," *Journal of Medical Internet Research*, vol. 24, no. 8, Article ID e36010, 2022.

[32] S. Ellahham, "Artificial intelligence: the future for diabetes care," *The American Journal of Medicine*, vol. 133, no. 8, pp. 895–900, 2020.

[33] B. Shamreen Ahamed and Dr. Meenakshi Sumeet Arya, "LGBM classifier based technique for predicting type-2 diabetes," *European Journal of Molecular & Clinical Medicine*, vol. 8, no. 3, pp. 454–467, 2021.

[34] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, vol. 10, no. 1, Article ID 11981, 2020.