Asia Pacific Academy of Science Pte. Ltd.

Computer and Telecommunication Engineering 2024, 2(1), 2446.
https://doi.org/10.54517/cte.v2i1.2446

Article

Prediction model for diabetes mellitus using machine learning algorithms for enhanced diagnosis and prognosis in healthcare

Prosanjeet Jyotirmay Sarkar[1], Satyanarayana Chanagala[2,*], George Chellin Jeya Chandra[3], Usha Ruby[3], Kavitha Manda[4]

[1] Department of Electronics and Communication Engineering, Dr. A. P. J. Abdul Kalam University, Indore 452016, India
[2] MAGNI5, Hyderabad 500073, India
[3] School of Computing Science and Engineering, VIT-BHOPAL University, Bhopal 466114, India
[4] Electronics and Communications, MNR College of Engineering and Technology, Sanga Reddy 502294, India
*** Corresponding author:** Satyanarayana Chanagala, scorppytek_satya@yahoo.com

CITATION

Sarkar PJ, Chanagala S, Chandra GCJ, et al. Prediction model for diabetes mellitus using machine learning algorithms for enhanced diagnosis and prognosis in healthcare. Computer and Telecommunication Engineering. 2024; 2(1): 2446. https://doi.org/10.54517/cte.v2i1.2446

ARTICLE INFO

Received: 25 December 2023
Accepted: 2 February 2024
Available online: 28 March 2024

Abstract: Diabetes mellitus (DM) affects the hormone insulin, which causes improper glucose metabolism and raises the body's blood sugar levels. With 4.2 million fatalities in 2019, DM is one of the top 10 global causes of mortality. Early detection of DM will aid in its treatment and avert complications. There must be a quick and simple technique to diagnose it. Such diseases can be managed and human lives can be saved with early diagnosis. Smart prediction techniques like Machine Learning (ML) have produced encouraging outcomes in predictive classifications. There has been a lot of interest in ML-based decision-support platforms for the prediction of chronic illnesses to provide improved diagnosis and prognosis help to medical professionals and the general population. By building predictive models using diagnostic medical datasets gathered from DM patients, ML algorithms efficiently extract knowledge that helps predict diabetic individuals. The association between DM and a healthy lifestyle is used in the model. In this study, the NHANES (National Health and Nutrition Examination Survey) data set is utilized, and five ML methods such as Artificial Neural Networks (ANN), CATBoost, XGBoost, XGBoost-histogram, and Light GBM to predict DM. The results of the experiment demonstrate that the XGB-h model outperformed other ML methods regarding area under the receiver operating characteristic curve (AUC-ROC), and accuracy. The most effective XGB-h framework can be used in a mobile app and a website to rapidly forecast DM. Real-time prediction using details delivered by the model at runtime can be developed as a whole bundle as a product. Clinicians can quickly determine who is likely to get diabetes using the proposed strategy, which will facilitate prompt intervention and caring.

Keywords: machine learning; fact-based filling; weighted-class training; Artificial Neural Networks; gradient boosting

1. Introduction

Diabetes Mellitus (DM) is a frequent illness in which the body's glucose amount is abnormally elevated for an extended time [1]. The effects of DM on many body organs cause damage to numerous bodily systems, including the blood vessels and nerve cells. If DM is not diagnosed at early stages, it can result in several significant, life-threatening consequences. A person's likelihood of having diabetes is shown to be influenced by the interactions of several risk factors, including their lifestyle, psychosocial environment, general health, demographics, and hereditary (genetic) composition. The prevalence of DM has dramatically risen due to the global population's advancing age and shifts in diet habits. Although the exact origins of diabetes are still unknown, many experts think that both inherited and environmental variables have a role [2]. Since DM is one of the world's most common and quickly progressing illnesses, several countries are working to prevent it from occurring as much as possible by predicting the signs of DM using diverse techniques. The DM condition can be managed and a person's life can be saved with prediction [3]. The knowledge that was retrieved can be utilized to predict DM patients with ease. According to the International Diabetes Federation (IDF), globally, 0.463 billion people will have diabetes.in 2019, 578 million in 2030, and 700 million in 2045 [4]. A tremendous improvement in the application of artificial intelligence (AI) has been shown recently in healthcare. A range of life-threatening disorders, such as breast cancer, diabetes, heart disease, etc., are predicted or diagnosed using AI algorithms in the medical field. Based on an enormous quantity of data, AI is capable of drawing sophisticated conclusions.

Deep learning (DL) and Machine learning (ML) which will be the mainstays of the AI boom in recent years have advanced significantly as a result of the rise in computational capacity and the correspondingly sharp growth in the performance of computers. In the present scenario, it is projected that massive quantities of organized data and a wealth of computing power will soon maximize the predictive capabilities of AI, resulting in a significant improvement in the accuracy of illness prediction algorithms for DM [5]. For predicting the risks and effects of the disease, ML models are employed [5–7]. An energy-efficient Wireless Sensor Network framework [8] is presented by the authors to detect Covid-19 patients. Machine learning (ML) algorithms such as Support Vector Machines(SVM) [9], Artificial Neural Networks (ANN) [10], J48 DecisionTree (DT), K-NearestNeighbors (KNN), CATBoost (CGB), Random Forest (RF), XGBoost (XGB) [11], XGBoost-histogram (XGB-h), Naive Bayes (NB), Logistic Regression (LR), Gradient Boost (GB), Multi-layer perception (MLP) and Light GBM (LGBM). To assess the models' efficacy, we consider their precision, accuracy, precision, accuracy, specificity, sensitivity, and area under the receiver operating characteristic curve (AUC-ROC), F1-Score, Negative Predictive Value (NPV), False Negatives (FN), Positive Predictive Value (PPV), False Positives (FP) and Matthew's correlation coefficient (MCC). The capacity to predict DM can be achieved using a variety of ML techniques. Choosing the optimal prediction method based on these variables is challenging [12]. The study makes use of five well-known ML algorithms, including ANN, CGB, XGB, XGB-h, and LGBM to predict DM from diagnostics medical data sets.

2. Related work

Chronic DM, which impairs the kidneys, eyes, nerves, heart, and various other systems, is brought on by continually elevated blood sugar levels. Early detection is essential for lowering the incidence and severity of diabetes, and knowing the major risk factors can encourage people to make lifestyle changes.

Agliata et al. [13] proposed an ML-based method to predict DM using SGD, RMSPROP, ADAM, and LM optimization algorithms. ADAM algorithm showed 86% accuracy and 0.934 AUC. The data set used was MIMIC-IV, NHANES, and MIMIC-III.

Faruque et al. [14] utilized DT, KNN, NB, and SVM algorithms for predicting DM. The accuracy of KNN is 70% and SVM is 79%.

Patil and Tamane [15] utilized LR, KNN, SVM, GB, DT, MLP, RF, and NB algorithms for predicting DM. The accuracy of KNN is 75% and SVM is 68%.

Iparraguirre-Villanueva et al. [16] proposed an ML-based DM prediction model which uses SVM, LR, DT, KNN, and Bernoulli Naïve Bayes (BNB) algorithms. The accuracy of BNB is 77.2% and KNN is 79.6%. Synthetic oversampling of minorities (SMOTE) was employed in this study.

Abegaz et al. [17] ML model utilized the AoU Research Program dataset. The algorithms used to create the model are RF, XGB, LR, and Weighted ensemble model (WEM). The ROC of RF is 0.77, LR is 0.7, and the accuracy of RF is 80%, and EGB is 77%.

Pranto et al. [18] put forth a model for predicting DM among Females in Bangladesh. They have utilized the PIMA Indian dataset (PID) dataset. KNN, RF, and NB are evaluated in this research, and 77.9% accuracy is obtained through RF with an AUC of 0.83.

Furthermore, they have employed a Kurmitola Hospital database in which the maximum accuracy is 81.2% by KNN and 0.84 AUC by Syed and Khan [19] proposed a method using various ML algorithms to design a prediction model. The dataset used is the NHANES dataset and the PID dataset. They have used the Western Region of Saudi Arabia dataset along with other datasets mentioned previously. The accuracy obtained from the proposed model is 82.1%, and AUC is 0.829.

Abdulhadi and Al-Mousa [20] put forth a model for predicting DM among females at an early stage using ML approaches. The model shows an accuracy of 82% using the RF algorithm.

Ahmed et al. [21] proposed a DM prediction model utilizing a fused ML methodology that uses ANN and SVM. The data collection used in the current investigation is split 70:30 between the training and testing parts. The input membership function of the fuzzy model is derived from the outputs of these models, and ultimately, fuzzy logic decides if the diagnosis of diabetes is positive or negative. The accuracy of predicting DM is 94.87%.

Manikandababu et al. [22] proposed a DM prediction model using the PID dataset of 500 non-DM individuals and 268 DM individuals. Ensemble and stacking methods are used to build the models. The model is built utilizing NB, SVM, and DT algorithms. The method achieved 94% accuracy.

Khanam et al. [23] put forth a model for predicting DM utilizing a PID data set with 768 individuals with nine attributes. ANN, LR, and SVM algorithms are used to build the model and the ANN model obtained 88.6% accuracy.

Hasan et al. [24] used the PID dataset (dataset size is 768) for DM prediction. The classification accuracy of the model is 78.9%.

Zou et al. [25] used the Luzhou and PID data set (dataset size is 68994, 768) for DM prediction. The model has an 80.84% classification accuracy. The data set included biochemical indicators such as HDL and LDL.

Maniruzzuman et al. [26] used the NHANES dataset (dataset size is 9858) for DM prediction. The classification accuracy of the model is 92.75%.

Jackins et al. [27] used the PID dataset (datasetsizeis768) for DM prediction. The classification accuracy of the model is 74.46%.

Sneha and Gangil [28] used the PIDD dataset (dataset size is 2500) for DM prediction. The classification accuracy of the model is 82.3%.

Bhaskar et al. [29] used the PIDD dataset (dataset size is 768) for DM prediction. The classification accuracy of the model is 77.5%.

Sisodia and Sisodia [30] used the PIDD dataset (dataset size is 768) for DM prediction. The model has a 76.3% classification accuracy.

Orabi et al. [31] used the Egyptian National Research Centre data set for DM prediction. The classification accuracy of the model is 84%.

Previous studies have connected lifestyle threat aspects for diabetes such as smoking, poor diet, BMI, drinking alcohol, and lack of exercise [32–36].

Juneja [37] utilized the PID dataset and also included biochemical indicators for DM prediction. They have compared supervised and unsupervised training models in their study and concluded that the supervised model outweighs the other model.

Tigga [38] used LR, KNN, SVM, NB, DT, and RF algorithms are used to construct the model for predictions. The data set employed is their dataset and PID dataset. The RF model’s accuracy obtained for the PID dataset is 75% and own dataset was 94.1% and the AUC is 1 for both data sets. A machine learning (ML) model is created in this work that more precisely predicts diabetes using a lifestyle variable that is more easily accessible. The proposed model is a user-friendly prediction model for DM and the model built is independent of biochemical indications.

3. Proposed method

The proposed model is shown in **Figure 1**. The flow graph of the same is shown in **Figure 2**.

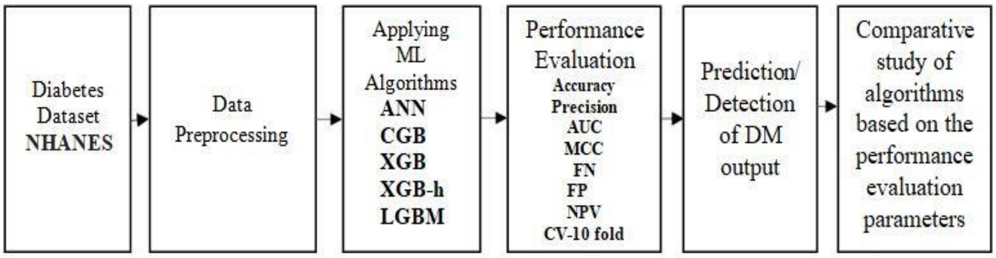

Figure 1. Proposed model’s process flow for DM prediction.

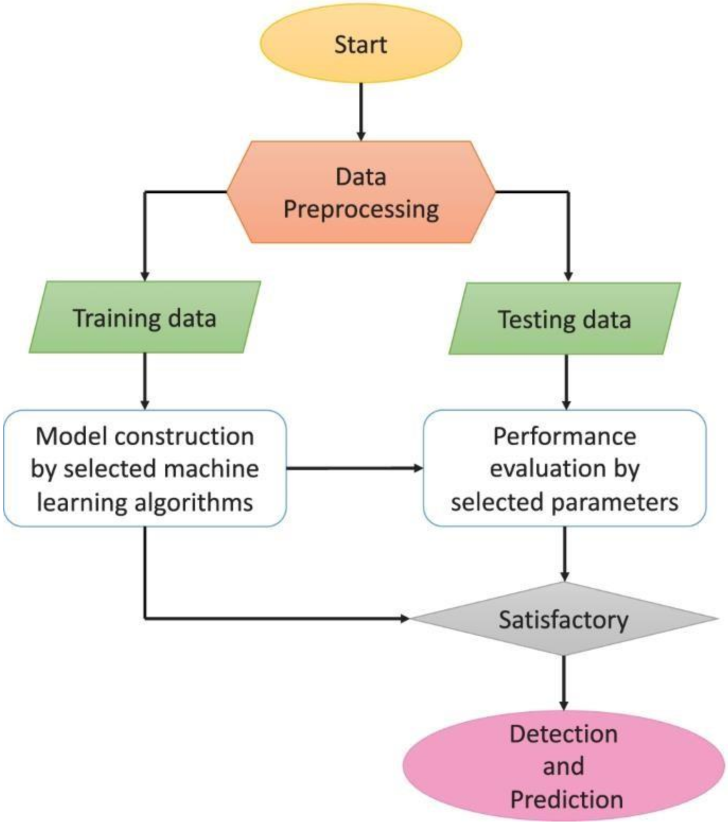
Start
Data Preprocessing
Training data
Testing data
Model construction by selected machine learning algorithms
Performance evaluation by selected parameters
Satisfactory
Detection and Prediction

Figure 2. The flow chart of the process flow.

3.1. Data set

Data from the NHANES dataset are gathered during data collection. The collected data is for 6 years which are 2007–2014, 2015–2016, and 2017–2018. The initial phase is to separate and identify the essential features because the majority of the over 1800 feature variables in the original dataset are irrelevant to the study. The primary components of health behaviors such as exercise, food habits, and smoking were selected to create the final dataset. A current study [39] that involved a million participants and found that those having elevated blood pressure increase the risk of developing Type-2 DM further influenced the decision to include the hypertension component because of its connection to DM. We chose to include drinking alcohol in the selected features since it is one of the most important parts of healthy behaviours, even though there are disagreements among the research that concentrated on the association between alcohol intake and diabetes. The investigation also considers additional demographic characteristics like racial or ethnic background, age, marriage status, level of educational attainment, annual earnings for families, and the ratio of family earnings to poverty thresholds. 55,939 observations comprising of 20 years to 60 years age group, and 30 features make up the final dataset and all of its values have been coded as numbers. **Table 1** shows the selected features for the study.

Table 1. Selected features for the study.

Sequence Number	Features selected
1	Gender
2	Date of Birth (Age)
3	Race
4	Smoker currently
5	high blood pressure (HP) details (Single time)
6	School/College education detail–Children (06–19 years)
7	Marriage details
8	Pregnancy information
9	Earnings of the family (Annually)
10	Cholesterol (milligram)
11	Dietary fibre (gram)
12	Carbohydrates (gram)
13	Overall fat (gram)
14	Protein (gram)
15	Overall sugars (gm)
16	Sodium (milligram)
17	BMI
18	high blood pressure (HP) details (>2 times)
19	Hypertension based on Age
20	Glucose (mmol/L)–Fasting- Status of Diabetes
21	Weight (Kilogram)
22	Exercises (Heavy)–Per week
23	Exercises (moderate)–Per week
24	Drinking habits (alcohol)
25	Smoking habits (100 cigarettes till now)
26	Family earnings to poverty thresholds
27	School/College education detail–Adults (>20 years)
28	Exercises (Heavy)–Daily (Minutes)
29	Exercises (moderate)–Daily (Minutes)
30	Potassium (milligram)

3.2. Preprocessing of dataset

The data is processed in the Data Processing stage to make it suitable for ML algorithm execution. This involves classifying data into appropriate categories, ensuring that the information is relevant, and eliminating any unnecessary data. Several of the readily available datasets contain missing or erroneous values. We have to clean the data set before feeding it to the model for training to create an effective ML model. Cleaning data refers to the procedure of dealing with incomplete and inaccurate information in the dataset, either by deleting them entirely or by using different methods to fill in the gaps. The dataset includes observations of individuals ranging in age from under 1 to 80. Given that DM is more prevalent in adults, individuals who had reached at least 18 years old were the focus of our investigation. The dataset size decreased to 35485 observations after every observation made by individuals under the age of 18 was eliminated. Furthermore, a large number of the variables in our data set are missing. The number of values that are missing for each variable and its ratio are displayed in **Table 2**. The last stage was to fill in the values that were missing, which required us to go through all of the features in the dataset that had missing values. Five features are overlooked in the dataset that had no missing values, according to **Table 2**, and concentrated on the remaining 25 features. The values that were missing were filled in using a variety of methods, including ML logic based on the knowledge of the dataset, and statistical procedures. The latter is the one that is most frequently employed. For certain characteristics, including Pregnancy information, Marriage details, and School/College education details, we utilized reasoning and fact-based filling techniques; for other features, such as weekly exercises (minutes), statistical techniques like the median are used. Other values that were missing had to be dropped because we were unable to deal with the min anyway.

Table 2. The number of values that are missing for each variable and its ratio.

Sequence Number	Features selected	Number of values that are missing	Number of values that are missing %
1	Gender	0	0%
2	DateofBirth (Age)	0	0%
3	Race	0	0%
4	Smoker currently	20545	57.90%
5	High Blood Pressure (HP) details (Single time)	0	0%
6	School/College education detail–Children (06–19 years)	1769	4.99%
7	Marriage details	1769	4.99%
8	Pregnancy information	28205	79.48%
9	Earnings of the family (Annually)	691	1.95%
10	Cholesterol (milligram)	3062	8.63%
11	Dietary fibre (gram)	3062	8.63%
12	Carbohydrates (gram)	3062	8.63%
13	Overall fat (gram)	3062	8.63%
14	Protein (gram)	3062	8.63%
15	Overall sugars (gm)	3062	8.63%
16	Sodium (milligram)	3062	8.63%
17	BMI	847	2.39%
18	high blood pressure (HP) details (>2 times)	23118	65.15%
19	Hypertension based on Age	23157	65.26%
20	Glucose (mmol/L)–Fasting- Status of Diabetes	0	0%
21	Weight (Kilogram)	789	2.22%
22	Exercises (Heavy)–Per week	27442	77.33%
23	Exercises (moderate)–Per week	27461	77.39%

Table 2. (*Continued*).

Sequence Number	Features selected	Number of values that are missing	Number of values that are missing %
24	Drinking habits (alcohol)	3062	8.63%
25	Smoking habits (100 cigarettes till now)	886	2.50%
26	Family earnings to poverty thresholds	3564	10.04%
27	School/College education detail–Adults (>20 years)	33716	95.01%
28	Exercises (Heavy)–Daily (Minutes)	27442	77.33%
29	Exercises (moderate)–Daily (Minutes)	27461	77.39%
30	Potassium (milligram)	3062	8.63%

3.3. Dataset after preprocessing (final dataset)

Following the preprocessing of the initial dataset, the final dataset has 14682 observations and 21 features in addition to the target variable (DM). The other 16 features have category values, whereas five of the features have numerical values. Categorical data are data kinds that can be stored and recognized depending on the names or labels given to them. The term “numerical data” describes information that takes the shape of numbers. **Table 3** shows the number of categories and types of data.

Table 3. Categories.

Sequence Number	Features selected	Number of categories	Type of data
1	Gender	2	Categorical
2	Date of Birth (Age)	-	Numerical
3	Race	5	Categorical
4	School/College education detail	5	Categorical
5	Marriage details	7	Categorical
6	Pregnancy information	4	Categorical
7	Earnings of the family (Annually)	-	Numerical
8	Cholesterol (milligram)	4	Categorical
9	Dietary fiber (gram)	4	Categorical
10	Carbohydrates (gram)	6	Categorical
11	Overall fat (gram)	5	Categorical
12	Protein (gram)	4	Categorical
13	Overall sugars (gm)	6	Categorical
14	Sodium (milligram)	5	Categorical
15	BMI	-	Numerical
16	Hypertension	2	Categorical
17	Glucose (mmol/L)–Fasting- Status of Diabetes	2	Categorical
18	Exercises-Per week	-	Numerical
19	Drinking habits (alcohol)	-	Numerical
20	Smoking habits	4	Categorical
21	Family earnings to poverty thresholds	12	Categorical
22	Potassium (milligram)	6	Categorical

3.4. ML model building

The different steps in the Machine Learning process are shown in **Figure 3**.

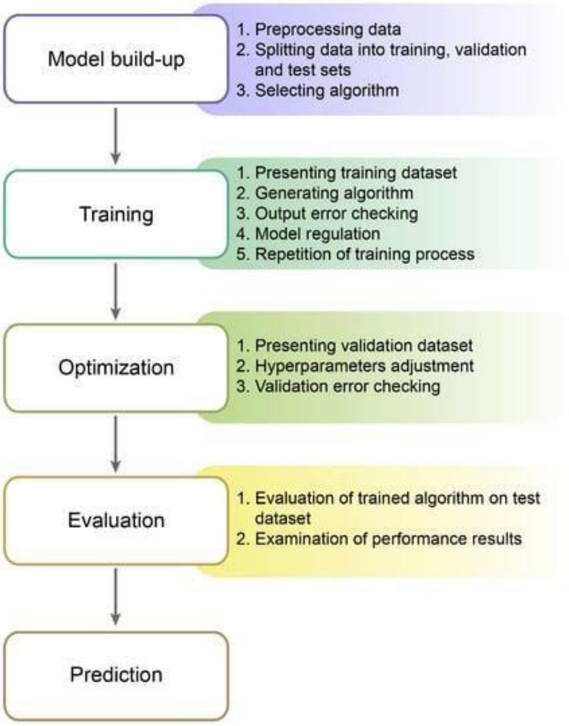
Model build-up
1. Preprocessing data
2. Splitting data into training, validation and test sets
3. Selecting algorithm
Training
1. Presenting training dataset
2. Generating algorithm
3. Output error checking
4. Model regulation
5. Repetition of training process
Optimization
1. Presenting validation dataset
2. Hyperparameters adjustment
3. Validation error checking
Evaluation
1. Evaluation of trained algorithm on test dataset
2. Examination of performance results
Prediction

Figure 3. ML process to predict DM.

Applying ML algorithms entails training the model and performing various ML techniques on the datasets. In the performance assessment step, the outcomes of the implemented algorithms are assessed, and each algorithm's performance is evaluated using precision, AUC, accuracy, etc. are examined. In the comparative evaluation stage, all of the applied ML algorithms are compared to see which one is the most effective at detecting DM and predicting blood glucose levels. In this study, the NHANES dataset is utilized and five ML methods such as Artificial Neural Networks (ANN), and gradient boosting such as CATBoost (CGB), XGBoost (XGB), XGBoost-histogram (XGB-h), and Light GBM (LGBM) to predict DM. Due to the population's mix of people with and without diabetes mellitus, practically every medical dataset is unbalanced. The target variable (DM) consists of 73.63% of participants without DM and 22.40% of participants with DM. When the proportion of the target's values differs by such a large amount, the resulting model will be incredibly biased and inadequate. Many approaches were put out to address the dataset imbalance, such as threshold-moving, over-sampling the minority class, weighted-class training, and under-sampling the majority class. In this research weighted-class, training method is employed. This approach was selected since it protects the dataset from alterations and has worked most effectively in this research. The data had to be divided into three sets before being fed to the model. Set-1 is for training the model, set-2 is for validating the model's accuracy through the process of training, and set-3 will assess the trained model's performance with never-before-tested data.

Data from the NHANES dataset are gathered during data collection. The collected data is for 6 years that are 2007–2014, 2015–2016, and 2017–2018. For training and validation, the data used from the years 2007–2016, out of which 20% of the data are for validation and the training of the model uses 80% of the data. For testing the model, the data was utilized from the year 2017–18. Five ML methods such as Artificial Neural Networks (ANN), CATBoost (CGB), XGBoost (XGB), XGBoost-histogram (XGB-h), and Light GBM (LGBM) are employed to predict DM. **Figure 4** shows the ML process to predict DM. To train the model, the following steps are followed (i) Enumerate through the selected algorithm, (ii) use the training set, train the model, (iii) validate the model with validating data, (iv) tune the model's hyperparameters, and (v) Utilizing the test data, assess the model's effectiveness. To choose the best model, we first had to train every model using the training dataset, then use the validation data set to identify the model's optimal hyperparameters, and then use the test data set to assess the model.

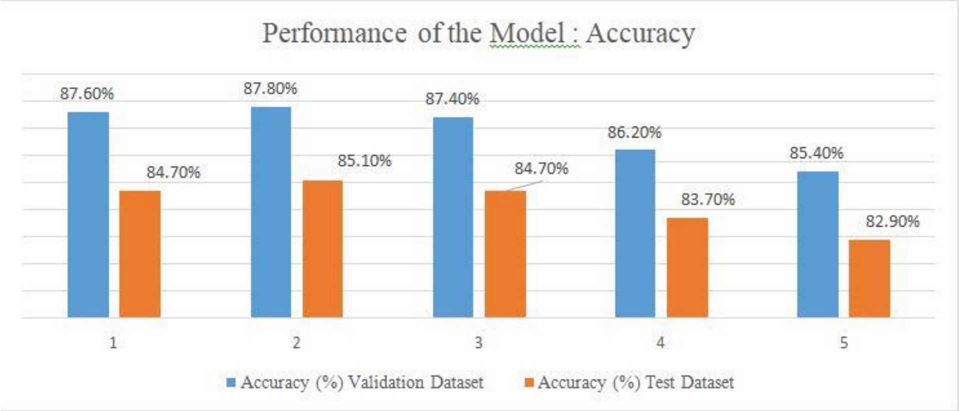

Figure 4. The accuracy of the classifiers for diabetic prediction is based on validation and test datasets.

4. Result

Following the training procedure, we identified the best model, which was created using the XGB-h. The model's cross-validation value was 0.864, and its overall accuracy with the validation dataset and test data set was 87.7% and 85%, respectively. Amongst all of the examined algorithms, the ANN model got the lowest ratings. Although ANN is well known for its cutting-edge performance across numerous ML tasks, it did not do the task as well as other algorithms. False Positives are low-risk errors in this study since patients can readily get tested for the disease and confirm that they don't have it, so even if the model incorrectly predicts that the patient has diabetes when he doesn't, it won't be an issue. The patient will believe he doesn't have diabetes and may not get checked for the disease, which could cause the diabetes problems to worsen. However, if the model predicts that the individual is diabetes-free when in fact he has diabetes, this will be an elevated-risk error. **Figure 4** shows the accuracy of the classifiers for diabetic prediction based on validation and test datasets. **Figure 5** shows the precision of the classifiers for diabetic prediction based on validation and test datasets. The specificity of the classifiers for diabetic prediction is based on validation and test datasets (**Figure 6**). The sensitivity of the classifiers for diabetic prediction based on validation and test dataset (**Figure 7**). **Figure 8** shows the F1-Score of the classifiers for diabetic prediction based on validation and test datasets. **Figure 9** shows the NPV of the classifiers for diabetic prediction based on validation and test datasets. **Figure 10** shows the MCC of the classifiers for diabetic prediction based on validation and test datasets. **Figure 11** shows the CV10-fold of the classifiers for diabetic prediction based on validation and test datasets. **Figure 12** shows the AUC of the classifiers for diabetic prediction based on validation and test data. **Figure 13** shows the FP of the classifiers for diabetic prediction based on validation and test data. **Figure 14** shows the FN of the classifiers for diabetic prediction based on validation and test data.

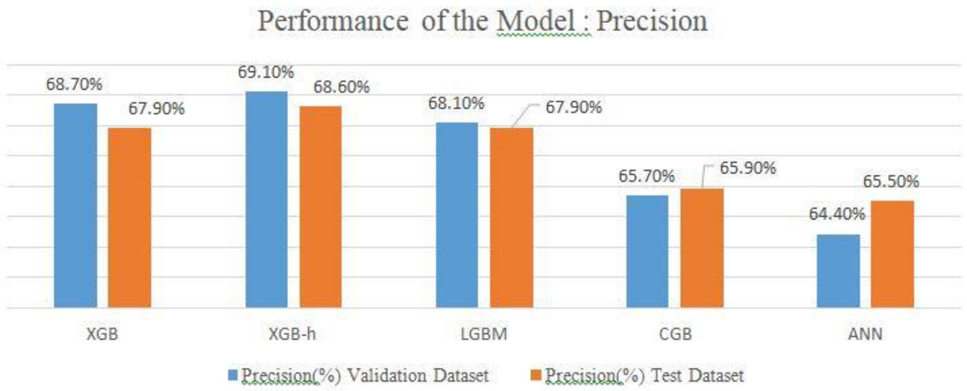

Figure 5. The precision of the classifiers for diabetic prediction based on validation and test dataset.

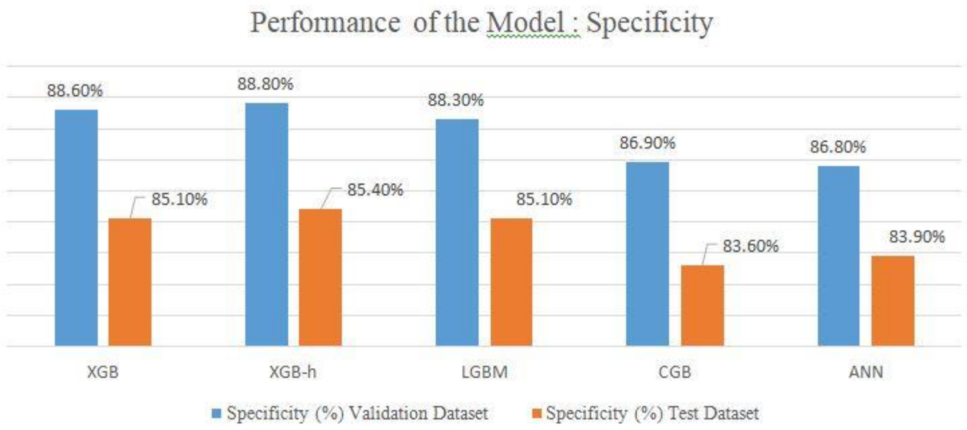

Figure 6. The specificity of the classifiers for diabetic prediction based on validation and test dataset.

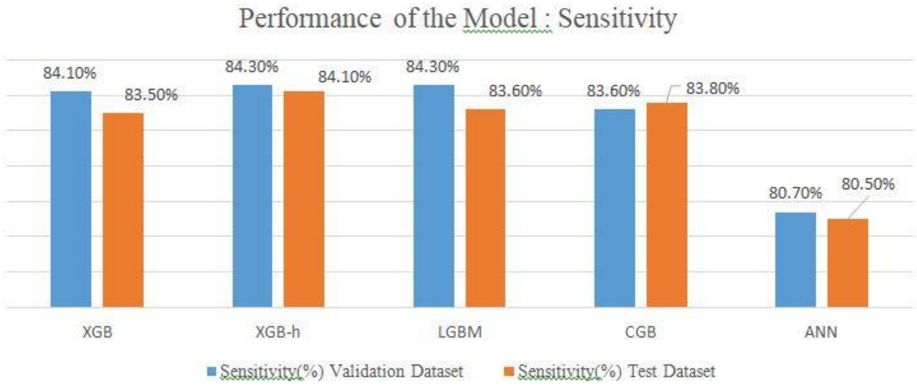

Figure 7. The sensitivity of the classifiers for diabetic prediction based on validation and test dataset.

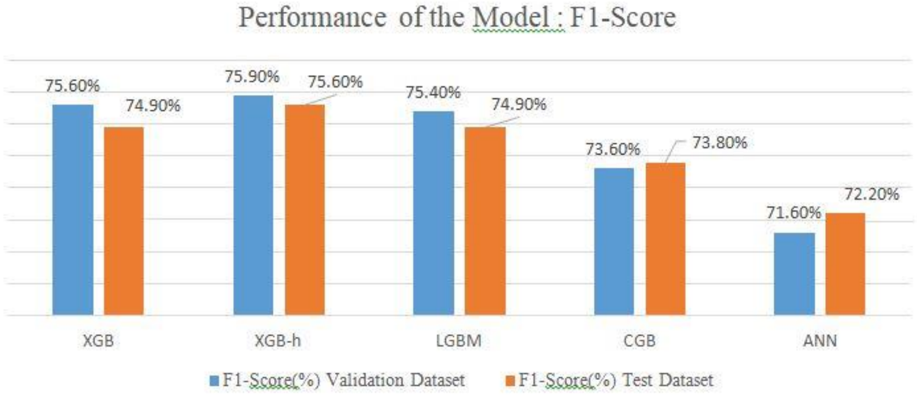

Figure 8. The F1-Score of the classifiers for diabetic prediction based on validation and test dataset.

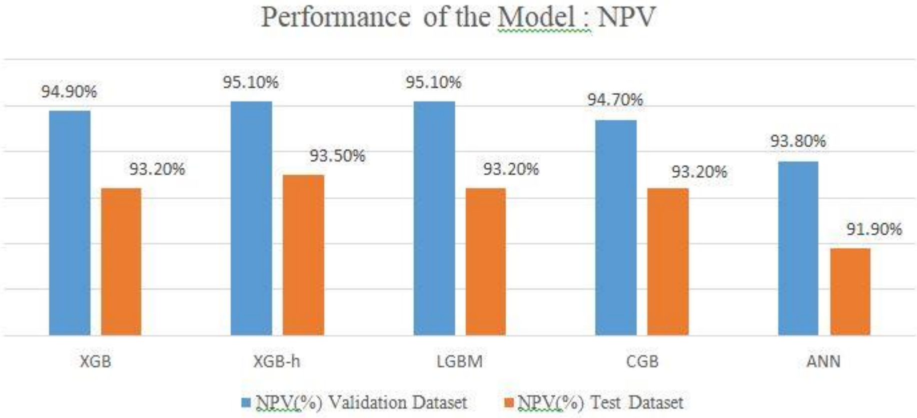

Figure 9. The NPV of the classifiers for diabetic prediction based on validation and test dataset.

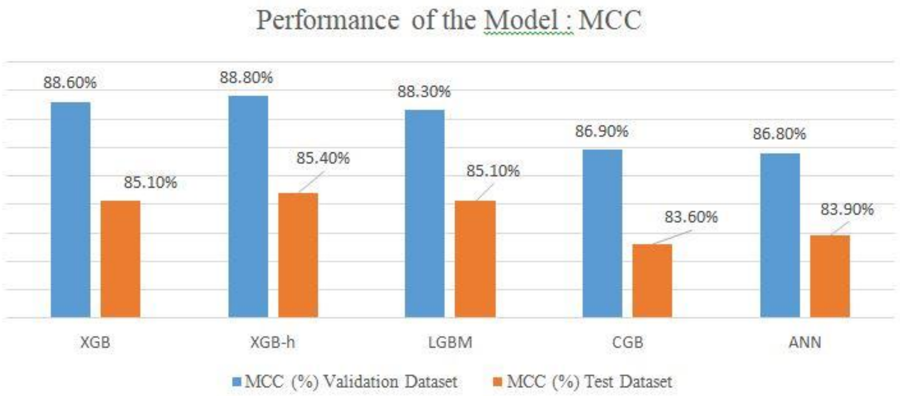

Figure 10. The MCC of the classifiers for diabetic prediction based on validation and test dataset.

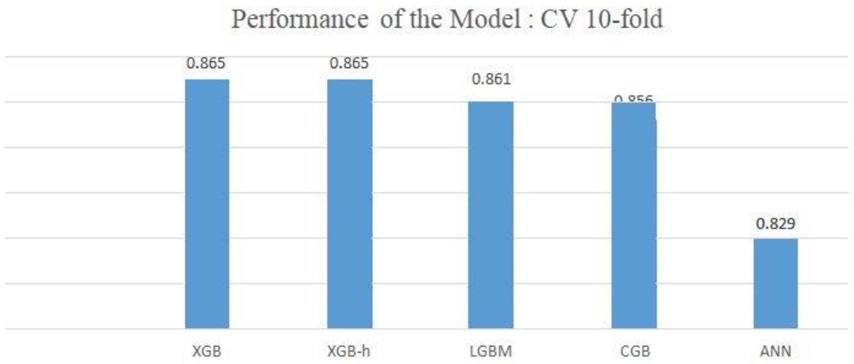

Figure 11. The CV10-fold of the classifiers for diabetic prediction based on validation and test dataset.

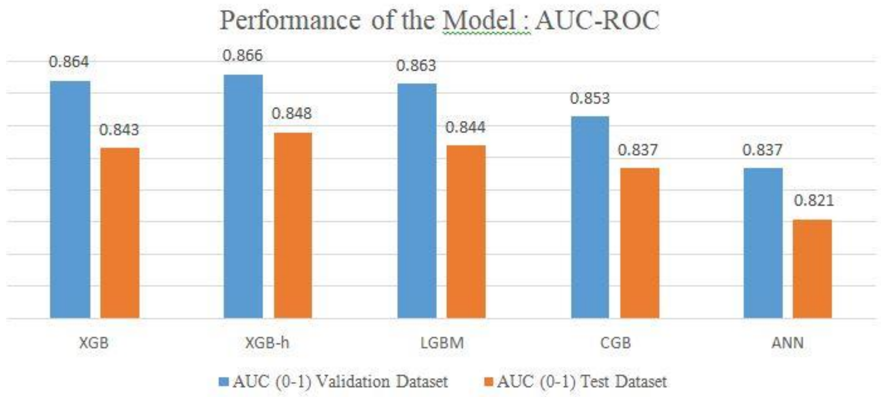

Figure 12. The CV10-fold of the classifiers for diabetic prediction based on validation and test dataset.

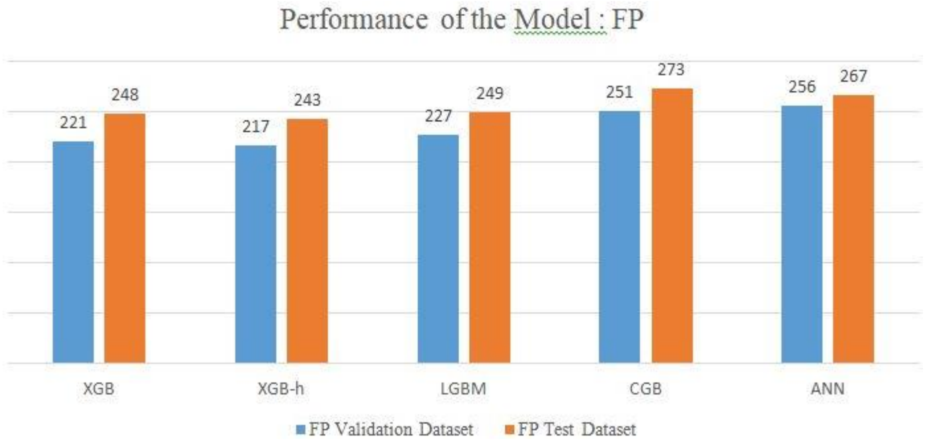

Figure 13. The FP of the classifiers for diabetic prediction based on validation and test dataset.

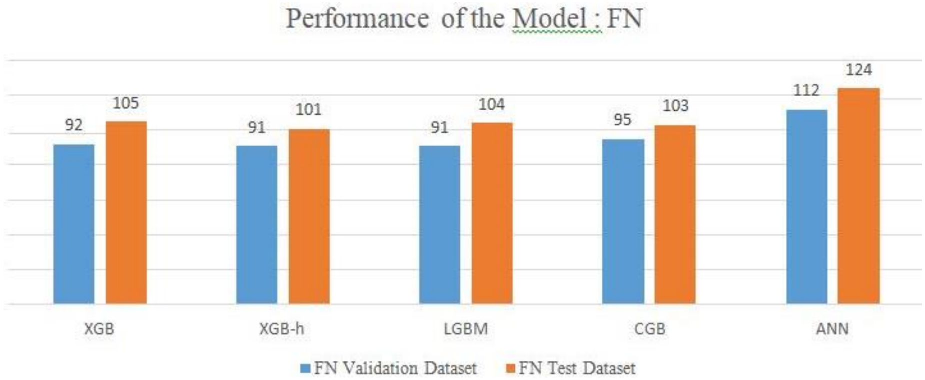

Figure 14. The FN of the classifiers for diabetic prediction based on validation and test dataset.

5. Conclusion

A smart expert system that uses ML methods to deliver improved outcomes using statistical metrics has been presented for enhanced DM prediction. The association between DM and a healthy lifestyle is used in the model. In this study, we use the NHANES dataset and five ML methods such as ANN, CGB, XGBoost XGB, XGB-h, and LGBM to predict DM. The results of the experiment demonstrate that the XGB-h model is superior to other ML methods in terms of accuracy and AUC. These algorithms will be applied to various, sizable, and real-time datasets as part of the present investigation to determine the effectiveness of the suggested solution. The most effective XGB-h framework can be used in a mobile app and a website to rapidly forecast DM. Real-time predictions using details delivered by the model at runtime can be developed as a whole bundle as a product Clinicians can quickly determine who is likely to get diabetes using the proposed strategy, which will facilitate prompt intervention and caring. The suggested approach makes it possible to predict the prevalence of DM more precisely. The application of ML techniques to predict the likelihood of developing any DM side effects, including retinopathy, kidney disease, and cardiovascular disease is also possible with the proposed method. The suggested approach will lessen the escalating cost of DM on the healthcare system while assisting in maintaining the quality of life for those with DM. The models are stored in a system that keeps data in the cloud for future use. The model uses the patient's most recent health information to decide whether they have diabetes. However, in the present work, the contribution of a particular feature towards DM is not considered.

Author contributions: Conceptualization, PJS and SC; methodology, PJS and SC; software, PJS; validation, PJS, SC, MK and GCJC; formal analysis, UR; investigation, PJS and SC; resources, PJS; data curation, PJS; writing—original draft preparation, PJS and SC; writing—review and editing, SC; visualization, PJS; supervision, PJS and SC; project administration, PJS; funding acquisition, PJS. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Mahajan S, Sarangi PK, Sahoo AK, et al. Diabetes Mellitus Prediction using Supervised Machine Learning Techniques. 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT); 5 May 2023. doi: 10.1109/incacct57535.2023.10141734
2. Charles RKJ, Mary AB, Jenova R, et al. VLSI design of intelligent, Self-monitored and managed, Strip-free, Non-invasive device for Diabetes mellitus patients to improve Glycemic control using IoT. Procedia Computer Science. 2019; 163: 117-124. doi: 10.1016/j.procs.2019.12.093
3. Balaji KV, Sugumar R. A Comprehensive Review of Diabetes Mellitus Exposure and Prediction using Deep Learning Techniques. 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI); 8 December 2022. doi: 10.1109/icdsaai55433.2022.10028832
4. Ansari RM, Harris MF, Hosseinzadeh H, et al. Application of Artificial Intelligence in Assessing the Self-Management Practices of Patients with Type 2 Diabetes. Healthcare. 2023; 11(6): 903. doi: 10.3390/healthcare11060903
5. Chaki J, Thillai Ganesh S, Cidham SK, et al. Machine learning and artificial intelligence-based Diabetes Mellitus detection and self-management: A systematic review. Journal of King Saud University - Computer and Information Sciences. 2022; 34(6): 3204-3225. doi: 10.1016/j.jksuci.2020.06.013
6. Charles Rajesh Kumar J, Baskar D, Mary Arunsi B, Vinod Kumar D. Energy-Efficient and Secure IoT Architecture Based on a Wireless Sensor Network Using Machine Learning to Predict Mortality Risk of Patients with CoVID-19. 2021 6th International Conference on Communication and Electronics Systems (ICCES); Coimbatore, India. 2021. pp. 1853-1861. doi: 10.1109/ICCES51350.2021.948895
7. Kumar JCR, Arunsi BM, Majid MA. A Machine Learning-driven IoT Architecture for Predicting the Growth and Trend of Covid-19 Epidemic Outbreaks to Identify High-risk Locations. 2023 20th Learning and Technology Conference (L&T); 26 January 2023. doi: 10.1109/lt58159.2023.10092331
8. Charles Rajesh Kumar J, Mary Arunsi B, Majid MA. Energy-Efficient IoT-Based Wireless Sensor Network Framework for Detecting Symptomatic and Asymptomatic COVID-19 Patients Using a Fuzzy Logic Approach. Contemporary Applications of Data Fusion for Advanced Healthcare Informatics. 2023; 25-51. doi: 10.4018/978-1-6684-8913-0.ch002
9. Lal ND, K H, T S, et al. An Effective Expectation of Diabetes Mellitus via Improved Support Vector Machine through Cloud Security. 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS); 24 February 2023. doi: 10.1109/icicacs57338.2023.10099888
10. R B, K TM, D J, et al. Diabetes Mellitus Diagnosis based on Tongue Images using Machine Learning. 2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS); 17 March 2023. doi: 10.1109/icaccs57279.2023.10112849
11. Laxmikant K, Bhuvaneswari R, Natarajan B. An Efficient Approach to Detect Diabetes Using XGBoost Classifier. 2023 Winter Summit on Smart Computing and Networks (WiSSCoN); 15 March 2023. doi: 10.1109/wisscon56857.2023.10133854
12. Malik A, Parihar V, Srivastava J, et al. Prognosis of Diabetes Mellitus Based on Machine Learning Algorithms. 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom); New Delhi, India. 2023. pp.

1466-1472.

13. Agliata A, Giordano D, Bardozzo F, et al. Machine Learning as a Support for the Diagnosis of Type 2 Diabetes. International Journal of Molecular Sciences. 2023; 24(7): 6775. doi: 10.3390/ijms24076775
14. Faruque MdF, Asaduzzaman, Sarker IH. Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus. 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE); February 2019. doi: 10.1109/ecace.2019.8679365
15. Patil R, Tamane S. A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes. International Journal of Electrical and Computer Engineering (IJECE). 2018; 8(5): 3966. doi: 10.11591/ijece.v8i5.pp3966-3975
16. Iparraguirre-Villanueva O, Espinola-Linares K, Flores Castañeda RO, et al. Application of Machine Learning Models for Early Detection and Accurate Classification of Type 2 Diabetes. Diagnostics. 2023; 13(14): 2383. doi: 10.3390/diagnostics13142383
17. Abegaz TM, Ahmed M, Sherbeny F, et al. Application of Machine Learning Algorithms to Predict Uncontrolled Diabetes Using the All of Us Research Program Data. Healthcare. 2023; 11(8): 1138. doi: 10.3390/healthcare11081138
18. Pranto B, Mehnaz SkM, Mahid EB, et al. Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh. Information. 2020; 11(8): 374. doi: 10.3390/info11080374
19. Syed AH, Khan T. Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study. IEEE Access. 2020; 8: 199539-199561. doi: 10.1109/access.2020.3035026
20. Abdulhadi N, Al-Mousa A. Diabetes Detection Using Machine Learning Classification Methods. 2021 International Conference on Information Technology (ICIT); 14 July 2021. doi: 10.1109/icit52682.2021.9491788
21. Ahmed U, Issa GF, Khan MA, et al. Prediction of Diabetes Empowered With Fused Machine Learning. IEEE Access. 2022; 10: 8529-8538. doi: 10.1109/access.2022.3142097
22. Manikandababu CS, IndhuLekha S, Jeniefer J, et al. Prediction of Diabetes using Machine Learning. 2022 International Conference on Edge Computing and Applications (ICECAA); 13 October 2022. doi: 10.1109/icecaa55415.2022.9936375
23. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021; 7(4): 432-439. doi: 10.1016/j.icte.2021.02.004
24. Hasan MdK, Alam MdA, Das D, et al. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. IEEE Access. 2020; 8: 76516-76531. doi: 10.1109/access.2020.2989857
25. Zou Q, Qu K, Luo Y, et al. Predicting Diabetes Mellitus With Machine Learning Techniques. Frontiers in Genetics. 2018; 9. doi: 10.3389/fgene.2018.00515
26. Maniruzzaman Md, Kumar N, Menhazul Abedin Md, et al. Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer Methods and Programs in Biomedicine. 2017; 152: 23-34. doi: 10.1016/j.cmpb.2017.09.004
27. Jackins V, Vimal S, Kaliappan M, et al. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. The Journal of Supercomputing. 2020; 77(5): 5198-5219. doi: 10.1007/s11227-020-03481-x
28. Sneha N, Gangil T. Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data. 2019; 6(1). doi: 10.1186/s40537-019-0175-6
29. Bhaskar MA, Dash SS, Das S, et al. International Conference on Intelligent Computing and Applications. Springer Singapore; 2019. doi: 10.1007/978-981-13-2182-5
30. Sisodia D, Sisodia DS. Prediction of Diabetes using Classification Algorithms. Procedia Computer Science. 2018; 132: 1578-1585. doi: 10.1016/j.procs.2018.05.122
31. Orabi KM, Kamal YM, Rabah TM. Early Predictive System for Diabetes MellitusDisease. Proceedings of the Industrial Conference on Data Mining; July 2017; New York, USA. Springer. pp. 420–427.
32. Baliunas DO, Taylor BJ, Irving H, et al. Alcohol as a Risk Factor for Type 2 Diabetes. Diabetes Care. 2009; 32(11): 2123-2132. doi: 10.2337/dc09-0227
33. Vazquez G, Duval S, Jacobs DR, et al. Comparison of Body Mass Index, Waist Circumference, and Waist/Hip Ratio in Predicting Incident Diabetes: A Meta-Analysis. Epidemiologic Reviews. 2007; 29(1): 115-128. doi: 10.1093/epirev/mxm008
34. Odegaard AO, Koh WP, Butler LM, et al. Dietary Patterns and Incident Type 2 Diabetes in Chinese Men and Women. Diabetes Care. 2011; 34(4): 880-885. doi: 10.2337/dc10-2350
35. Smith AD, Crippa A, Woodcock J, et al. Physical activity and incident type 2 diabetes mellitus: a systematic review and

dose–response meta-analysis of prospective cohort studies. Diabetologia. 2016; 59(12): 2527-2545. doi: 10.1007/s00125-016-4079-0

36. Pan A, Wang Y, Talaei M, et al. Relation of active, passive, and quitting smoking with incident type 2 diabetes: a systematic review and meta-analysis. Lancet Diabetes Endocrinol. 2015; 3(12): 958-967. doi: 10.1016/S2213-8587(15)00316-2
37. Juneja A, Juneja S, Kaur S, et al. Predicting Diabetes Mellitus With Machine Learning Techniques Using Multi-Criteria Decision Making. International Journal of Information Retrieval Research. 2021; 11(2): 38-52. doi: 10.4018/ijirr.2021040103
38. Tigga NP, Garg S. Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science. 2020; 167: 706-716. doi: 10.1016/j.procs.2020.03.336
39. Priyanka S, Kavitha C, Kumar MP. Deep Learning based Approach for Prediction of Diabetes. 2023 2nd International Conference for Innovation in Technology (INOCON); 3 March 2023. doi: 10.1109/inocon57975.2023.10101241