

Machine Learning-Based Approaches for Type 2 Diabetes Mellitus Classification Early-stage Identification and Prediction

Satish Singh Mekale, Institute of Engineering & Management Kolkata, University of Engineering and Management, Kolkata, India
Maumita Chakraborty, Institute of Engineering & Management Kolkata, University of Engineering and Management, Kolkata, India
Chiradeep Mukherjee, Institute of Engineering & Management Kolkata, University of Engineering and Management, Kolkata, India

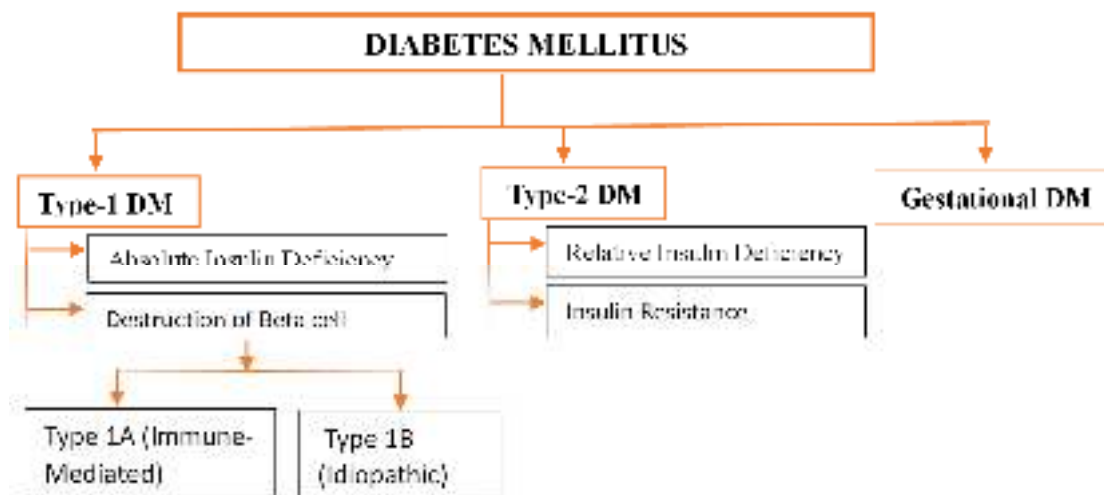
Abstract: Diabetes mellitus (DM) is a leading cause of death because it worsens the consequences of other potentially fatal conditions, such as those affecting the heart, kidneys, and nerves. High blood glucose levels, brought about by insulin obstruction or resistance and insufficient insulin secretion by the pancreas, are characteristic of the chronic condition known as Type 2 Diabetes Mellitus (T2DM). The present work provides an analysis and evaluation of the Machine Learning (ML)-Based Diabetes Mellitus Type 2 Classification and Prediction. Numerous classifiers, including Random Forests (RF), Decision Trees (DT), Support Vector Machines (SVM), Extreme Gradient Boosting (XG-Boost), Logistic Regression (LR), and Naive Bayes (NB), have been used to categorize cases of diabetes. In this study, the author research on many types of Diabetes Mellitus (DM), as well as the symptoms and treatment options for diabetes mellitus. This research covers 50 states, the District of Columbia, and three U.S. territories, has also undergone thorough analysis by the author and it include 253,680 samples with 21 feature variables, feature engineering will be applied to it for better model design. experimental analysis and evaluation, having 768 records and 9 attributes. According to the findings of the experiment, Bagged Decision Tree (BDT) schemes have a better level of accuracy (99.14%), Precision (98.98%), Recall (99.32%), and F-score (99.14%) than other methods.

Keywords: Diabetes Mellitus, Machine Learning, Type 2 Diabetes, Symptoms

Introduction

DM is a condition that is regarded to be chronic and has been affecting individuals of all ages. The precise reason that individuals are becoming sick is still a mystery. On the other hand, some of the variables or causes include becoming older, having a history of the illness in the family, having other diseases in the family, being pregnant, having fluctuating levels of glucose, blood pressure, and so on [1]. Diabetes is a condition that can be treated with prescription; however, a full cure with medications is not yet attainable. The four primary types of diabetes include type-1 diabetes, type-2 diabetes, gestational diabetes, and pre-diabetes [2]. Figure 1 depicts the many ways in which diabetes mellitus can be categorized [3].

Figure 1. Diabetic disease classification



Some other varieties are categorized under these four main ones. When the functionality of the insulin-releasing cell is compromised, it becomes incapable of producing insulin, a condition known as "Type-1 diabetes" or "insulin-dependent diabetes" develops [4]. In the case of "Type 2" diabetes, the body's insulin production is insufficient [5]. This often occurs at an average age of 40 or later. Most cases of "gestational diabetes" manifest themselves during a woman's pregnancy. The last of the four basic types, "prediabetes," is a situation when levels of blood sugar are greater not quite high enough but higher than usual to be categorised as type-2 diabetes [6-7].

In the present study of this work, we discuss an ML model for predicting T2D. The input data instance is classified by the prediction models as usual(non-diabetic), diabetes, or prediabetes depending on the model's output [8]. We examined the accuracy of several prediction models, including those based on LR, SVM, RF, XG-Boost, and DT.

Diabetes Mellitus

Diabetes is portrayed by raised degrees of blood glucose (hyperglycemia), increased thirst (polydipsia), and excessive hunger (polyphagia). It ranks among the most prevalent metabolic disorders, and its prevalence continues to rise worldwide. [9-10]. Within only 34 years, the quantity of individuals with diabetes has significantly increased over the years, with the worldwide populace impacted ascending from 109 million of every 1982 to 425 million out of 2014. Additionally, the prevalence of diabetes among individuals aged 18 and older has also shown a notable rise, escalating from 4.7% in 1980 to 8.5% in 2014. In 2030, Diabetes is predicted to overtake smoking 7th in terms of causes of death, according to a fact given by the World Health Organisation (WHO). [11].

Classification of Diabetes Mellitus

Four main groups make up the current and most frequently recognized categorization of diabetes (Table 1).

Type 1 diabetes mellitus

Type 1 diabetes, otherwise called insulin-reliant or adolescent diabetes is characterized by beta cell loss brought on by an autoimmune reaction, which frequently results in complete insulin deficiency. T1DM is characterised by the presence of antibodies against glutamic acid decarboxylase, islet cells, or insulin indicates the occurrence of autoimmune

mechanisms leading to the loss of beta cells. Insulin therapy is necessary for people with type 1 diabetes need to direct their glucose levels within the normal range [12].

Type 2 diabetes mellitus

Diabetes mellitus, as characterized by the WHO, is an ongoing metabolic problem described by raised blood glucose levels that steadily make harm the eyes, veins, kidneys, heart, and nerves. The hormone insulin, produced by your pancreas, acts as a crucial key that enables the entry of blood sugar into the cells of your body, serving as a source of fuel. Insulin resistance, which causes more than 90% of all examples of diabetes mellitus, insulin resistance refers to the condition where cells do not respond adequately to insulin.[13]. Hyperglycemia occurs when the illness progresses to the point that insulin secretion is inadequate help keep glucose levels in balance. T2DM patients are frequently characterised as overweight or having a larger amount of body fat, particularly in the abdomen. Due to global increases in obesity, unhealthy habits such as high-calorie diets, and ageing populations, the incidence and prevalence of T2DM have quadrupled. [14-15].

Gestational Diabetes Mellitus (GDM)

Women who acquire diabetes mellitus when pregnant are classified into the operational categorization of gestational diabetes mellitus (rather than a pathophysiologic disease). GDM is a classification given to pregnant women who either acquire Type 1 diabetes mellitus or who's asymptomatic T2DM is detected during pregnancy. Most cases of GDM appear in pregnant women during their third trimester [Baynes].

Table 1. Classification of DM [16-17]

Classification	Essential Features
T1DM	Death of Type 2 pancreatic beta cells due to the immune system Absolute or significant insulin insufficiency Consists of LADA (Adult-Onset Autoimmune Diabetes) Type B (idiopathic) diabetes was included
T2DM	Beta cell death and malfunction together Insulin resistance varies in severity
Specific types	Dissimilar peoples Beta cell malfunction or loss is the most common cause. Because of inherent or environmental factors
GDM	Pregnancy-related onset

Sign and Symptoms of DM

Diabetes can trigger a variety of symptoms and indicators, and there could be no symptoms at all. Polyuria, polydipsia, fatigue, and frequent urination are just a few of the common symptoms, increased desire, Loss of appetite, Exhaustion, and Weakness, symptoms such as rapid breathing (kussmaul), lack of interest or focus (hyperventilation), Symptoms include as skin redness and itching, a tingling or numb feeling in the extremities, Reduced visibility, Migraine, Recurring Illnesses, Minor cuts and scrapes, nausea, vomiting, and abdominal pain (often misdiagnosed as the common cold)[18]. Type 1 diabetes often progresses rapidly and sensationally, but T2DM symptoms can be mild or non-existent, making this form of diabetes harder to differentiate from Table 2.

Table 2. Type 1 and Type 2 diabetes sufferers exhibit varied clinical characteristics [19].

Features	Type 1	Type 2
Treatment	Insulin	Weight reduction, thiazolidinedione, metformin, insulin
Insulin sensitivity N	Normal	Reduced
Glucagon in Plasma	heightened, suppressible	High, suppressible resistance
Body mass	Low (wasted) to normal	Obese
Glucose in Plasma	Increased	Increased
Insulin in Plasma	Low or absent	the beginning, normal to high
Age of onset	typically under 20 years	Typically, older than 30 years

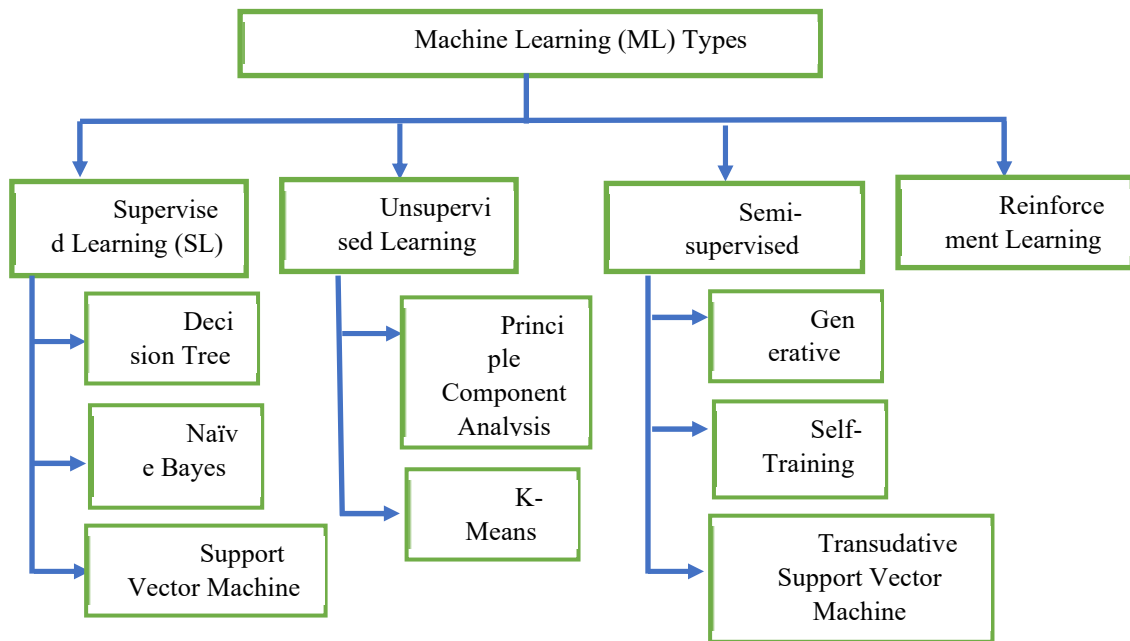
Brief Description of ML Classification Techniques

In this part, the authors offered a concise description of the machine-learning (ML) strategies that are utilized in the treatment of diabetes mellitus.

Machine Learning

ML is useful to instruct machines in data management. The end goal of ML is to improve data. There have been a lot of investigations into how to give robots the ability to learn on their own. To solve the problem, several mathematicians and computer scientists use various methods. Figure 2 shows off a few of them.

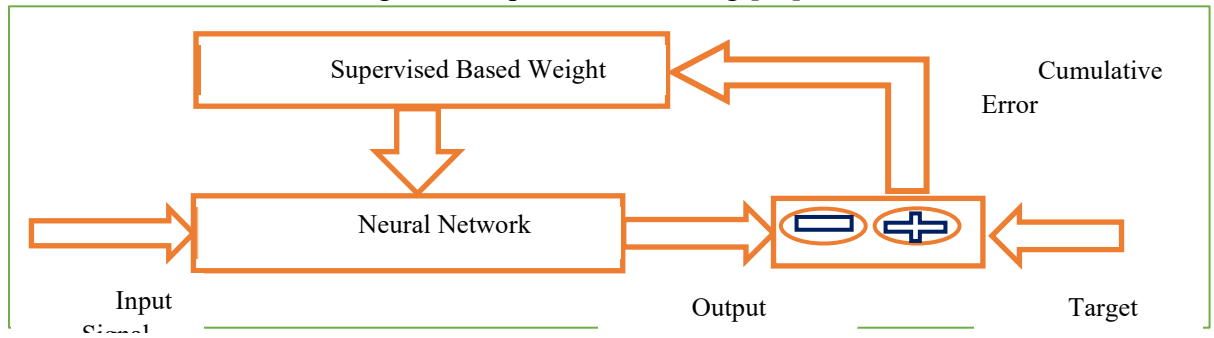
Figure 2. Various types of ML techniques [20-23]



Supervised Learning (SL)

A function that translates input to output is often learned via ML from sample input-output pairs [24]. It infers a function from marked preparing information and a bunch of preparing models. SL is implemented in a task-driven setting, where certain outputs are sought from a given set of inputs [25]. Common supervised tasks include "classification," which involves grouping data, and "regression," which involves transforming data. Figure 3 indicates the SL model.

Figure 3: Supervised Learning [26]



Unsupervised Learning (UL)

UL describes how a network can be trained to represent a subset of input plans in a manner that reliably reproduces the mathematical course of action of the full arrangement of information plans [27]. The calculation gets as information sources a lot of information and a quality of every perception except not the planned result. UL is often used (for example, clustering) to divide images into two groups or clusters based on intrinsic qualities of the images like as color, size, shape, and furthermore. Figure4 describes the training network connection.

Figure 4. Unsupervised Learning [28]

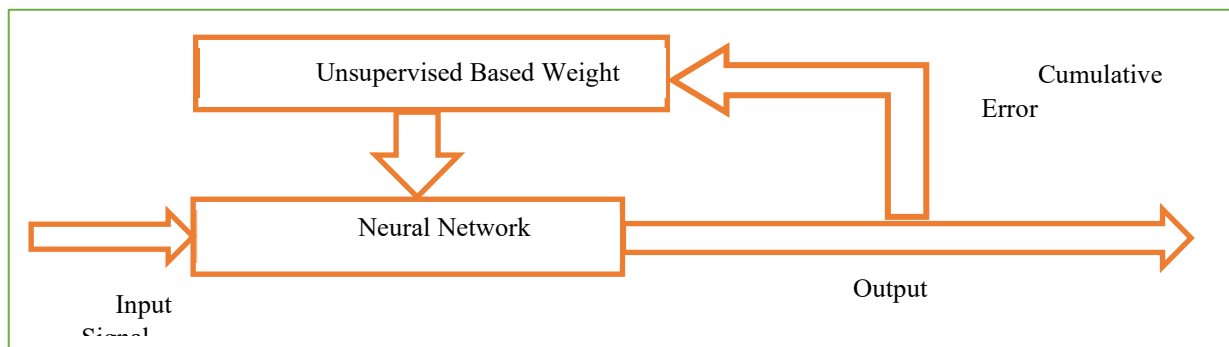


Table 3. Types of ML [29-30]

Types	Types of data provided	Examples of model
SL	The data are annotated with the various classifications and results. This enables the algorithm to derive a forecast more accurately.	Various ML techniques, such as RF, NB networks, and ANN, are included in the category.
UL	As the data is unlabeled, it can be called agnostic. The program detects critical correlations in datasets.	Various Methods like Principal Component Analysis (PCA), ranked clustering, and k-means are all included.
SSL	A method of learning that can be applied to datasets having both labeled and unlabeled columns.	This includes outcomes and classifications that are both labeled and unlabeled. Mostly used for image and voice recognition.
RL	A technique related to human psychology employs reward criteria.	It is increasingly widespread in diagnostic imaging, analytics-based illness screening, and medication choice.

LR, SVM, RF, Decision tree, NBC, and KNN are Base of the most current ML approaches today.

Naive Bayes (NB) Classifier

The Bayesian classifier, often identified as the NBC, is centered on Bayes' theorem. One method of text categorization is to allocate to a certain document d the class $c^* = \arg \max_c p(c|d)$.

It develops the NB classifier by first reflecting that by Bayes' rule,

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)}, \quad (1)$$

where $P(d)$ plays no role in choosing c^* . To approximate the term $P(d|c)$, NBC it by supposing the f_i 's are provisionally individual given d 's class:

$$P_{NB}(c|d) := \frac{P(c)(\prod_{i=1}^m P(f_i|c)^{n_i(d)})}{P(d)} \quad (2)$$

Random Forest (RF)

This is one of the ensemble approaches that is exclusively utilized to increase the success and accuracy of ML algorithms in artificial intelligence. Scikit-learn uses Gini Significance to estimate the importance of each node in a DT, given as there exist only a pair of child nodes:

$$n_j = w_j C_j - w_{\text{left}(j)} C_{\text{left}(j)} - w_{\text{right}(j)} C_{\text{right}(j)} \quad (1)$$

Logistic Regression

The ML method of LR is called SL. The sigmoid function (logarithmic) provides the foundation since it takes a real number and returns a value between 0 and 1. An N -sample training dataset is given to the algorithm during the training phase. Each instance has certain X properties and a Y label that specifies the way it should be categorized. As a result of the training process, the system generates a model that can be used to label data that was not involved in the training set [31].

Support Vector Machine

The SVM is another popular cutting-edge ML method. Its primary function is to organize data. As an algorithm, SVM relies on the concept of margin calculation. Essentially, it establishes artificial barriers between economic classes. Classes are separated from the margin as much as possible to provide the lowest possible classification error in the process [32].

K- Nearest Neighbor (KNN) Classifier

The KNN approach can be used for both kind of problem statement either regression or classification issues, however, it is more frequently utilized to address the latter in the corporate world. Its primary key benefit is the Sentences can be translated in a short period of time. In Figure 8, the points (2.5, 7) and (5.5, 4.5) would be allocated to any one of the clusters. The KNN employs the Euclidean distance function to determine how far away a new data point is from the pre-existing data points.

Extreme gradient boosting (XG-Boost)

It is heavily utilized for the Higgs sub-signal identification challenge on Kaggle competition. It has lately gained widespread notice for its exceptional efficiency and great forecast accuracy. Indeed, XG-Boost is an upgraded Gradient Boosted Decision Tree (GBDT)

method, which is generally utilized around classification and regression and consists of several decision trees [33].

Related Work

Buyrukoğlu et al., (2022)[34] Using the advantages of the Correlation Matrix (CM), Sequential Forward Selection (SFS), and Heatmap (HM), a novel hybrid feature engineering selection approach that was developed to categorise the most helpful qualities for the diagnosis of diabetes. To assess the best features, classifiers such as SVM, Random Forests, and ANN were employed. Accuracy (99.2%), F-measure (99.2%), precision (99.4%), recall (99.2%), and area under the curve (AUC) (99.8%) were the five metrics with the best results for ANN.

Ganieet al., (2022)[35] developed a novel system based on ensemble learning for early detection of T2DM using dietary and physical activity data. Bagging, Boosting, and Voting are just a few of the many ensemble learning methods used. In terms of accuracy (99.54%), precision (99.53%), recall (94.83%), specificity (99.51%), F1-score (99.55%), misclassification rate (MCR) (0.86%), and AUC (99.07%), the BDT performed better than the other approaches. The proposed methodology could be applied to the early detection of diabetes in the healthcare industry.

Lu et al., (2022)[36] create an ensemble of novel patient networks and ML techniques for illness prediction using a actual-world administrative claim dataset for T2DM. We use eight ML models (LR, SVM, DT, XG-Boost, KNN, NB, RF, and ANN) to make predictions about the risk of chronic illness based on the extracted characteristics. Extensive studies reveal that the AUC for the proposed framework using ML classifiers is between 0.79 and 0.91.

Zhanet al., (2022)[37] a T2DM classification model based on a Hybrid Random Forests Linear Model (HRFLM) was developed, which combines LR with 4 RF. Standardising data and using exploratory data analysis (EDA) in the data preparation process. The author additionally used the optional ML model to automatically adjust some of the parameters. Accuracy is measured in terms of percentage for each of the suggested approaches, and it is shown that the HRFLM provides the maximum accuracy (86.4%).

Rashidet al., (2022)[38] examine the feasibility of using a cutting-edge ML strategy for making such forecasts based on patient demographics, clinical characteristics, and test results. Automatic clinical testing for microvascular problems in diabetic patients was then achieved by the creation of an ML strategy predicated on LR, RF, and SVM algorithms. Gender, albumin-creatinine ratio, and diastolic blood pressure are used. RF obtained the highest prediction accuracies for testing Cardiac Autonomic Neuropathy (CAN); using Haemoglobin A1C testing criteria include microalbuminuria, history of smoking, and Diabetic Peripheral Neuropathy (DPN); and using these same factors for testing diabetic Retinopathy (RET) (84.38%).

Ganieet al., (2022)[39] conducted a performance assessment of numerous indicators for the early prediction of T2DM and compared the results from different supervised ML classifiers. The experimental work has been effectively applied to six ML classification techniques (LR, NB, SVM, DT, RF, and ANN). When compared to other classifiers, RF has the highest success rate (93.75%) for predicting type 2 diabetes. In addition, the class biases in the dataset have been removed using the ten-fold cross-validation approach.

Liu et al., (2022)[40] construct reliable ML-based methods for predicting T2DM risk in Chinese retirees. It built the models for making predictions using four different ML methods: LR, DT, RF, and XG-Boost. The effectiveness of the model was measured

concerning several various execution measurements, like AUC, exactness, explicitness, and responsiveness, are utilized to assess the exhibition of the model. The best results were found using the XG-Boost model (AUC = 78.05%, sensitivity = 64.52%, specificity = 75.77%, accuracy = 75.03%).

Al-Sadi et al., (2023)[41] provide AI and ML prediction models for improved accuracy and reduced processing time in detecting type 2 diabetes mellitus (T2DM) in the Omani population-NN, SVM, NB, DT, RF, Linear Discriminant Analysis, and ANN are only some of the ML algorithms that were implemented in MATLAB. For the Oman data set, the RF and DT models outperformed all other methods with an accuracy of 98.38%. The precision achieved using the Oman dataset was an increase of 9.1 percentage points greater than with Pima Indian Diabetes (PID) using the same model and the same no. of features. Table 4 indicates the evaluation of the literature of review.

Table 4. Comparison of Literature of Review

Authors	Techniques Used	Outcomes
Buyrukoğlu et al., (2022) [34]	SVM, RF, ANN	The findings from this study demonstrate that the utilization of the ANN algorithm to generate a feature set using the suggested hybrid feature selection technique yields the highest classification accuracy (99.1%).
Ganieet al., (2022) [35]	BDT, RF, LR, SVM, and DT	The BDT was the most reliable method overall, with a 99.41% success rate, a 99.13% precision rate, a 95.13% recall rate, a 99.11% specificity rate, an F1-score of 99.15, an MCR of 0.86%, and a ROC curve of 99.07%.
Lu et al., (2022) [36]	LR, KNN, SVM, NB, DT, RF, XG-Boost and ANN	In comparison to the other seven models, RF has the greatest accuracy at 84.95%, followed by ANN and XG-Boost at 82.52% each. Obtaining an AUC in the experimental range of 0.79 to 0.91 demonstrates the efficiency of the models.
Zhanet al., (2022) [37]	HRFLM, RF, LR and XG-Boost	The findings show that while utilizing the PIMA Indian Dataset, the HRFLM generated the greatest accuracy with a percentage of 86.4% when the testing set was adjusted by 10%, demonstrating the efficacy of HRFLM.
Rashidet al., (2022) [38]	LR, RF, and SVM	At weights of 1.05 and 0.9, we discovered that RF was the most accurate model. The precision, awareness, and explicitness of the model were determined to be 98.68%, 98.48%, and 100% respectively.
Ganieet al., (2022) [39]	LR, NB, SVM, DT, RF and ANN	According to the experimental results, RF has the most accurate classifier overall on the dataset, with values of (99.71% for the training set) and (93.80% for the testing set).
Liu et al., (2022) [40]	LR, DT, RF, and XG-Boost	The XG-Boost model had the greatest AUC value (0.7901 on the test set), and its sensitivity (0.6453), specificity (0.7577), and accuracy (0.7503) were the best.
Al Sadi et al., (2022) [41]	K-NN, SVM, NB, DT, RF, LDA, and ANN	After training the model with seven different classification methods, they found that RF and DT both achieved the best accuracy (98.37%).

Experiment setup

The accompanying presentation or metrics for evaluation are used to assess the effectiveness of the suggested technique.

The formulas for these are as follows:

Accuracy

The percentage of all correctly predicted values to the sum of the true positive and true negative.

$$\text{Accuracy} = \frac{T_P + T_N}{T_{\text{Total_Images}}} \times 100(1)$$

Recall (or Sensitivity) and Precision

It can be defined as true positive rate (TPR). It can be as the ratio of genuine positives prediction to total test observation. The recall is determined as follows:

$$\text{Recall} = \left(\frac{\text{true positive}}{\text{true positive} + \text{false negative}} \right) \quad (2)$$

The precision is calculated as

$$\text{Precision} = \left(\frac{\text{true positive}}{\text{true positive} + \text{false positive}} \right) \quad (3)$$

F-score

It can be defined as the combination of together the precision metrics and recall metrics, it is harmonic mean of the formula to evaluate the performance of ML model.

$$F - \text{score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \quad (4)$$

Data Availability and Description

The dataset used to support are available at <https://www.kaggle.com/uciml/pima-indians-diabetes-database>.

Also, study one more dataset that is available at <https://www.cdc.gov/brfss/index.html>

Figure 5. describe PIMA dataset Heat map correlation Matrix and Figure 6. Describe BFRSS dataset Heat map correlation Matrix.

Figure 5. Heatmap for PIMA Dataset

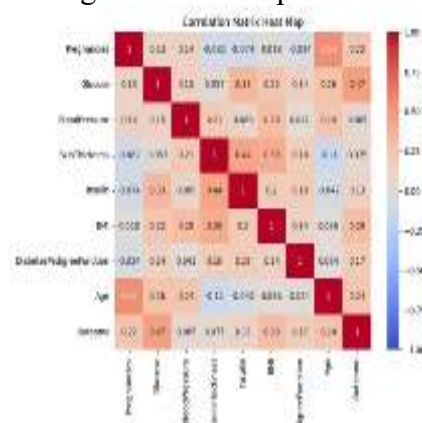
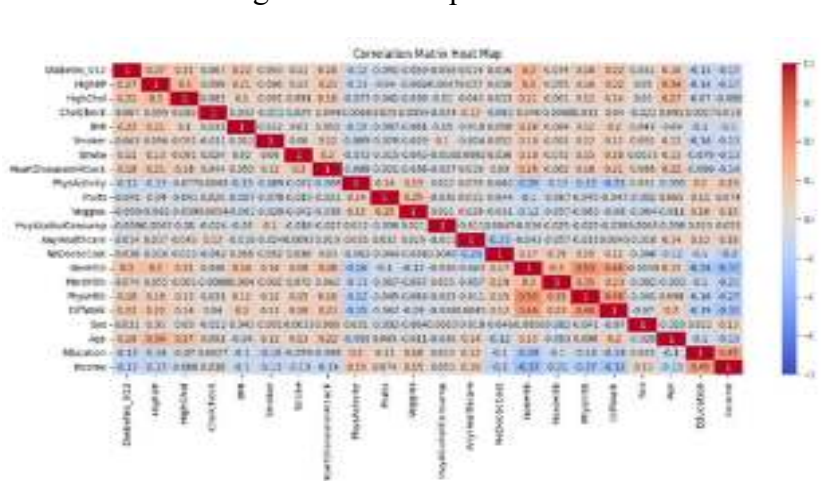


Figure 6. Heatmap for BRFSS Dataset



Comparative Analysis

This section of the review presents a comparative analysis of the literature that has been reviewed previously. It assessed the accuracy of prior reviewed research papers by utilizing diverse datasets. The results of the comparison show that BDT schemes perform better than their competitors in terms of accuracy (98.14%), precision (98.98%), recall (97.32%), and F-score (98.14%). Table 5 indicates the comparative analysis of LOR and Figure 7 shows a

graph of comparative analysis based on accuracy and precision. Figure 8 similarly depicts the graph of a comparison study based on recall and F-score.

Table 5. Comparative analysis of several techniques

Authors [Reference]	Techniques Used	Dataset Used	Accuracy	Precision	Recall	F-score
Buyrukoğlu et al., (2022) [34]	ANN	PID	99.1	99.3	99.1	99.1
Ganieet al., (2022) [35]	BDT	Lifestyle datasets from geographical regions	99.14	98.98	99.32	99.14
Lu et al., (2022) [36]	RF	PID	84.95	85.97	84.95	84.79
Zhanet al., (2022) [37]	HRFLM	PIMA Indian	94.9	68.6	97.3	86
Rashidet al., (2022) [38]	RF	PID	98.67	98.48	100	99.2
Ganieet al., (2022) [39]	RF	Training and testing	93.79	89.1	95.3	92.1
Liu et al., (2022) [40]	XG-Boost	Training and testing	75.03	75.77	64.52	69.69
Al Sadi et al., (2022) [41]	RF	PID	98.37	84.1	98.01	90.5

Figure 7. Graph of several techniques based on Accuracy and Precision

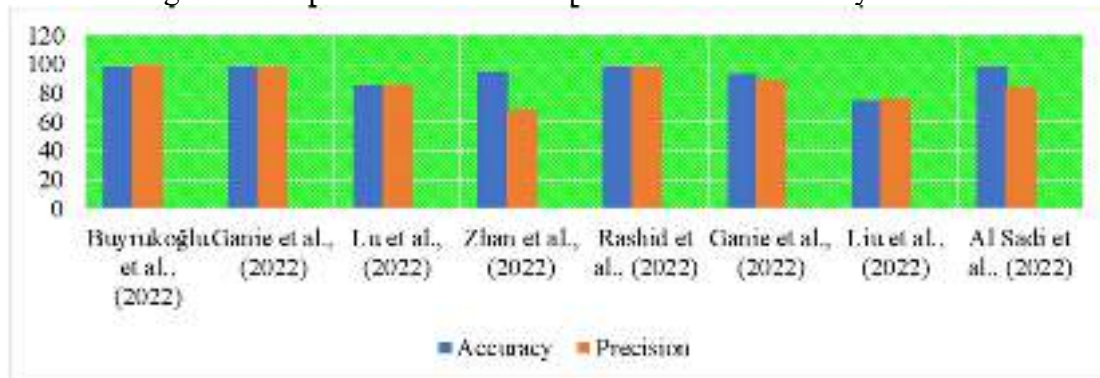
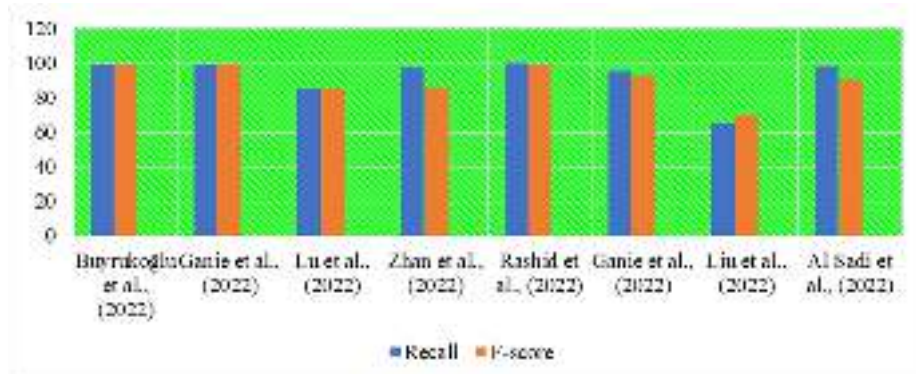


Figure 8. Graph of several techniques based on Recall and F-score.



Conclusion

The early detection of diabetes risk is one of the world's major health challenges. In India, diabetes affects over 30 million people and puts many more at risk. Consequently, timely identification and intervention are crucial in order to forestall diabetes and the related wellbeing complications it entails. Our study presents an analysis and evaluation of the ML-Based Diabetes Mellitus Type 2 Classification and Prediction. Six distinct Strategies of classification were implemented, and the results of these strategies were evaluated using various metrics of statistics. The several datasets were subjected to a variety of tests. The author has also carefully examined the 253,680 samples with 21 feature variables of the Behavioural Risk Factor Surveillance System (BRFSS). All of the models that were used, which included six different ML algorithms, gave excellent findings related to certain metrics, including precision, F-score, accuracy, and recall, among others. According to the findings of the experiment on the PIMA Indian Diabetes dataset having 768 records and 9 attributes the BDT schemes have a greater level of accuracy (99.14%), Precision (98.98%), Recall (99.32%), and F-score (99.14%) than the alternatives.

This finding has the potential to be utilized in the future to diagnose any other illness. This research does not rule out the possibility of more research and development in the future, including the use of other ML algorithms to diagnose diabetes or any other condition. Deep learning models can be utilized on PIMA and BRFSS in future work to predict other diseases.

REFERENCES

Journal

- [1] Dash, S., Shakyawar, S. K., Sharma, M., and Kaushik, S. (2019). Big data in healthcare: management, analysis, and prospects. *J. Big Data* 6, 54. doi: 10.1186/s40537-019-0217-0
- [2] Nibareke, T., and Laassiri, J. (2020). Using big data-machine learning models for diabetes prediction and flight delay analytics. *J. Big Data* 7, 78. doi: 10.1186/s40537-020-00355-0
- [3] Ng, Li Ching, and Manish Gupta. "Transdermal drug delivery systems in diabetes management: A review." *Asian Journal of pharmaceutical sciences* 15, no. 1 (2020): 13-25.
- [4] Martinsson, J., Schliep, A., Eliasson, B., and Mogren, O. (2020). Blood glucose prediction with variance estimation using recurrent neural networks. *J. Healthc. Inform. Res.* 4, 1–18. doi: 10.1007/s41666-019-00059-y
- [5] Wang, F., Stiglic, G., Obradovic, Z., and Davidson, I. (2015). Guest editorial: special issue on data mining for medicine and healthcare. *Data Min. Knowl. Disc.* 29, 867–870. doi: 10.1007/s10618-015-0414-1
- [6] Mujumdar, A., and Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Proc. Comp. Sci.* 165, 292–299. doi: 10.1016/j.procs.2020.01.047
- [7] Ahamed, B. Shamreen, Meenakshi Sumeet Arya, and Auxilia Osvin Nancy V. "Prediction of Type-2 Diabetes Mellitus Disease Using Machine Learning Classifiers and Techniques." *Frontiers in Computer Science* 4 (2022).
- [8] Deberneh, Henock M., and Intaek Kim. "Prediction of type 2 diabetes based on a machine learning algorithm." *International Journal of environmental research and public health* 18, no. 6 (2021): 3317.

Website

- [9] World Health Organization (WHO). Global Report on Diabetes; WHO: Geneva, Switzerland, 2017; Available online: <http://www.who.int/diabetes/global-report/en/> (accessed on 22 September 2018).

Journal

- [10] Mathers, C.D.; Loncar, D. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med.* 2006, 3, e442.
- [11] Alam, Saruar, Md Kamrul Hasan, Sharif Neaz, Nazmul Hussain, Md Faruk Hossain, and Tania Rahman. "Diabetes Mellitus: insights from Epidemiology, biochemistry, risk factors, diagnosis, complications, and comprehensive management." *Diabetology* 2, no. 2 (2021): 36-50.
- [12] Baynes, Habtamu W. "Classification, pathophysiology, diagnosis and management of diabetes mellitus." *J diabetes metab* 6, no. 5 (2015): 1-9.
- [13] Galicia-Garcia, Unai, Asier Benito-Vicente, Shifa Jebari, Asier Larrea-Sebal, Haziq Siddiqi, Kepa B. Uribe, Helena Ostolaza, and César Martín. "Pathophysiology of type 2 diabetes mellitus." *International Journal of molecular sciences* 21, no. 17 (2020): 6275.
- [14] Chatterjee, S.; Khunti, K.; Davies, M.J. Type 2 diabetes. *Lancet* 2017, 389, 2239–2251.
- [15] NCD Risk Factor Collaboration. Worldwide trends in diabetes since 1980: A pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* 2016, 387, 1513–1530.
- [16] American Diabetes Association. 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2021. *Diabetes Care* 2021, 44, S15–S33.
- [17] Masini, Matilde, Mara Suleiman, Michela Novelli, Lorella Marselli, Piero Marchetti, and Vincenzo De Tata. "Mast Cells and the Pancreas in Human Type 1 and Type 2 Diabetes." *Cells* 10, no. 8 (2021): 1875.
- [18] Mungrue, Kameel, Lesley-Ann Roper, and Tamika Chung. "Assessment of weight loss in the management of patients with type 2 diabetes mellitus in primary care in Trinidad." *J Diabetes Metab* 2, no. 120 (2011): 2.
- [19] Nisha, K. "Diabetes mellitus: Classification, symptoms, and management: A review." *Wiener klinische Wochenschrift* (2016).
- [20] Sarker, Iqbal H. "Machine learning: Algorithms, real-world applications, and research directions." *SN computer science* 2, no. 3 (2021): 160.

Conference

- [21] Dutta, Nabanita, S. Umashankar, VK Arun Shankar, Sanjeevikumar Padmanaban, Zbigniew Leonowicz, and Patrick Wheeler. "Centrifugal pump cavitation detection using machine learning algorithm technique." In *2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe)*, pp. 1-6. IEEE, 2018.

Journal

- [22] Sundararajan, Karpagam, Lalit Garg, Kathiravan Srinivasan, Ali Kashif Bashir, Jayakumar Kaliappan, Ganapathy Pattukandan Ganapathy, Senthil Kumaran Selvaraj, and T. Meena. "A contemporary review on drought modeling using machine learning approaches." *CMES-Computer Modeling in Engineering and Sciences* 128, no. 2 (2021): 447-487.
- [23] Gupta, Manju M., and Akshat Gupta. "Survey of artificial intelligence approaches in the study of anthropogenic impacts on symbiotic organisms—a holistic view." *Symbiosis* 84, no. 3 (2021): 271-283.
- [24] Han, J. "M. Kamber in J. Pei, Data mining: Concepts and techniques: Concepts and techniques, 3. izd." (2011).

- [25] Sarker, Iqbal H., A. S. M. Kayes, Shahriar Badsha, Hamed Alqahtani, Paul Watters, and Alex Ng. "Cybersecurity data science: an overview from a machine learning perspective." *Journal of Big Data* 7 (2020): 1-29.

Conference

- [26] Yusof, Yusman, HM Asri H. Mansor, and Adizul Ahmad. "Utilizing unsupervised weightless neural network as autonomous states classifier in the reinforcement learning algorithm." In *2017 IEEE 13th International Colloquium on Signal Processing & its Applications (CSPA)*, pp. 264-269. IEEE, 2017.
- [27] Dike, Happiness Ugochi, Yimin Zhou, Kranthi Kumar Deveerasetty, and Qingtian Wu. "Unsupervised learning based on the artificial neural network: A review." In *2018 IEEE International Conference on Cyborg and Bionic Systems (CBS)*, pp. 322-327. IEEE, 2018.

Journal

- [28] Mahesh, Batta. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR)*. [Internet] 9 (2020): 381-386.
- [29] Shameer, Khader, Kipp W. Johnson, Benjamin S. Glicksberg, Joel T. Dudley, and Partho P. Sengupta. "Machine learning in cardiovascular medicine: are we there yet?" *Heart* 104, no. 14 (2018): 1156-1164
- [30] Johnson, Kipp W., Jessica Torres Soto, Benjamin S. Glicksberg, Khader Shameer, Riccardo Miotto, Mohsin Ali, Euan Ashley, and Joel T. Dudley. "Artificial intelligence in cardiology." *Journal of the American College of Cardiology* 71, no. 23 (2018): 2668-2679.
- [31] Waring, Jonathan, Charlotta Lindvall, and Renato Umeton. "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare." *Artificial intelligence in medicine* 104 (2020): 101822.
- [32] Mahesh, Batta. "Machine learning algorithms-a review." *International Journal of Science and Research (IJSR)*. [Internet] 9 (2020): 381-386.
- [33] Xu, Yonghui, Xi Zhao, Yinsheng Chen, and Zixuan Yang. "Research on a mixed gas classification algorithm based on extreme random tree." *Applied Sciences* 9, no. 9 (2019): 1728.
- [34] BUYRUKOĞLU, Selim, and Ayhan AKBAŞ. "Machine Learning based early prediction of type 2 diabetes: a new hybrid feature selection approach using correlation matrix with heatmap and SFS." *Balkan Journal of Electrical and Computer Engineering* 10, no. 2 (2022): 110-117.
- [35] Ganie, Shahid Mohammad, and Majid Bashir Malik. "An ensemble machine learning approach for predicting type-II diabetes mellitus based on lifestyle indicators." *Healthcare Analytics* 2 (2022): 100092.
- [36] Lu, Haohui, Shahadat Uddin, Farshid Hajati, Mohammad Ali Moni, and Matloob Khushi. "A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus." *Applied Intelligence* 52, no. 3 (2022): 2411-2422.

Conference

- [37] Zhan, Weiyi. "A Comparative Study on Machine Learning Based Type 2 Diabetes Mellitus Prediction." In *2022 International Conference on Computer Science, Information Engineering and Digital Economy (CSIEDE 2022)*, pp. 859-871. Atlantis Press, 2022.

Journal

- [38] Rashid, Mamunur, Mohanad Alkhodari, Abdul Mukit, Khawza Iftekhhar Uddin Ahmed, Raqibul Mostafa, Sharmin Parveen, and Ahsan H. Khandoker. "Machine Learning for Screening Microvascular Complications in Type 2 Diabetic Patients Using

- Demographic, Clinical, and Laboratory Profiles." *Journal of Clinical Medicine* 11, no. 4 (2022): 903.
- [39] Ganie, Shahid Mohammad, and Majid Bashir Malik. "Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus." *International Journal of Medical Engineering and Informatics* 14, no. 6 (2022): 473-483.
- [40] Liu, Qing, Miao Zhang, Yifeng He, Lei Zhang, Jingui Zou, Yaqiong Yan, and Yan Guo. "Predicting the risk of incident type 2 diabetes mellitus in Chinese elderly using machine learning techniques." *Journal of Personalized Medicine* 12, no. 6 (2022): 905.
- [41] Al-Sadi, Khoula, and Wamadeva Balachandran. "Prediction Model of Type 2 Diabetes Mellitus for Oman Prediabetes Patients Using Artificial Neural Network and Six Machine Learning Classifiers." *Applied Sciences* 13, no. 4 (2023): 2344.