Multimedia Tools and Applications
https://doi.org/10.1007/s11042-023-16745-4

Diabetes prediction model using machine learning techniques

Sandip Kumar Singh Modak[1] · **Vijay Kumar Jha**[2]

Received: 23 May 2023 / Revised: 17 August 2023 / Accepted: 31 August 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
Diabetes has emerged as a significant global health concern, contributing to various severe complications such as kidney disease, vision loss, and coronary issues. Leveraging machine learning algorithms in medical services has shown promise in accurate disease diagnosis and treatment, thereby alleviating the burden on healthcare professionals. The field of diabetes forecasting has rapidly evolved, offering the potential for early intervention and patient empowerment. To this end, our study presents an innovative diabetes prediction model employing a range of machine learning techniques, including Logistic Regression, SVM, Naïve Bayes, and Random Forest. In addition to these foundational techniques, we harness the power of ensemble learning to further enhance prediction accuracy and robustness. Specifically, we explore ensemble methods such as XGBoost, LightGBM, CatBoost, Adaboost, and Bagging. These techniques amalgamate predictions from multiple base learners, yielding a more precise and resilient final prediction. Our proposed framework is developed and trained using Python, utilizing a real-world dataset sourced from Kaggle. Our methodology is rigorously examined through performance evaluation metrics, including the confusion matrix, sensitivity, and accuracy measurements. Among the ensemble techniques tested, CatBoost emerges as the most effective, boasting an impressive accuracy rate of 95.4% compared to XGBoost's 94.3%. Furthermore, CatBoost's higher AUC-ROC score of 0.99 reinforces its potential superiority over XGBoost, which achieved an AUC-ROC score of 0.98.

Keywords Diabetes · Machine learning · SVM · Random Forest and Naïve Bayes

✉ Sandip Kumar Singh Modak
modaknit@gmail.com

Vijay Kumar Jha
vkjha@bitmesra.ac.in

[1] Department of Computer Science & Engineering, Sarla Birla University, Ranchi, Jharkhand, India

[2] Department of Computer Science & Engineering, Birla Institute of Technology, Mesra, Ranchi, India

Published online: 05 October 2023

1 Introduction

Diabetes is a persistent condition described by a raised blood glucose level. Diabetes causes moderate kidney, eye, and heart harm after some time [1]. Early detection of diabetes is a difficult task. Diabetes sickness can be categorized into three classes: Type 1 diabetes, also known insulin-dependent diabetes or juvenile diabetes, happens when the body's safe framework harms insulin-delivering cells, ending insulin creation [2]. More than 90% of DM cases are those of type 2 diabetes. As of late, the frequency of type 2 diabetes has been expanding significantly every year [3]. As per the most recent overview delivered by the Global Diabetes League in 2019, diabetes has a commonness of 9.3%, and it affects roughly 463 million grown-ups around the world. It is normal that the quantity of affected people will arrive at 578 million (10.2%) by 2030 and 700 million (10.9%) by 2045 [4]. Diabetes is especially hazardous for pregnant ladies, and unborn youngsters are probably going to be impacted by this infection. By and large, assuming the glucose level in the blood transcends the ordinary worth, the individual is viewed as diabetic. This is because of the powerlessness of the pancreas in the human body to play out its errand completely. The individual's glucose rises on the off chance that the pancreas can't use the insulin it delivers or doesn't make enough of it [5]. A strategy called, Predictive Analysis consolidates a different type of machine learning algorithm. Data mining techniques and statistical method that utilizes current and past information to track down information and foresee future occasions. By applying Predictive Analysis on medical services information, huge choices can be taken and forecasts can be made [6]. Data mining and machine learning have been developing, reliable, and supporting tools in the medical domain in recent year [7]. Machine Learning algorithms were introduced to automate the working model of healthcare systems to improve the disease prediction accuracy. Hadoop cluster based distributed computing framework was introduced to support efficient processing and storing large data in cloud environment. The new machine learning algorithm was introduced in hadoop based clusters for performing the diabetes prediction [8]. In recent years, plenty of methods have been proposed and published for diabetes prediction [9]. There were various ML-based systems used to classify and predict of diabetic disease like linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), naïve Bayes (NB), support vector machine (SVM), artificial neural network (ANN), feed-forward neural network (FFNN), Adaboost (AB), decision tree (DT), J48, random forest (RF), Gaussian process classification (GPC), logistic regression (LR), and k-nearest neighborhood (KNN) and so on [10]. The rest of this paper is summarized as follows: in Section 2 discuss different machine learning technique; Section 3 gives insights about related work; Section 4 features about the propose mechanism in details; Section 5 discuss result and analysis of the proposed mechanism; Section 6 highlights the comparative analysis of different machine learning algorithm. And finally concludes the papers in Section 7.

2 Machine learning technique

Data Mining and Artificial Intelligence (AI) plays an important role in the prediction of diabetes. With the continuous development of artificial intelligence and data mining technology, researchers begin to consider using machine learning and deep learning techniques to search for the characteristics of diabetes. Machine learning techniques

can find implied pathogenic factors in virtue of analyzing and using diabetic data, with a high stability and accuracy in diabetes diagnosis. Therefore, machine learning techniques which can find out the reasonable threshold risk factors and physiological parameters provide new ideas for screening and diagnosis of diabetes [11]. Diabetes is a very serious disease that, if not treated properly and on time, can lead to very serious complications, including death. This makes diabetes, one of the main priorities in medical science research, which in turn generates huge amounts of data. Constantly increasing volumes of data are very well suited to be processed using data mining that can readily handle them. Using data-mining methods in diabetes research is one of the best ways to utilize large volumes of available diabetes-related data for extracting knowledge. Both descriptive (association and clustering) and predictive (classification) data-mining methods are used in the process. These data-mining methods are different from traditional statistic approaches in many ways [12]. A large area of Artificial Intelligence is Machine Learning. By this, it means that AI uses machine learning algorithms for its intelligent behavior. A computer is said to learn from some task if the error continuously decreases and if it matches the performance as desired. Machine learning will study algorithms that will perform the task of extraction automatically. Machine learning comes from statistics, but it is not actually. Similar to AI, machine learning also has a very broad scope. Deep Learning is a subset of machine learning. It works in the same way on the machine just like how the human brain processes information. Like a brain can identify the patterns by comparing it with previously memorized patterns, deep learning also uses this concept.

- **Classification:** Classification is a machine learning capability that can be relegated to target classifications for data set objects. This approach can be utilized as a prehandling step before to putting away information in the classification model. It gives the prediction of Yes or No, for example, “Is this tumour cancerous?”, and “Does this cookie meet our quality standards?” Common classification approaches include artificial neural network, back propagation, decision tree, support vector machines, Naive Bayes classifier, K-Nearest Neighbors (K-NN), Random forest [13]. Classification is used to classify data into predefined categorical class labels. “Class” in classification, is the attribute or feature in a data set, in which users are most interested. It is defined as the dependent variable in statistics. To classify data (or records), a classification algorithm creates a classification model consisting of classification rules. For example, banks have constructed classification models to categorize the bank loan and mortgage applications into risky or safe. In the medical field, classification can be used to help define medical diagnosis and prognosis based on symptoms and health conditions.
- **Regression:** Regression analysis consists of a set of machine learning methods that allow us to predict a continuous outcome variable (y) based on the value of one or multiple predictor variables (x). Briefly, the goal of regression models is to build a mathematical equation that defines y as a function of the x variables. Next, this equation can be used to predict the outcome (y) on the basis of new values of the predictor variables (x). Linear regression is the most simple and popular technique for predicting a continuous variable. It assumes a linear relationship between the outcome and the predictor variables. Regression analysis is a way of fitting a “best” line through a series of observations. With “best” line we mean that it is fitted in such a way that it minimizes the sum of squared differences between the observations and the line itself. Regression analysis can also help to make predictions. For example, if we have estimated a regression model using data on sales, prices, and promotional activities, the results from this

regression analysis could provide a precise answer to what would happen to sales if prices were to increase by 5% and promotional activities were to increase by 10% [14].

- **Decision Tree (DT):** Decision trees are one of the most effective methods for data mining; they have been widely used in several disciplines because they are easy to be used, free of ambiguity, and robust even in the presence of missing values. Both discrete and continuous variables can be used either as target variables or independent variables. More recently, decision tree methodology has become popular in medical research [15]. An example of the medical use of decision trees is in the diagnosis of a medical condition from the pattern of symptoms.
- **Knowledge Discovery Dictionary (KDD):** It is a data mining technique, which deals with process of data and discovery of information. It involves information handling, information selection, information preparation, information establishment on data sets, and interpretation of the most effective ways based on observed findings [16]. It comprises of an iterative information integration sequence and acknowledgment of Data Mining designs.
- **Support Vector Machine (SVM):** The support vector machine is a statistical based machine learning technique used in various research area. Due to great advancement and a higher exactness, support vector machine has received the most attention from the machine learning community [17]. Initially, support vector machines is used for class classification problem, but with the rapid development of computer technology, network technology, database technology, it is used for the classification and management of large amounts of information, the classification problem can no longer meet people's needs. This method will be extended to multi-class classification problem. It is currently a hot research topic.
- **K-Nearest Neighbor (KNN):** KNN algorithms are supervised non-parametric learning algorithms that learn the relationship between input and output observations. A new input instance is classified by assigning the output class of the K most similar neighbors, where similarity is defined according to a distance metric. KNN algorithm is used for classifying samples of the data and based on the principle that a sample which has more similar properties with K sample are put into the same category[18]. In a sense, K of samples is recognized for the new sample, and the label of the category which is more repetitive among all these samples is identified as the category of its result.
- **Random Forest (RF):** Random Forest is an ensemble method, which predicts based on the results of a collection of Decision Trees. Resampling using the bootstrap approach is used for the creation of each tree in the "forest." Also, on each node split a subset of features is selected randomly and the selection of the split variable occurs over this subset. The predicted value is the majority vote, for classification, and the average, for regressions [19]. Essentially, there are two parameters for tuning on Random Forest models: mtry–the number of randomly selected features to consider in each split; and ntree – the number of trees in the model.
- **Naive Bayes Classification:** Naive Bayesian classification algorithm (NBC) is one of the classic Bayesian classification algorithms, which has a simple algorithm structure and high computational efficiency. One advantage of a naive Bayes classifier is that it only needs to estimate the necessary parameters (mean and variance of variables) based on a small amount of training data. Due to the assumption of independent variables, only the method of estimating each variable is needed, and the whole covariance matrix is not needed [20].
- **K-Mean Clustering:** Clustering is a process of grouping data objects into disjointed clusters so that the datas in the same cluster are similar, yet datas belonging to differ-

ent cluster differ. The demand for organizing the sharp increasing datas and learning valuable information from data. K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results. K-means is a numerical, unsupervised, non-deterministic, iterative method [21]. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good clustering results.

- **Association Rule:** Association discovery is one of the most common data mining techniques that are used to extract interesting knowledge from large datasets. Much effort has been made to use its advantages for classification under the name of associative classification. Association discovery aims to find interesting relationships between the different items in a database, while classification aims to discover a model from training data that can be used to predict the class of test patterns [22]. Both association discovery and classification rules mining are essential in practical data mining applications and their integration could result in greater savings and convenience for the user.
- **Artificial Neural Network:** ANNs are artificial adaptive systems that are inspired by the functioning processes of the human brain. They are systems that are able to modify their internal structure in relation to a function objective. They are particularly suited for solving problems of the nonlinear type, being able to reconstruct the fuzzy rules that govern the optimal solution for these problems. The base elements of the ANN are the nodes, also called processing elements (PE), and the connections. Each node has its own input, from which it receives communications from other nodes and/or from the environment and its own output, from which it communicates with other nodes or with the environment [23]. Finally, each node has a function f through which it transforms its own global input into output.

2.1 Application of machine learning in healthcare

The increasingly growing number of applications of machine learning in healthcare allows us to glimpse at a future where data, analysis, and innovation work hand-in-hand to help countless patients without them ever realizing it.

- **Identifying Diseases and Diagnosis:** One of the main applications of ML in healthcare is the identification and diagnosis of diseases and ailments which are otherwise considered hard-to-diagnose. This can include anything from cancers which are tough to catch during the initial stages, to other genetic diseases.
- **Smart Health Records:** The main role of machine learning in healthcare is to ease processes to save time, effort, and money. Document classification methods using vector machines and ML-based OCR recognition techniques are slowly gathering steam, such as Google's Cloud Vision API and MATLAB's machine learning-based handwriting recognition technology.
- **Clinical Trial and Research:** Machine learning has several potential applications in the field of clinical trials and research. As anybody in the pharma industry would tell you, clinical trials cost a lot of time and money and can take years to complete in many cases. Applying ML-based predictive analytics to identify potential clinical trial candidates can help researchers draw a pool from a wide variety of data points, such as previous doctor visits, social media, etc.

- **Crowdsourced Data Collection:** Crowdsourcing is all the rage in the medical field nowadays, allowing researchers and practitioners to access a vast amount of information uploaded by people based on their own consent. This live health data has great ramifications in the way medicine will be perceived down the line.
- **Outbreak Prediction:** AI-based technologies and machine learning are today also being put to use in monitoring and predicting epidemics around the world. Today, scientists have access to a large amount of data collected from satellites, real-time social media updates, website information, etc. Artificial neural networks help to collate this information and predict everything from malaria outbreaks to severe chronic infectious diseases.

3 Related work

Yang et al. [24] proposed a novel methodology utilizing machine learning procedures. Improving the prediction model's accuracy and summing up the model's forecasts past a solitary informational collection are essential objectives. The model is parted into two sub-parts: an improved K-mean algorithm and a logistic regression method, the two of which depend on an assortment of pre-handling steps. In order to forecast diabetes risk, Islam et al. [25] need a dataset that incorporates data about individuals who are recently determined to have diabetes or who are in danger of fostering the condition. They utilized a 520-item dataset that was collected from surveys administered to Sylhet Diabetes Hospital patients in Bangladesh. Moreover, they applied Naive Bayes, Logistic Regression, and Random Forest algorithms to predict the model. Similarly, Woldemichael and Menaria [26] introduced a model utilizing machine learning technique to anticipate instances of diabetes, foreseeing whether an individual has diabetes utilizing a back propagation algorithm. Predictions of diabetes were made utilizing various strategy, including J48, naive bayes, and support vector machine. Similarly, a prediction model was designed by Fiarni et al. [27] to estimate the event of three significant entanglements of diabetes in Indonesia, and key variables related with these complexities are recognized. The seven risk factors for diabetes were recognized as age, gender, BMI, family history of diabetes, blood pressure, length of time diabetic, and blood glucose level. Therefore, k-means clustering and the Naive Bayes Tree classification strategies were used to look at this informational collection. Moreover, application programming created by Aldallal et al. [28] is utilized by specialists and other clinical experts to predict the beginning or repeat of persistent illnesses (NCDs). In this undertaking, they used an information mining method that could predict future results. Data about patients from the Bahrain Protection Power Medical Clinic was utilized to test the program. Additionally, Khan et al. [29] go into the area of glycemic management for diabetes and research data mining-based diagnosis and prediction solutions. Likewise, Kavakiotis et al. [30] completed a precise examination of diabetes research utilizing machine learning, data mining approaches, and tools using three distinct data mining organization methods: Naive Bayes (NB), Support Vector Machine (SVM), and Decision Tree. Kumar et al. [31] assessed and examined the imminent ways to forecast the chance of heart disease for diabetic patients based on their predictive accuracy. Moreover, the goal of Mahesh et al. [32]'s blended ensemble learning (EL)-based forecasting system was to identify the best classifier for evaluating clinical outcomes through a standardised set of metrics. In this article, authors suggest an EL that uses Bayesian networks and

radial basis functions. Standard predictive methods, such as K-nearest neighbour (KNN) and logistic regression, are used in Oza and Bokhare [33]. By comparing the different ways that machine learning can be used, a model is proposed to improve performance and measure accuracy. Prediction of diabetes mellitus using data mining is well reviewed by Anil et al. [34]. The goal is to examine and compare the predicted accuracy of the various analytical approaches currently employed in this sector through a study of studies that have used these approaches. Logistic regression analysis with data mining techniques such as decision trees using the J48 and LMT algorithms, Naive Bayes, and Artificial Neural Networks (ANN) was proposed by Paisanwarakiat et al. [35]. The risk of developing diabetes was predicted using Random Forest, KNN, and Support Vector Machine by Arumugam et al. [36]. As seen in the outcomes, the support vector methodology is very trustworthy. A variety of classifiers have been presented, and their structures make it possible to choose the most appropriate one for future data analysis and interpretation. Abdollahi and Moghaddam [37] employed a genetic algorithm-based ensemble training methodology to effectively identify and predict the consequences of diabetes mellitus. Experimental data and genuine data on Indian diabetes from the University of California's website are used in this study. Luo et al. [38] proposed a deep learning method to assist with chronic atrophic gastritis diagnosis using white light images. In the external test set, the diagnostic accuracy of model 1 for detecting gastric antrum atrophy was 0.890. The identification accuracies for the severity of gastric antrum atrophy were 0.773 and 0.590 in the internal and external test sets, respectively. Luo et al. [39] developed a deep learning-based method for COVID-19 pneumonia detection from CT images. In this study, an AI system for the diagnosis of COVID-19 pneumonia based on ResNet-50 and U-Net were developed. In the test set, the sensitivity of the deep learning model in diagnosing normal cases, CAP, and COVID-19 patients was 98.03%, 89.28%, and 92.15%, respectively. Zamzami et al. [40] proposed a Machine learning algorithms for smart and intelligent healthcare system in Society 5.0. In this work author explain and compare the different algorithms of ML which could be helpful in detecting different disease at earlier stage. It summarizes the algorithms and different steps involved in ML to extract information for betterment of the society which is already exposed to the world of data. Ahmed et al. [41] explain the use of deep learning approach for augmented detection of Coronavirus disease. The proposed deep learning modalities are based on convolutional neural network (CNN) and convolutional long short-term memory (ConvLSTM). The proposed deep learning modalities are tested on both X-ray and CT images as well as a combined dataset that includes both types of images. They achieved an accuracy of 100% and an F1 score of 100% in some cases. The simulation results reveal that the proposed deep learning modalities can be considered and adopted for quick COVID-19 screening. Kuldeep et al. [42] proposed a Cloud-Based Predictive Model for the Detection of Breast Cancer. In this research, random forests, logistic regression, decision trees, and SVM are employed, and the authors assess the performance of various algorithms using confusion measures and AUROC to choose the best machine learning model for breast cancer prediction. Precision, recall, accuracy, and specificity are used to calculate results. Confusion matrix is based on predicted cases. The ML model's performance is evaluated. For simulation, the authors used the Wisconsin Dataset of Breast Cancer (WDBC). Through experiments, it can be seen that the SVM model reached 98.24% accuracy with an AUC of 0.993, while the logistic regression achieved 94.54% accuracy with an AUC of 0.998.Hammad et al. [43] proposed a deep learning Models for Arrhythmia Detection in IoT Healthcare Applications. In this paper, novel convolutional neural network (CNN) and convolutional long

short-term (ConvLSTM) deep learning models (DLMs) are presented for automatic detection of arrhythmia for IoT applications. Overall accuracies of 97%, 98%, 94% and 91% are obtained on spectrograms of MIT-BIH dataset, compressed MIT-BIH dataset, PhysioNet 2016 dataset, and PhysioNet 2018 dataset, respectively.

4 Proposed model

The proposed model of diabetic prediction using different machine learning technique is comprise of selecting of dataset, prepare datasets for training purpose, extraction of feature or feature extraction which include elimination of unwanted features, apply different machine learning algorithm for classification, validation of the model and finally test the model. The proposed model is shown in Fig. 1.

- **Selection of Dataset:** In this proposed model we use the dataset named "Diabetic2" which is downloaded from kaggle an online community of data scientist and machine learning engineers.
- **Data preprocessing:** Clean the dataset by removing any duplicates, outliers, or missing values. Also, normalize or standardize the data to ensure that all features are on the same scale.
- **Feature selection:** Use feature selection techniques to identify the most important features that contribute to diabetes prediction. This could involve using correlation analysis, principal component analysis (PCA), or other methods.
- **Model selection:** Choose an appropriate machine learning algorithm for the problem at hand. Popular choices for classification problems like this include logistic regression, decision trees, random forests, support vector machines (SVMs), and neural networks.
- **Model training:** Split the dataset into training and testing sets. Use the training set to train the model using the chosen algorithm and the selected features.
- **Model evaluation:** Use the testing set to evaluate the performance of the model. Common evaluation metrics for binary classification problems like this include accuracy, precision, recall, and F1 score.

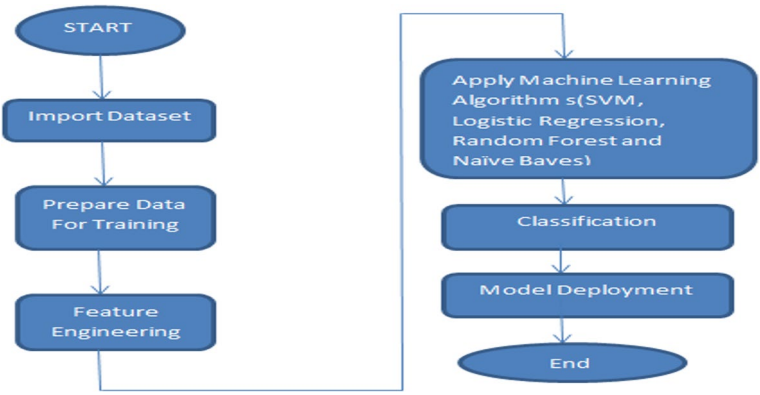

Fig. 1 Proposed model

- **Model optimization:** Fine-tune the model parameters and/or try different algorithms or features to improve the model performance.
- **Model deployment:** Once the model has been optimized and validated, it can be deployed in a real-world setting to make predictions on new patient records.

Description of Dataset used:

Dataset Name: Diabetic2 Dataset
Total Number of Instance: 5000
Features Available:
PatientID: Unique Ids of patient
Pregnancies: Number of times pregnant
PlasmaGlucose: Plasma glucose concentration a 2 h in an oral glucose tolerance test
BloodPressure: Diastolic blood pressure (mm Hg)
SkinThickness: Triceps skin fold thickness (mm)
SerumInsulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)^2)
DiabetesPedigreeFunction: Diabetes pedigree function
Age: Age (years)
Diabetic: Class variable (0 or 1)

Histogram is one kind of plot which helps in effective visualization of numeric attributes. It helps in understanding the distribution of a numeric data into series of intervals, also termed as 'bins'. The focus of histogram is to plot ranges of data values (acting as 'bins'), the number of data elements in each range will depend on the data distribution. Based on that, the size of each bar corresponding to the different ranges will vary. The Histogram representation of diabetic dataset is shown in Fig. 2. From the figure it is found that the nature of 'Pregnancies' attribute is Right Skewed, whereas PlasmaGlucose in Symmetrical and Unmoral in nature and BloodPressure in Bimodal in nature.

A box plot is an extremely effective mechanism to get a one-shot view and understand the nature of the data. The box plot (also called box and whisker plot) gives a standard visualization of the five-number summary statistics of a data, namely minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. BoxPlot Representation of Diabetic Dataset is shown in Fig. 3. A scatter plot helps in visualizing bivariate relationships, i.e. relationship between two variables. It is a two-dimensional plot in which points or dots are drawn on coordinates provided by values of the attributes. Scatter Plot Representation of Diabetic Dataset is shown in Fig. 4.

Proposed Algorithm:

✓ START : Diabetic Dataset, Machine_Learning_model
✓ IMPORT : Dataset = Diabetic2.csv
✓ TRAIN MODEL: Training Dataset: Training_data
✓ EXTRACT: Features -> from training data
✓ APPLY : Machine Learning Model -> new_data=Machine_Learning_model(Training_data)
✓ CLASSIFY:new_data
✓ DEPLOYMENT : new_data-> dataset
✓ END

The datasets are gathered from the information base. In stage two, the information will be pre-handled, which will incorporate information cleaning, mixing, and changing. By

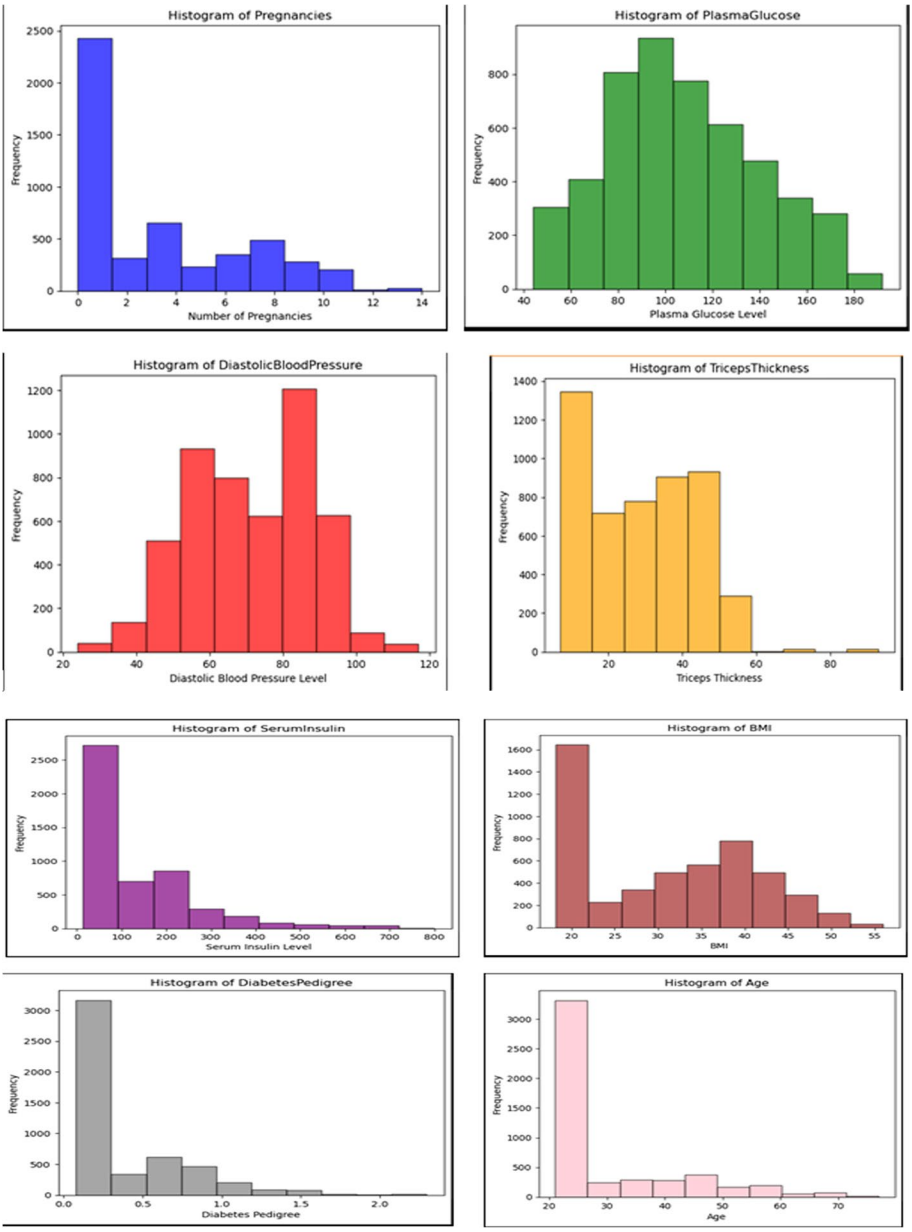

Fig. 2 Histogram representation of diabetic dataset

utilizing ensemble learning, we can achieve better precision when compared with other calculations.

The information was acquired from Kaggle. It was gathered and, along these lines, the data was inputted as instructing tests and back-to- back investigated to supply a decent model. Data variety is the arrangement of accommodating pertinent data that is assembled

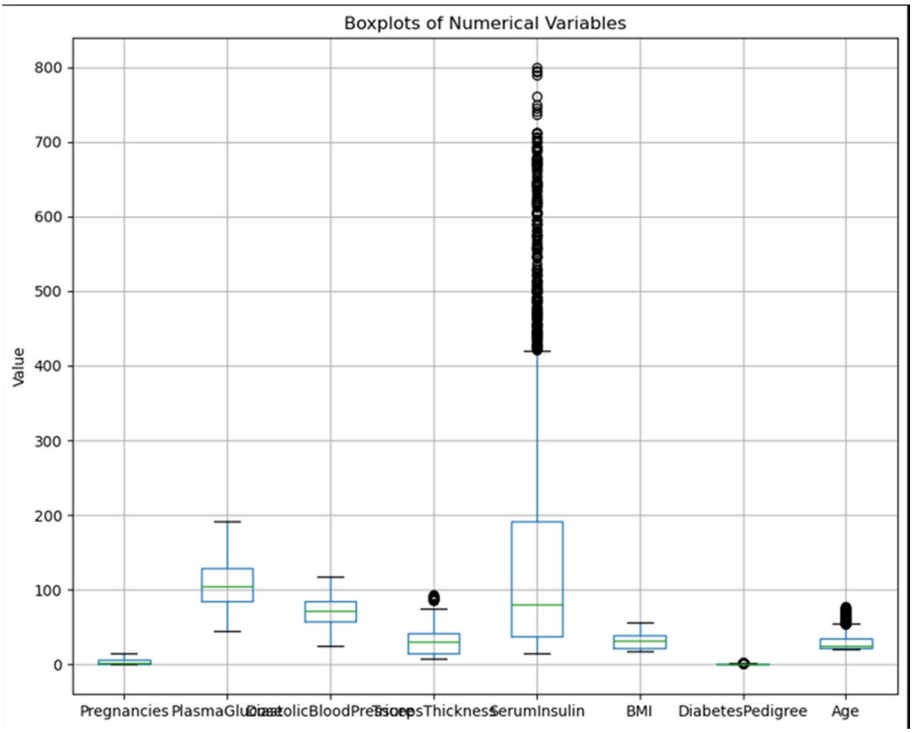

Fig. 3 Boxplot representation of diabetic dataset

through an exploitation question measure. The information is separated in a very specific way, with a number of very specific categories. Pair Plot Representation of Diabetic Dataset is shown in Fig. 5.

4.1 Feature engineering steps

Feature Engineering is the process of selecting, transforming, and creating relevant features from raw data to improve the performance of machine learning models. It is a critical step in the machine learning (ML) pipeline and has a significant impact on the model's ability to learn patterns and make accurate predictions.

- Data Preprocessing: Load the diabetic dataset into a Pandas DataFrame.
- Data Cleaning: In this phase we address the missing value in the diabetic dataset. In our case we don't find any null value in diabetic dataset, so skip this phase.
- Categorical Encoding: This phase is deal with to convert categorical variables into numerical representations using one-hot encoding.
- Feature Scaling: Feature scaling is a data preprocessing technique used to standardize or normalize the numerical features in a dataset to a common scale. There are two

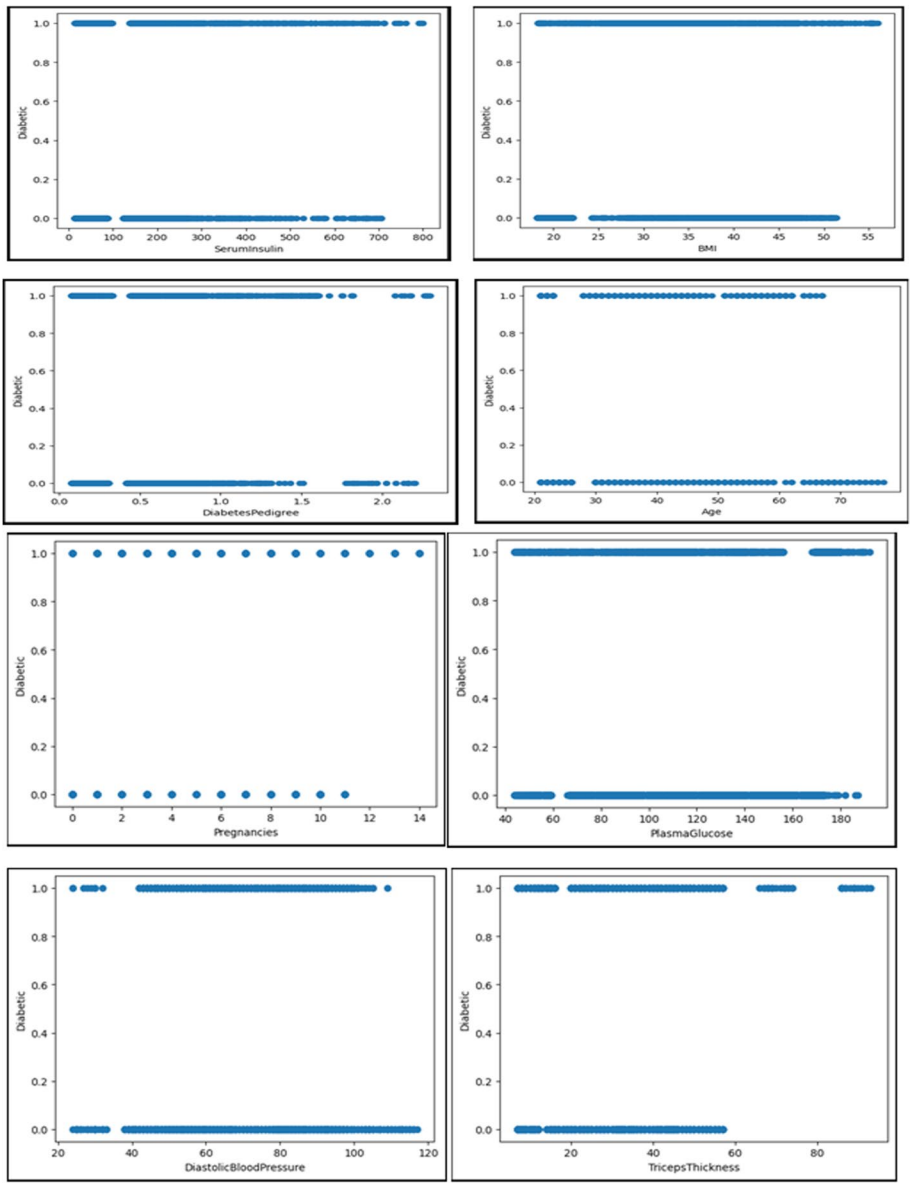

Fig. 4 Scatter plot representation of diabetic dataset

common methods of feature scaling; these are Normalization (Min–Max Scaling) and Z-score scaling. Min–Max Scaling technique is used in this work for scaling purpose.

- Model Training: Split the dataset into training and testing sets and training our machine learning model using 'Diabetic' as target variable and rest features as fea-

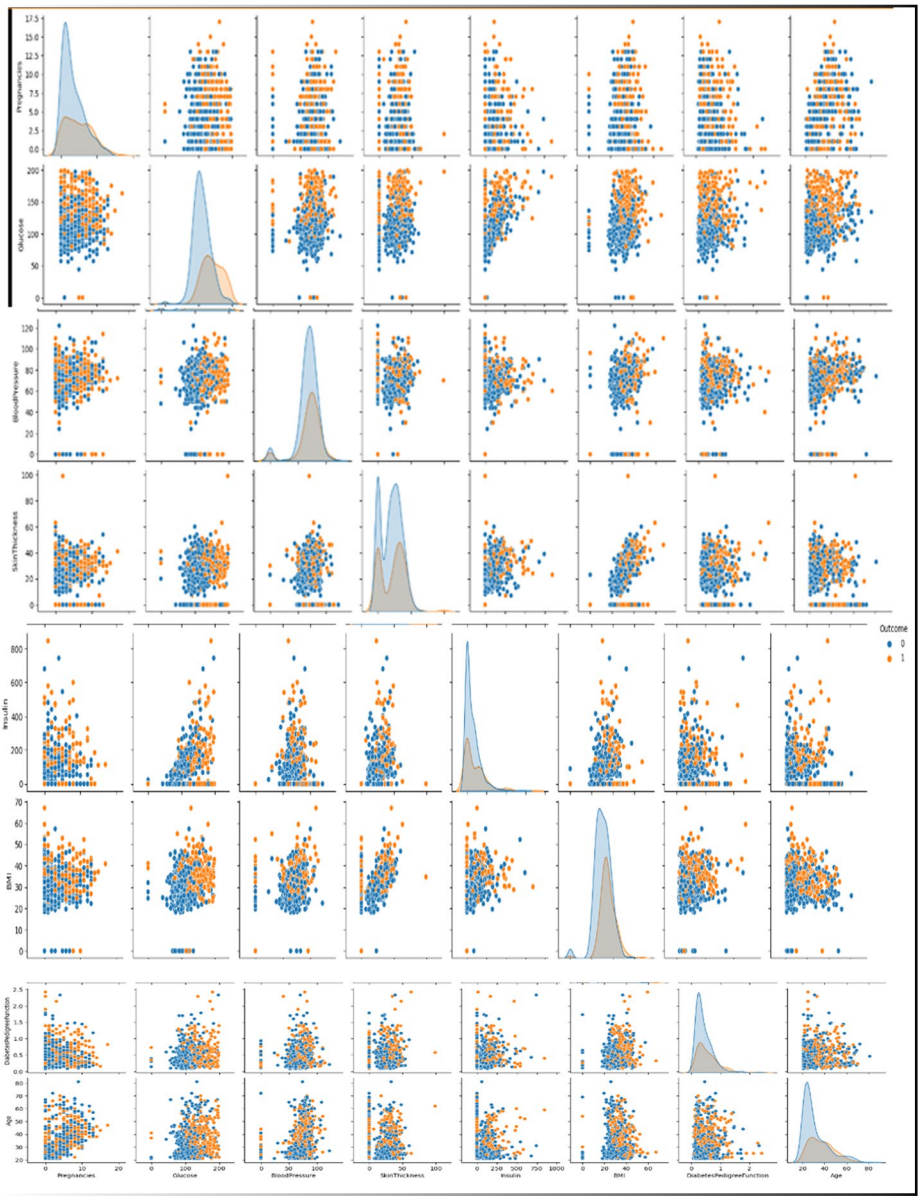

Fig. 5 Pair plot representation of diabetic dataset

tures variable. The whole dataset is divided into two parts, first 80% of the datasets consider as training data and remaining 20% is reserve for testing.

- Model evaluation**:** Use the testing set to evaluate the performance of the model. Common evaluation metrics for binary classification problems like this include accuracy, precision, recall, and F1 score.

- Model optimization**:** Fine-tune the model parameters and/or try different algorithms or features to improve the model performance. In this work we use hyper parameter tuning and Cross validation (tenfold) techniques for the improvement of the model in terms of Accuracy.

4.2 Clinical application of the proposed work

The clinical application of an AI model for diabetes detection can offer significant value, even if traditional diagnostic methods exist. Here are a few reasons why the proposed model could still have clinical application value:

- Early Detection and Prevention: AI models have the potential to analyze a wider range of data points beyond blood glucose levels. This can include genetics, lifestyle factors, and other biomarkers that may contribute to early detection of diabetes risk. Detecting diabetes at an earlier stage allows for timely intervention and lifestyle changes that could potentially prevent or delay the onset of the disease.
- Risk Stratification: AI models can help stratify patients based on their risk of developing diabetes. This can aid healthcare professionals in identifying high-risk individuals who might benefit from more intensive monitoring, lifestyle modifications, or preventive measures.
- Personalized Medicine: AI models can provide personalized recommendations based on an individual's unique health profile. This could include tailored dietary and exercise advice, medication management, and other interventions to manage diabetes effectively.
- Comprehensive Data Analysis: AI can process and analyze large volumes of data from various sources, potentially uncovering patterns and correlations that might be missed by human clinicians. This comprehensive analysis can lead to more accurate and reliable predictions.
- Supporting Clinical Decision-Making: AI models can serve as decision support tools for healthcare professionals. They can provide additional insights and information that aid doctors in making informed clinical decisions, ultimately leading to better patient care.
- Reducing Workload: While simplification is important, AI can also help streamline and automate certain aspects of clinical practice, allowing healthcare providers to focus on more complex tasks and patient interactions.

5 Result

In this section, results and analysis have been discussed. The proposed mechanism is implemented using Python and Jupyter notebooks. It is an open-source programming language. The execution of a programme is fast in Python. It can provide in-built library files to run users' programs. Python is best suited for data-driven machine learning algorithms. The heat map of the diabetic dataset is shown in Fig. 6.

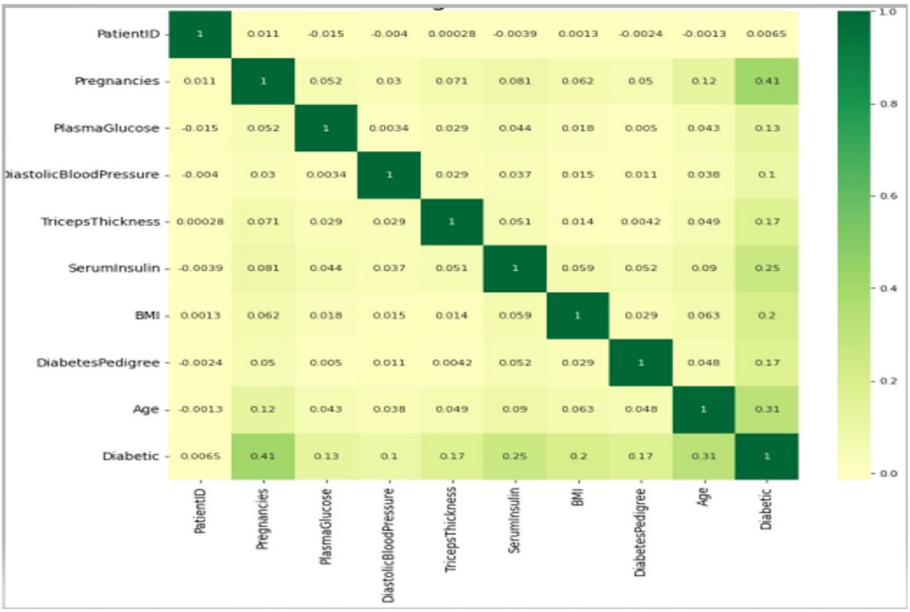

Fig. 6 Heatmap of diabetic dataset

5.1 Performance evaluation

In the implementation sections, different types of performance operators in conjunction with validation have been used without an in detail explanation of how these operators really function.

- **Confusion Matrix:** Classification performance is best described by an aptly named tool called the confusion matrix or truth table. Understanding the confusion matrix requires becoming familiar with several definitions.

  The predicted class is Y, and the actual class is also Y - this is a True Positive or TP
  The predicted class is Y, and the actual class is N - this is a False Positive or FP
  The predicted class is N, and the actual class is Y - this is a False Negative or FN
  The predicted class is N, and the actual class is also N - this is a True Negative or TN

- **Sensitivity:** Sensitivity is the ability of a classifier to select all the cases that need to be selected. A perfect classifier will select all the actual Y's and will not miss any actual Y's. In other words it will have no FNs. In reality, any classifier will miss some true Y's, and thus, have some FNs. Sensitivity is expressed as a ratio (or percentage) calculated as follows:

$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

- **Accuracy:** Accuracy is defined as the ability of the classifier to select all cases that need to be selected and reject all cases that need to be rejected. For a classifier with 100% accuracy, this would imply that FN = FP = 0, Accuracy is given by

$$\text{Accuracy} = (\text{TP} + \text{TN})/\text{TP} + \text{TN} + \text{FP} + \text{FN}$$

Finally, error is simply the complement of accuracy, measured by (1- accuracy)

5.2 Description of dataset used

To examine the proposed system, we use the diabetes dataset downloaded from Kaggle. The dataset contains information on the upsides of patients having an age greater than 20. The anticipated result, i.e., positive or negative, has been determined based on different boundaries, for example, pregnancy, sugar level, BMI, and so on.

So here we have nine rules that can be utilized to gauge and anticipate diabetes. These rules are pregnancy (that is, the condition of women when giving birth to a child), quantity of glucose, the number or value of blood pressure, the thickness of human skin, body mass ratio, also called BMI, the function of the diabetes pedigree, the year to which the human body has gone through disease, and the age. These are all collectively going to predict diabetes. The collective dataset of features is going to give crucial information about the estimation of diabetes in a human being. All the features of our dataset play a significant role in the prediction of diabetes. Calculating the importance of each feature will help us determine how relevant each feature is to finding the output of our model. Table 1 shows the comparison of ensemble learning technique (XGBoost, LightGBM, CatBoost, Adaboost and Bagging), SVM, RF, logistic regression, and naive bayes algorithm in terms of their confusion matrix. A confusion matrix is a table that is used to evaluate the performance of a classification model. It presents the true and predicted classifications of the model's predictions on a test dataset.In the case of CatBoost Algorithm the number of true positives is 320, meaning 320 instances that actually belong to the positive class were correctly predicted as positive. The number of true negatives is 634, meaning 634 instances that actually belong to the negative class were correctly predicted as negative. The number of false positives is 24, meaning 24 instances that actually belong to the negative class were wrongly predicted as positive. The number of false negatives is 22, meaning 22 instances that actually belong to the positive class were wrongly predicted as negative. Likewise, in the case of XGBoost the number of true positives is 308, meaning 308 instances that actually belong to the positive class were correctly predicted as positive. The number of true negatives is 635, meaning 635 instances that actually belong to the negative class were correctly

Table 1 Performance comparison based on confusion matrix

Model	Confusion matrix	Model	Confusion matrix
XGBOOST	[[635 23] [34 308]	Random forest	[[626 32] [40 302]
LightGBM	[[636 22] [31 311]	SVM	[[595 63] [127 215]
CatBoost	[[634 24] [22 320]	Naïve bayes	[[530 128] [89 253]
Adaboost	[[636 22] [32 310]	Logistic Regression	[[582 76] [137 205]
Bagging	[[618 40] [51 291]		

predicted as negative. The number of false positives is 23, meaning 23 instances that actually belong to the negative class were wrongly predicted as positive. The number of false negatives is 34, meaning 34 instances that actually belong to the positive class were wrongly predicted as negative.

Furthermore, Table 2 depicts the comparison of different machine learning techniques in terms of accuracy. Hyperparameter tuning and cross-validation are essential techniques in machine learning to improve the performance and generalization of models. Hyperparameter tuning involves systematically searching for the optimal hyperparameter values to achieve the best model performance. Grid search, random search, Bayesian optimization, and genetic algorithms are some common methods for hyperparameter tuning. Whereas Cross-validation is a technique used to assess a model's performance and generalization on unseen data. The idea is to divide the dataset into multiple subsets (folds) to train and evaluate the model iteratively.

Cross-validation helps provide a more reliable estimate of a model's performance compared to using a single train-test split.The most common form of cross-validation is k-fold cross-validation, where the data is split into k subsets (or folds). We have used tenfold cross validation technique for the assessment of the model performance. From Table 2 it is clear that CatBoost gives the highest accuracy of 95.4% which has outperformed the performance of XGBoost with an accuracy rate of 94.3%. After the hyperparameter tuning, the accuracy rate of XGBoost has reached up to 95.8% with the best hyperparameter: {'subsample': 1.0, 'n_estimators': 200, 'max_depth': 3, 'learning rate': 0.1, 'colsample_bytree': 0.9}, and the accuracy rate of CatBoost has reached up to 95.7% with the best hyperparameter: {'depth': 4, 'iterations': 300, 'learning rate': 0.1}. Similarly, after tenfold cross-validation, the accuracy rate of XGBoost has attained up to average accuracy of 0.9494 with Cross-Validation Accuracy Scores: [0.948 0.942 0.958 0.928 0.95 0.958 0.962 0.948 0.954 0.946] and CatBoost has attained up to average accuracy of 0.9558 with Cross-Validation Accuracy Scores: [0.958 0.94 0.962 0.938 0.958 0.97 0.968 0.95 0.958 0.956]. It is crucial to provide a prompt and correct diagnosis while dealing with disorders like diabetes. If proper care is not taken, a delay in diabetes diagnosis can have devastating effects on health. Therefore, the accuracy of machine learning algorithms used to anticipate a patient's status must be maximized. A false negative has a much greater cost impact than a false positive. If the subject is given a false diagnosis, they may try to relax despite the significance of their situation.

Table 2 Comparison based on accuracy

Algorithm	Accuracy	Accuracy (Hyperparameters)	Accuracy (tenfold)
XGBOOST	0.943	0.958	0.9494
LightGBM	0.947	0.948	0.952
CatBoost	0.954	0.957	0.9558
Adaboost	0.946	0.9442	0.9454
Bagging	0.909	0.93	0.919
Random forest	0.928	0.93	0.9304
SVM	0.81	0.863	0.8076
NB	0.783	0.81	0.793
LR	0.787	0.787	0.79

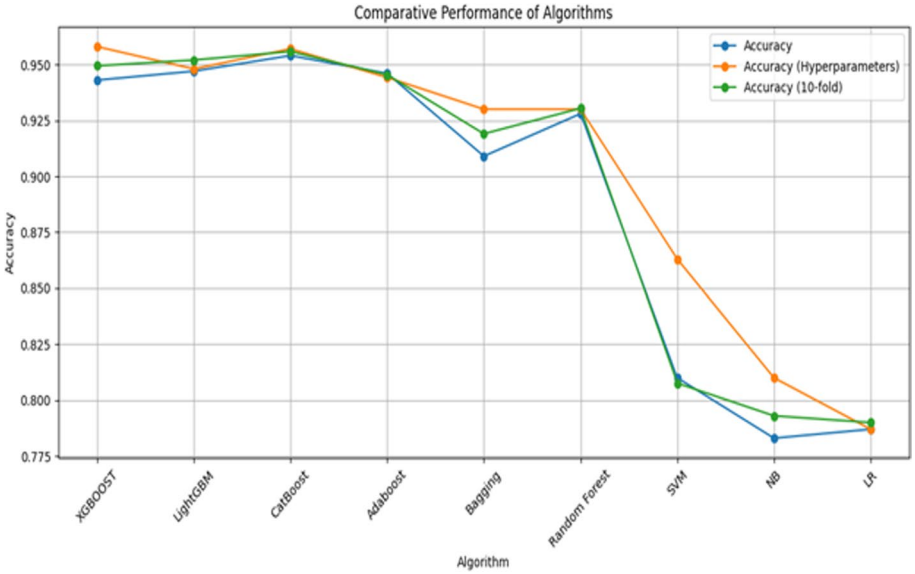

Fig. 7 Comparison of algorithm accuracy

Figure 7 depicts the line graph of the proposed model in terms of accuracy. In the Random forest model, the accuracy is 93.00% after applying hyperparameter tuning and 93.04% after applying tenfold cross validation technique. Whereas in Logistic regression the accuracy is 78.7% after applying hyperparameter tuning and 79% after applying tenfold cross validation technique.

Furthermore, Table 3 depicts the comparison of different machine learning techniques in terms of precision, sensitivity and f-measure. It is clear from the table that XGBoost has a precision of 0.9305, sensitivity of 0.900 and f-measure of 0.915, whereas CatBoost has a precision of 0.9255, sensitivity of 0.9093 and f-measure of 0.9174. The lowest precision of 0.664 has achieved by the Naïve bayes algorithm, the lowest sensitivity of 0.5994 has achieved by logistic regression and the lowest f-measure of 0.658 has achieved by logistic regression.

Figure 8 shows the comparative analysis of different algorithm in terms of accuracy after hyperparameter tuning and accuracy after tenfold cross validation technique. It is

Table 3 Comparison based on precision, sensitivity & f-measure

Model	Precision	Sensitivity or Recall	F-Measure
XGBOOST	0.9305135951661632	0.9005847953216374	0.91530460642407131
LightGBM	0.933933933933934	0.9093567251461988	0.9214814814816
CatBoost	0.9255952380952381	0. 9093567251461988	0.9174041297935103
Adaboost	0.9319526627218935	0.9210526315789473	0.9264705882352942
Bagging	0.8792	0.8509	0.8648
Random forest	0.9050445103857567	0.8918128654970761	0.8983799705449191
SVM	0.7733812949640287	0.6286549707602339	0.6935483870967741
NB	0.6640419947506562	0. 7397660,818713451	0.6998616874135546
LR	0.7295373665480427	0.5994152046783626	0.6581059390048153

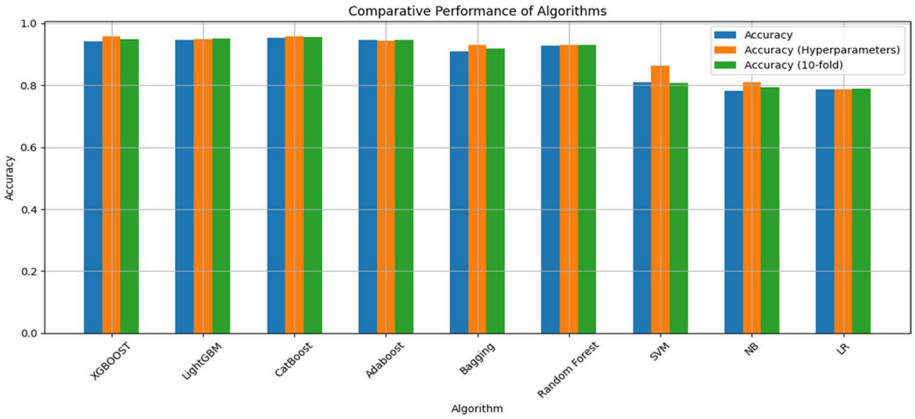

Fig. 8 Comparison of algorithm accuracy in terms of hyperparameter tuning and tenfold cross validation technique

clear from the figure that XGBOOST perform well after hyperparameter tuning with 95.8% of accuracy, whereas CatBoost provide the accuracy result of 95.7%. In the case of tenfold cross validation technique, CatBoost provide the accuracy result about 95.5%, which is outperform than the performance of XGBOOST by 94.9%. Figure 9 shows the comparative analysis of different algorithm in terms of precision, sensitivity and f-measure. CatBoost provides precision of 0.925, sensitivity of 0.909 and f-measure of 0.917, whereas XGBOOST provides precision of 0.930, sensitivity of 0.90 and f-measure of 0.915.

6 Discusion and comparision

This section focuses primarily on a comparative analysis of several diabetic prediction schemes, including the type of dataset used, the size of the dataset, the algorithm type used, and the accuracy rate, sensitivity, precision, and f-measure of the model for each. It also discusses the effectiveness of the various algorithms used in each diabetic prediction

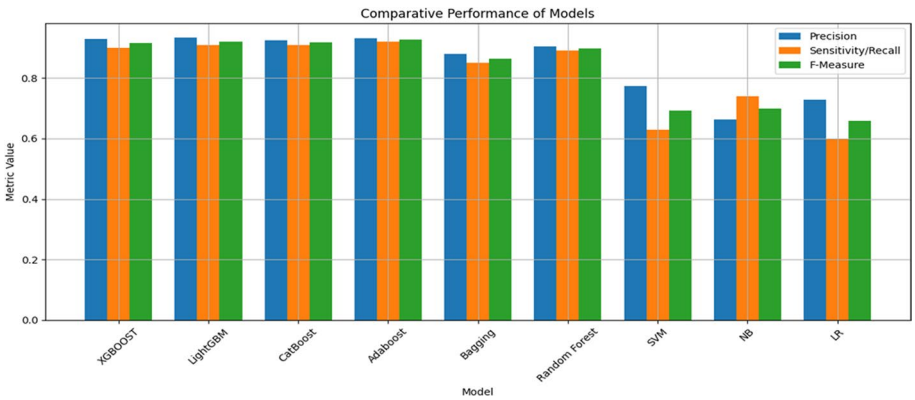

Fig. 9 Comparison of algorithms precision, sensitivity and f-measure

Table 4 Comparison of the proposed system with similar diabetes prediction works

Year	Authors	Algorithm	Dataset (size)	Accuracy	precision	Sensitivity	F-measure
2020	Chatrati et al. [48]	SVM	PIMA	75	0.72	0.75	0.73
2020	Hasan et al. [2]	XGBoost	PIMA	88.8	0.84	0.79	0.81
2020	Maniruzzaman et al. [46]	LR and RF	National health and nutrition examination	94.25	NA	NA	NA
2021	Kumari et al. [49]	Soft voting	PIMA (768)	79.1	0.73	0.72	0.72
2022	Oza et al. [33]	KNN	768	77.27	NA	0.4375	NA
2023	Rastogi et al. [44]	Logic Regression	PIMA (768)	82.46	NA	0.68	NA
2023	Febrian et al. [45]	Naïve bayes	PIMA (768)	76.07	0.73	0.71	NA
2023	Tasin et al. [47]	XGBOOST	PIMA	88.5	0.82	0.8	0.81
2023	This work	CatBoost	5000	95.4	0.92	0.90	0.91

scheme. Table 4 shows the comparison of the proposed system with similar diabetes prediction works.

In 2020, the related works on diabetic prediction based on machine learning technique was proposed by [2, 46, 48]. Chatrati et al. [48] proposed a Smart home health monitoring system for predicting type 2 diabetes and hypertension. This proposed work develops an application for a home health monitoring system with a user friendly, easy to use graphical user interface to diagnose blood pressure and diabetes. The SVM algorithm with an accuracy of 75% was found to predict with highest accuracy as compared to other algorithms like KNN and decision tree. In [2] author uses ensembling of different machine learning classifiers for diabetic prediction. Proposed ensembling classifier (adaptive (AB) and gradient (XB)) for predicting diabetes is a better diagnosis, with an AUC of 0.950, when the AUC weighted soft voting and proposed preprocessing pipeline were employed compared to others, which is more reliable than the system proposed by the authors in [48].

In [46] authors proposed a ML based system using the LR-RF combination for feature selection technique and classifier gave the highest classification accuracy. The results demonstrated that the proposed combination reached an accuracy of 94.25% for K10 protocol. It is also observed that the AUC of RF-based classifier for K10 protocol is 0.95 while NB, DT and AB are 0.82, 0.78, and 0.90, respectively. The accuracy rate achieved by the proposed method is lower than the accuracy rate achieved by authors in [48], which is 95.0% by using ensembling classifier (adaptive (AB) and gradient (XB)) for predicting diabetes.

In 2023, the related works on diabetic prediction based on machine learning technique was proposed by [44, 45, 47]. In [44] authors proposed a diabetic prediction model using different machine learning technique. The results of the proposed model reveal that in the logistic regression, the accuracy is high, i.e., 82.46% as compared to other algorithms. Whereas in SVM, the accuracy is low, i.e., 79.22% as compared to other models. Furthermore, the sensitivity is slightly higher in RF, 68.88% in comparison to other models, and SVM is lower, 59.99% in comparison to other models such as naive Bayes, logistic regression and random forest. In [45] authors proposed a diabetic prediction model using supervised machine learning technique, According to the results of proposed method the Naive Bayes algorithm outperforms KNN, with an average value of 76.07 percent accuracy, 73.37 percent precision, and 71.37 percent recall in Naive Bayes and an average value of 73.33 percent accuracy, precision 70.25 percent, and recall of 69.37 percent in KNN. In [47] proposed a diabetes prediction using machine learning and explainable AI techniques, the proposed system provided the best result in the XGBoost classifier with the ADASYN approach with 81% accuracy, 0.81 with F1 coefficient and an AUC of 0.84.

7 Research question or hypothesis

- Which machine learning algorithm performs best in classifying diabetic patients from the dataset?

Null Hypothesis (H0): There is no significant difference in performance among the SVM, LR, NB, KNN, XGBoost, LightGBM, CatBoost, Adaboost, and Bagging models in classifying diabetic patients.

Alternative Hypothesis (H1): There are significant differences in performance among the algorithms, and some algorithms outperform others in diabetes prediction.

- How does the model's performance vary with different feature engineering techniques?

Null Hypothesis (H0): Feature engineering techniques do not significantly impact the performance of the ML models on the diabetic dataset.

Alternative Hypothesis (H1): Feature engineering techniques have a significant impact on the performance of the ML models, leading to improved diabetes prediction.

- Can we identify the most important features that contribute to diabetes prediction?

Null Hypothesis (H0): All features have equal importance in predicting diabetes.

Alternative Hypothesis (H1): Some features are more important than others in predicting diabetes, and identifying these features improves the model's accuracy.

- How does the model's performance change with varying dataset sizes (e.g., subsampling)?

Null Hypothesis (H0): Model performance remains consistent regardless of dataset size.

Alternative Hypothesis (H1): Model performance varies with dataset size, and larger datasets lead to better generalization.

- Is it possible to improve model performance through hyperparameter tuning?

Null Hypothesis (H0): Hyperparameter tuning does not significantly affect the performance of the ML models.

Alternative Hypothesis (H1): Hyperparameter tuning improves the performance of the ML models, resulting in better diabetes prediction.

- How sensitive is the model to imbalanced classes, and can techniques like SMOTE improve its performance?

Null Hypothesis (H0): Imbalanced classes do not significantly impact model performance.

Alternative Hypothesis (H1): Imbalanced classes negatively affect model performance, and resampling techniques like SMOTE can lead to better results.

- Does the addition of ensemble methods (Bagging and Boosting) enhance the overall predictive accuracy?

Null Hypothesis (H0): Ensemble methods do not significantly improve the performance of individual ML models.

Alternative Hypothesis (H1): Ensemble methods, such as Bagging and Boosting, improve the overall predictive accuracy of the ML models for diabetes classification.

ROC (Receiver Operating Characteristic) curve is a graphical representation used to evaluate the performance of binary classification models. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. In addition to the ROC curve, the area under the ROC curve (AUC-ROC) is a common metric used to summarize the model's performance. AUC-ROC provides a single value that represents the overall performance of the classifier. The higher the AUC-ROC,

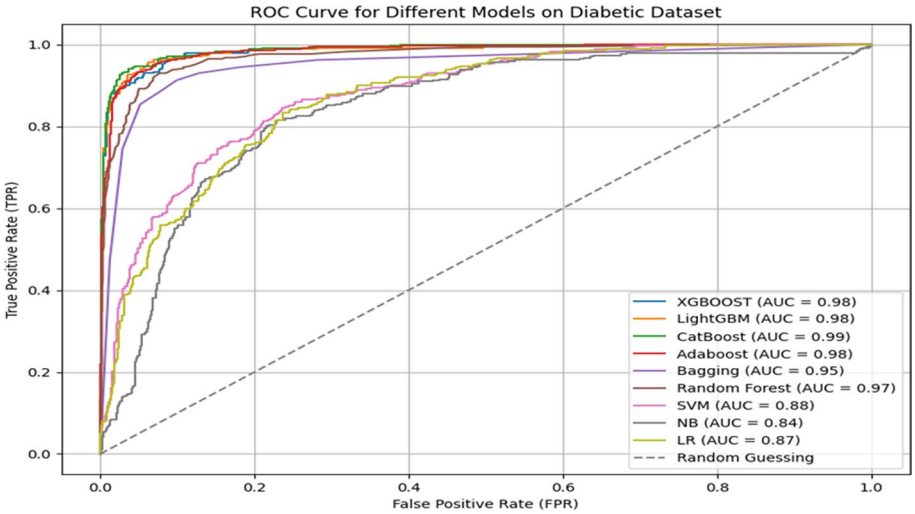

Fig. 10 ROC curve for different algorithm

the better the model's ability to distinguish between the two classes. Figure 10 shows the ROC curve for different algorithms utilized in this proposed work, it is clear from the figure that CatBoost provide higher AUC-ROC score of 0.99, whereas XGBoost with AUC-ROC score of 0.98 and Random forest with AUC-ROC score of 0.97. The results suggest that CatBoost it might be performing slightly better than XGBoost.

Diabetes, a global health crisis with cascading complications such as kidney disease, vision impairment, and coronary disorders, demands innovative solutions for early detection and management. Machine learning has shown promise in transforming medical services, yet a pressing need remains for robust and accurate diabetes prediction models. This study takes a pioneering step towards addressing this critical gap by presenting a comprehensive ensemble approach for diabetes prediction. The prevailing challenge lies in the timely identification of individuals at risk of diabetes, enabling proactive interventions and lifestyle adjustments. Existing research provides individual insights into various machine learning algorithms and their predictive efficacy. However, a holistic model that seamlessly integrates multiple techniques to enhance accuracy and reliability is lacking. This study, therefore, aims to fill this void by proposing a unified framework that amalgamates Logistic Regression, SVM, Naïve Bayes, Random Forest, and cutting-edge ensemble methods. Our approach's value proposition is multifold. Firstly, it exhibits a remarkable enhancement in diabetes prediction accuracy, with CatBoost achieving an impressive 95.4% accuracy compared to XGBoost's 94.3%. This heightened accuracy translates to earlier and more precise identification of individuals at risk, facilitating timely interventions and improved clinical outcomes. Moreover, CatBoost's superior AUC-ROC score of 0.99 further underlines its potential to outperform existing methods. Beyond improved accuracy, our research offers novel insights into diabetes prediction. By dissecting the strengths and weaknesses of various machine learning techniques and their ensemble combinations, we shed light on the intricate dynamics influencing prediction outcomes. This comprehensive analysis equips healthcare practitioners, researchers, and analysts with a nuanced understanding of predictive modeling in diabetes. The industry significance of our work extends beyond medical centers. Accurate diabetes prediction translates to optimized resource allocation,

streamlined patient care, and reduced healthcare costs. In the realm of predictive analytics, our research provides a blueprint for effectively leveraging ensemble techniques to enhance accuracy. This has far-reaching implications for sectors reliant on data-driven decision-making, from pharmaceutical research to insurance underwriting.

The practical utility of our approach for medical centers is evident. By minimizing false alarms, healthcare practitioners can allocate resources more efficiently, ensuring that interventions are focused on individuals truly at risk. Patients benefit from reduced anxiety and a more targeted approach to managing their health. Furthermore, medical centers can optimize their workflows, leading to enhanced patient outcomes and cost savings.

In summary, our research presents a transformative approach to diabetes prediction, offering both enhanced precision and practicality. By addressing the issue of false alarms, we pave the way for more effective resource allocation and patient care in medical centers. Moreover, our findings resonate across industries, highlighting the potential of ensemble techniques to reshape data-driven decision-making. This study marks a significant step towards improving healthcare outcomes and advancing predictive analytics across sectors.

8 Conclusion and future work

These days, machine learning assumes a pivotal part in diabetes forecast in the medical services framework. Diabetes is a significant wellbeing challenge on the planet. An early forecast of diabetes will bring about superior outcomes. This paper proposes a diabetes forecast model with the assistance of machine learning strategies. In this paper, we propose a diabetes forecast model utilizing different machine learning techniques such as Logistic regression, SVM, Naïve Bayes and Random forest. In addition, we have used various ensemble learning techniques like XGBoost, LightGBM, CatBoost, Adaboost and Bagging, which combine the predictions of multiple base learners (weak learners) to create a more accurate and robust final prediction. The proposed mechanism is trained using Python and analysed with a real dataset, which is collected from Kaggle. Furthermore, the performance of the proposed mechanism is analysed using the confusion matrix, sensitivity and accuracy performance metrices. CatBoost provides the best accuracy of 95.4% as compared to XGBoost with 94.3%. CatBoost higher AUC-ROC score 0.99 suggests that it might be performing slightly better than XGBoost with an AUC-ROC score of 0.98. Later on, it is expected to continue working on it and apply different machine learning or deep learning algorithm to anticipate diabetes datasets. It is likewise intended to recommend a better approach to make expectations about diabetes results more precise. However, there is still room for improvement in the accuracy rate. In this paper, we explore future work that could further enhance the system's accuracy and generalization. In future we can use advance feature engineering techniques like Polynomial Features, Group Statistics, Time-Series Transformations and Feature Decomposition. Advanced feature engineering goes beyond simple transformations like scaling or one-hot encoding. Instead, it involves more sophisticated techniques that aim to capture intricate patterns, relationships, and interactions within the data. The goal is to provide the machine learning model with more informative and discriminating features, thereby enhancing its ability to make accurate predictions. Consider experimenting with deep learning architectures like convolutional neural networks (CNNs) or recurrent neural networks (RNNs) to see if they can provide further accuracy improvements. Further, preprocessing techniques may be used to replace missing values

by various missing value imputation techniques like: mean or median, expectation maximization (EM) algorithm, K-nearest neighbors (KNN), fuzzy K-means (FKM), and singular value decomposition (SVD). Moreover, there are various techniques of feature extraction, feature selection (PCA, different statistical tests, FDR, RF, etc.), and classifiers, namely: NN, GPC, SVM, deep learning (DL) and so on.

Data availability https://github.com/MicrosoftLearning/DP100/blob/master/data/diabetes.csv

Declarations

Conflict of interest There is no conflict of interest in the current research.

References

1. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Procedia Comput Sci 132:1578–1585
2. Hasan MK, Alam MA, Das D, Hossain E, Hasan M (2020) Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 8:76516–76531
3. Saeedi P, Petersohn I, Salpea P et al (2019) Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas. Diabetes Res Clin Pract 157:107843
4. Idicula-Thomas S, Kulkarni AJ, Kulkarni BD, Jayaraman VK, Balaji PV (2006) A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. Bioinformatics 22(3):278–284
5. Mellitus D (2005) Diagnosis and classification of diabetes mellitus. Diabetes Care 28:S5–S10
6. Kalyankar GD, Poojara SR, Dharwadkar NV (2017) Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop. International Conference On I-SMAC, 978–1–5090–3243–3
7. Khanam JJ, Foo SY (2021) A comparison of machine learning algorithms for diabetes prediction. ICT Exp 7(4):432–439
8. Seka S, Pon K, Shakila S (2021) Machine Learning Based Diabetic Disease Prediction With Big Healthcare Data. Webology (ISSN: 1735–188X) 18.6
9. Hasan MK et al (2020) Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 8:76516–76531
10. Maniruzzaman M et al (2020) Classification and prediction of diabetes disease using machine learning paradigm. Health Information Sci Syst 8:1–14
11. Sun YL, Zhang DL (2019) Machine learning techniques for screening and diagnosis of diabetes: a survey. Tehnički vjesnik 26(3):872–880
12. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L (2012) Data mining in healthcare and biomedicine: a survey of the literature. J Med Syst 36(4):2431–2448
13. Kaur H, Kumari V (2020) Predictive modelling and analytics for diabetes using machine learning approach. Appl Comput Informatics 18:90–100
14. Sarstedt M, Mooi E (2014) Regression Analysis. https://doi.org/10.1007/978-3-642-53965-7_7
15. Song Y-Y, Ying LU (2015) Decision tree methods: applications for classification and prediction.". Shanghai Arch Psychiatry 27(2):130
16. Mavrogiorgou A, Kiourtis A, Manias G, Kyriazis D (2021) An optimized KDD process for collecting and processing ingested and streaming healthcare data, in: 2021 12th International Conference on Information And Communication Systems (ICICS), IEEE, pp 49–56
17. Zhang Y (2012) Support vector machine classification algorithm and its application. Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings, Part II 3. Springer Berlin Heidelberg pp 179–186
18. Lee M, Gatton TM, Lee KK (2010) A monitoring and advisory system for diabetes patient management using a rule-based method and KNN. Sensors 10(4):3934–3953
19 Resende PAA, Drummond AC (2018) A survey of random forest based methods for intrusion detection systems. ACM Comput Surv (CSUR) 51(3):1–36

20 Chen H et al (2021) Improved naive Bayes classification algorithm for traffic risk management. EURASIP J Adv Signal Process 2021(1):1–12
21. Na S, Xumin L, Yong G (2010) Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, Jian, China, pp 63-67. https://doi.org/10.1109/IITSI.2010.74
22. Alcalá-Fdez J, Alcala R, Herrera F (2011) A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning. IEEE Trans Fuzzy Syst 19(5):857–872
23. Grossi E, Buscema M (2007) Introduction to artificial neural networks. Eur J Gastroenterol Hepatol 19(12):1046–1054
24. Wu H, Yang S, Huang Z, He J, Wang X (2018) Type 2 diabetes mellitus prediction model based on data mining. Inform Med Unlocked 10:100–107
25. Islam MM, Ferdousi R, Rahman S, Bushra HY (2020) Likelihood prediction of diabetes at early stage using data mining techniques. In: Computer Vision and Machine Intelligence in Medical Image Analysis, Springer, Singaporep, pp 113–125
26. Woldemichael FG, Menaria S (2018) Prediction of diabetes using data mining techniques. In: 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, pp 414–418.
27. Fiarni C, Sipayung EM, Maemunah S (2019) Analysis and prediction of diabetes complication disease using data mining algorithm. Procedia Comput Sci 161:449–457
28. Aldallal A, Al-Moosa AAA (2018) Using data mining techniques to predict diabetes and heart diseases. In: 2018 4th International Conference on Frontiers Of Signal Processing (ICFSP), IEEE, pp 150–154
29. Khan FA, Zeb K, Al-Rakhami M, Derhab A, Bukhari SAC (2021) Detection and prediction of diabetes using data mining: a comprehensive review. IEEE Access 9:43711–43735
30. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I (2017) Machine learning and data mining methods in diabetes research, Comput. Struct. Biotechnol J 15:104–116
31. Kumar A, Kumar P, Srivastava A, Ambeth Kumar VD, Vengatesan K, Singhal A (2020) Comparative analysis of data mining techniques to predict heart disease for diabetic patients, in: International Conference on Advances In Computing And Data Sciences, Springer, Singapore, pp 507–518
32. Mahesh TR, Kumar D, Vinoth Kumar V, Asghar J, Mekcha Bazezew B, Natarajan R, Vivek V (2022) Blended Ensemble Learning Prediction Model for Strengthening Diagnosis and Treatment of Chronic Diabetes Disease, vol. 2022, Computational Intelligence and Neuroscience
33. Oza A, Bokhare A (2022) Diabetes prediction using logistic regression and K-nearest neighbor, In: Congress on Intelligent Systems, Springer, Singapore, pp 407–418
34. Anil KS, Jain R (2022) Data mining techniques in diabetes prediction and diagnosis: a review. In: 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, pp 1696–1701
35. Paisanwarakiat R, Na-udom A, Rungrattanaubol J (2022) Combining logistic regression analysis with data mining techniques to predict diabetes. In: International Conference on Computing and Information Technology, Springer, Cham, pp 88–98
36. S.S. Arumugam, V. Kuppan, V. Chakravarthi, K. Palaniappan, An accurate diagnosis of diabetes using data mining, in: AIP Conference Proceedings, vol. 2405, AIP Publishing LLC, 2022, April, 1, p. 020017.
37. J. Abdollahi, B. Nouri-Moghaddam, Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction, Iran J. Comput. Sci. (2022) 1–16.
38. Luo J, Cao S, Ding N, Liao X, Peng L, Xu C (2022) A deep learning method to assist with chronic atrophic gastritis diagnosis using white light images. Dig Liver Dis 54(11):1513–1519
39. Luo J, Sun Y, Chi J, Liao X, Xu C (2022) A novel deep learning-based method for COVID-19 pneumonia detection from CT images. BMC Med Inform Decis Mak 22(1):1–7
40 Zamzami IF, Pathoee K, Gupta BB, Mishra A, Rawat D, Alhalabi W (2022) Machine learning algorithms for smart and intelligent healthcare system in Society 5.0. Int J Intell Syst 37(12):11742–11763
41. Sedik A, Hammad M, Abd El-Samie FE, Gupta BB, Abd El-Latif AA (2021) Efficient deep learning approach for augmented detection of Coronavirus disease. Neural Comput Appl 1–18
42. Pathoee K, Rawat D, Mishra A, Arya V, Rafsanjani MK, Gupta AK (2022) A cloud-based predictive model for the detection of breast cancer. Int J Cloud Appl Comput (IJCAC) 12(1):1–12
43. Hammad M, Abd El-Latif AA, Hussain A, Abd El-Samie FE, Gupta BB, Ugail H, Sedik A (2022) Deep learning models for arrhythmia detection in IoT healthcare applications. Comput Electr Eng 100:108011
44. Rastogi R, Bansal M (2023) Diabetes prediction model using data mining techniques. Measurement: Sensors 25:100605

45. Febrian ME, Ferdinan FX, Sendani GP, Suryanigrum KM, Yunanda R (2023) Diabetes prediction using supervised machine learning. Proc Comput Sci 216:21–30
46. Maniruzzaman M, Rahman MJ, Ahammed B, Abedin MM (2020) Classification and prediction of diabetes disease using machine learning paradigm. Health Inform Sci Syst 8:1–14
47. Tasin I, Nabil TU, Islam S, Khan R (2023) Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technol Lett 10(1–2):1–10
48. Chatrati SP, Hossain G, Goyal A et al (2020) Smart home health monitoring system for predicting type 2 diabetes and hypertension. J King Saud Univ Comput Inf Sci 34(3):862–870
49. Kumari S, Kumar D, Mittal M (2021) An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. Int J Cognit Comput Eng 2:40–46

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.