BMC Medical Research Methodology

RESEARCH

Open Access

An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques

Abdulaziz Altamimi[1], Aisha Ahmed Alarfaj[2], Muhammad Umer[3], Ebtisam Abdullah Alabdulqader[4], Shtwai Alsubai[5], Tai-hoon Kim[6*] and Imran Ashraf[7*]

Abstract

Diabetes is thought to be the most common illness in underdeveloped nations. Early detection and competent medical care are crucial steps in reducing the effects of diabetes. Examining the signs associated with diabetes is one of the most effective ways to identify the condition. The problem of missing data is not very well investigated in existing works. In addition, existing studies on diabetes detection lack accuracy and robustness. The available datasets frequently contain missing information for the automated detection of diabetes, which might negatively impact machine learning model performance. This work suggests an automated diabetes prediction method that achieves high accuracy and effectively manages missing variables in order to address this problem. The proposed strategy employs a stacked ensemble voting classifier model with three machine learning models. and a KNN Imputer to handle missing values. Using the KNN imputer, the suggested model performs exceptionally well, with accuracy, precision, recall, F1 score, and MCC of 98.59%, 99.26%, 99.75%, 99.45%, and 99.24%, respectively. In two scenarios one with missing values eliminated and the other with KNN imputer, the study thoroughly compared the suggested model with seven other machine learning techniques. The outcomes demonstrate the superiority of the suggested model over current state-of-the-art methods and confirm its efficacy. This work demonstrates the capability of KNN imputer and looks at the problem of missing values for diabetes detection. Medical professionals can utilize the results to improve care for diabetes patients and discover problems early.

Keywords Diabetes detection, Missing values, Healthcare, KNN imputer, Ensemble learning

*Correspondence:
Tai-hoon Kim
323020@zust.edu.cn
Imran Ashraf
imranashraf@ynu.ac.kr
[1] College of Computer Science and Engineering, University of Hafr Al-Batin, Hafr Al-Batin 39524, Saudi Arabia
[2] Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[3] Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur 63100, Pakistan
[4] Department of Information Technology, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia
[5] Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[6] School of Electrical and Computer Engineering, Chonnam National University, 50, Daehak-ro, Yeosu-si 59626, Republic of Korea
[7] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

Introduction

Medical experts aid in the diagnosis and treatment of illnesses, wounds, and other deficiencies in people's bodies and minds. Surgeons, physicians, dentists, and their associates comprise the group of health experts. In addition, physical therapy, pharmacy, nursing, optometry, and athletic training are all included in healthcare. It is believed that the health care system is essential to preserving people's bodily and emotional well-being. Diagnosing a disease is the first and most important step in treating it; in the case of diabetes mellitus (DM), an early diagnosis may be crucial. It is a disorder in which the body either produces too little insulin or does so inefficiently. The hormone that controls blood sugar levels is called insulin. High blood sugar, or hyperglycemia, is the result of uncontrolled diabetes. Severe damage to key organs and their functioning caused by hyperglycemia disrupts linked systems, particularly blood vessels and neurons [1]. 8.5% of people over the age of 18 had diabetes in 2014. Approximately 1.6 million fatalities worldwide in 2016 were attributed to diabetes mellitus (DM) compared to 2.2 million deaths worldwide in 2012 [1–4]. The number of deaths attributed to DM surpassed 1.5 million in 2019 [5]. Growing numbers of DM patients create a major financial burden on the world's medical aid programs as well as a general financial predicament. The estimated annual cost of care for DM patients worldwide is close to 825 billion dollars [6]. According to statistics, there will be 629 million DM sufferers worldwide by the end of 2045 [7].

DM is classified into four [8] types: Diabetes mellitus Type-1 (juvenile diabetes), Type 2 (non-insulin-dependent diabetic mellitus), Type 3 (gestational), and Type 4 (prediabetes) are different forms of the disease. Type 1 diabetes is caused by insufficient production of insulin by the human body. Insulin from an outside source needs to be injected for this type of diabetes. Diabetes type-2, or non-insulin dependent diabetes, does not necessitate external insulin injections. The body's incapacity to utilize insulin as intended is what defines this disorder. The third type of diabetes that raises blood sugar levels in pregnant women without any prior diabetic symptoms is called gestational diabetes. The normal range for glucose concentration is 100-125 mg/dL. The WHO lists the following few issues as potential causes of diabetes mellitus:

- Elevated blood glucose levels that are higher than usual
- Persistently higher than average fasting blood sugar readings
- Elevated blood triglyceride levels
- People who are over 45 and rarely work out
- Elevated blood pressure
- Pregnant ladies over the age of thirty
- Body mass index greater than 24 kg/m2
- Family history of diabetes

The WHO's statistics indicate that the number of diabetic patients is rising daily. Both in industrialized and emerging nations, a sizable portion of the population is becoming less active. Moreover, DM may also result from eating behaviors such as obesity, consuming fast food and junk food, eating irregularly, etc. Depression, worry at work, and pressure from the job upset the stomach and can lead to DM. Diabetes disorders are becoming harder to maintain and control due to insufficient abilities and understanding of how to use the current technology to live a mote healthy lifestyle. Regular exercise and eating habits can help prevent diabetic problems. To alter the existing way of life, deliberate efforts are required. Examples of these include meal recommendation systems, physical activity tracking and monitoring systems, drug warning systems, and interactive chatbots.

Diabetes management is a multifaceted challenge that necessitates comprehensive strategies beyond medical interventions alone [9]. A holistic approach, therefore, that will incorporate the use of technology and lifestyle changes will have to be used in effectively curbing the rising number of diabetics. For instance, the use of meal recommendation systems can significantly help in directing people to healthy dietary habits that best suit them based on their nutritional requirements. Through physical activity tracking and monitoring systems, exercise can be appropriately implemented, which is one of the significant aspects of blood glucose control [10]. Drug warning systems also ensure that medication is taken on time and not late, which may cause complications. Chatbots are interactive and give real-time support and information, closing the gaps in patient education and engagement. In this regard, by combining such technological aids with an active lifestyle change, the management of diabetes by an individual will improve dramatically, and their health will be positively enhanced.

This will need intentional effort by the individual and the healthcare givers to incorporate the technological advances in their daily lives [11]. There is a need for the provision of education through campaigns that will sensitize the community on the benefits of using such devices and how to use them appropriately to be in charge of their health. Healthcare professionals should be trained on how to apply these technologies when training them to offer complete guidance and support to their patients [12]. To this end, technology developers, healthcare providers, and patients must collaborate on how the most user-friendly solutions can be formulated and implemented. Policymakers also have a role in ensuring that

these technologies are available and affordable, especially in developing countries, which are experiencing a surge in the prevalence of diabetes. The wide diffusion of such resources could create a supportive environment in making significant strides against the diabetes epidemic and further promoting the quality of life for people with this condition.

In health analytics and medicine-related databases, data mining stands as an imperative attribute for a prior positive help in sensitivity and specificity for the diagnosis of diseases [13]. Generally speaking, this technology and tools can deliver more valuable and exact results overall [14]. Diabetes is another disease in which data mining and ML approaches are extremely useful for faster and more accurate diagnosis. Finer and more automatic results are generated using machine learning models. Many strategies and models have been used to produce the most accurate categorization. The subcategory classification information used in these models was collected from various sources, and such records may be incomplete or null, among other effects that lead to less efficient models. The main goal of this work is to help in developing the model used in predicting drug delivery with machine-learning approaches. This work explores how missing data affects the diagnosis of diabetes mellitus and adds the following:

- A new ensemble model for diabetes prediction is presented. The machine learning methods random forest (RF), extreme gradient boosting (XGB), and extra tree classifier (ETC) are applied in the suggested ensemble model to maximize their combined strengths. A majority vote process is used to integrate these models and determine the final prediction.
- To complete the values that are missing, experiments are carried out using the KNN imputer in addition to the original data. Model performance is carried out both with and without KNN.
- The performance of the proposed ensemble model is evaluated utilizing stochastic gradient descent (SGD) and decision trees (DT) as examples of machine learning models, logistic regression (LR), random forest (RF), support vector classifier (SVC), gradient boosting machine (GBM), Gaussian Naive Bayes (GNB), and ETC. Additionally, the suggested model's performance is contrasted with cutting-edge methods.

The following is the order of this paper: A review of earlier DM literature is given in Related work section. The experimentation methodologies and dataset are explained in Materials and methods section. Additionally, it offers the specifics of the recommended procedure. The results are displayed in Results and discussion section, and the conclusion and next steps are presented in Conclusion section.

Related work

Data mining and machine learning together offer a potential remedy for a number of issues. Because of its massive feature set and sheer bulk, medical data is challenging to analyze. Machine learning has proven to be important in many fields, including medicine. It has made it possible to preserve sensitive medical data while simultaneously developing precise and dependable systems for medical applications. Similarly, early-stage DM hazards have been identified through the use of machine learning algorithms. Risk variables that are frequently considered include blood pressure, insulin, glucose, and body mass index (BMI).

In [15], the authors introduce a machine learning (ML) system to forecast, given the current year's (y) variable, the likelihood of developing Type-2 diabetes in the next year (y + 1). The collection is based on electronic health records (EHR) that were acquired from a hospital that is private between 2013 and 2018. There are two methods used for feature elimination: chi-square and analysis of variance (ANOVA). The system achieves 81% accuracy by using an ensemble classifier that combines soft voting (SV) and confusion-matrix-based integration (CIM). A method for diabetes detection using ensemble learning was presented by Rupapara et al. [16]. In the study, the authors made use of a publically accessible dataset (Pima India). They also used the eight separate ML models to assess the system's effectiveness. They used original features, Chi-2, and principal component analysis (PCA) in a number of their experiments. The results demonstrate that the Chi-2 features perform better than other features, and the suggested ensemble model was able to predict diabetes with an accuracy score of 85%.

The study by Saad et al. [17] focuses on developing an improved deep learning model for detecting COVID-19 from chest X-ray images. The study utilizes a convolutional neural network (CNN) architecture and incorporates transfer learning to enhance the model's performance. The authors highlight the importance of accurate and rapid detection of COVID-19 to aid in early diagnosis and subsequent treatment. They evaluate their model on a dataset comprising COVID-19 positive and negative cases, demonstrating promising results in terms of accuracy and sensitivity. Another study by Ozturk et al. [18] presents a deep learning approach specifically designed for the detection of COVID-19 from chest X-ray images. The study proposes a deep convolutional neural network (DCNN) architecture tailored for this purpose, aiming

to contribute to the efforts in combating the COVID-19 pandemic through early and accurate diagnosis. The authors evaluate their model using a dataset of chest X-ray images obtained from COVID-19-positive and negative patients, demonstrating competitive performance in terms of classification accuracy and sensitivity.

Transfer learning and data augmentation approaches were used by Deng et al. [19] to overcome issues caused by uneven datasets and a lack of training data. Three neural network topologies, augmentation techniques, transfer learning techniques, as well as various loss functions, such as mix-up and generative models were all examined by the writers. An analogous network architecture for Type-1 DM detection was developed using the public OhioT1DM dataset, and the International Review Board and Beth Deaconess Medical Center accepted the research. The authors report a 95% accuracy rate. Similar to this, Butt et al. [20] described a machine learning method for determining and categorizing diabetes in its early stages. The authors also suggest a fictitious Internet of Things (IoT) system for tracking blood glucose (BG) levels in both healthy people and diabetics. The multilayer perceptron (MLP), LR, and RF were employed.

Oversampling and feature augmentation were proposed as part of an ML technique by Shamreen Ahamed et al. [21]. They made use of the DMS and Pima benchmark datasets. The pretreatment and enhancement scenarios were the two in which the authors conducted their research. A comparison of the results reveals that the accuracy according to the DMS dataset is 99.45%, while the Pima Indian dataset has the maximum accuracy of 92.5%. In a similar vein, Pethunachiyar [22] unveiled an ML-based approach for DM patient classification. The study demonstrated that support vector machines (SVM) with linear functions enabled performed better than classifiers utilizing DT, Naive Bayes, and neural networks. Nevertheless, There is no state-of-the-art comparison in the article, and the parametric data for the models is absent.

An ensemble learning system was proposed by Laila et al. [23] for the prediction of diabetes in its early stages. The authors used the UCI diabetes dataset, which has 17 attributes for each record. Chi-2 is also employed in the feature selection process. AdaBoost, bagging, and RF are three predictive models that are employed in the trials. According to experimental results, the RF model performed better than other models, achieving an accuracy of 97%. Madan et al. [24] have presented a deep learning-based ensemble method for Type 2 DM detection. Both ensemble and standalone deep learning models are used in the experiments. The study's findings demonstrated that, with an accuracy of 88.37%, compared to the other models, the ensemble model CNN-BiLSTM fared better.

Kannadasan et al. [25] developed a DNN model to classify Type-2 diabetes. The model applies a softmax function to the classification, uses stacked encoders for feature engineering, and fine-tunes the model by applying backpropagation. The training is based on 786 PIDD patient records. It is noted that the model classifies accurately 86.26% of the patients. Dutta et al. [26] proposed an automated system to predict diabetes earlier. This work applies the methodology proposed to use these five machine-learning algorithms with tuned parameters for optimal results. The pipeline also consists of K-fold cross-validation, feature selection, and imputation for missing values. As confirmed by ANOVA testing, the proposed weighted ensemble approach, in cooperation with the preprocessing techniques (RF, DT, LGB, XGB), significantly improves performance in predicting diabetes. 73.5% accuracy was reported by the study when employing the ensemble technique.

Tasin et al. [27] proposed a machine learning-based system for diabetes detection. They used the explainable AI methods SHAP and LIME for better insights into the prediction process. Results of the study show that the XGB outperforms the other learning models and achieved the highest accuracy score of 81%. For the efficient prognosis of diabetes an ensemble learning approach is proposed by the S.M Ganie et al. [28]. For the predictive analysis, they used multiple preprocessing steps such as data normalization, feature selection, upsampling methods, etc. Results of the study show that the GBC achieved the highest accuracy score of 92.85%. In a similar fashion study [29] proposed a weighted ensemble model for diabetes prediction. The proposed weighted ensemble model achieved the highest accuracy score of 88.84% on the PIMA Indian diabetes dataset.

Torkey et al. [30] explore the use of machine learning models to enhance cancer diagnosis through RNA sequencing (RNAseq) microarray data. The primary focus of their research is the development of a function named Feature Reduction Classification Optimization (FeRCO). FeRCO integrates various machine learning techniques to predict cancer presence based on minimal and optimized features. Patil and Sherekar [31] provide a comparative analysis of various machine-learning algorithms for cancer prediction. The study aims to identify the most effective algorithms for predicting cancer-based on historical patient data. The study compares algorithms such as Decision Trees, Naive Bayes, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM). Evaluation of these algorithms is based on accuracy, precision, recall, F-measure, and ROC curves. The findings indicate that SVM and KNN algorithms generally outperform

others in terms of accuracy and reliability in cancer prediction. Decision Trees and Naive Bayes also show strong performance but are slightly less accurate compared to SVM and KNN.

Notwithstanding the aforementioned research works' excellent performance, a number of drawbacks are evident. For instance, the problem of missing values is not examined, despite the fact that some studies employ data augmentation to address the problem of unequal class distribution. It is crucial to remember that the conclusions of some of the models previously presented may not be as generalizable because they were assessed using sparse data. In order to increase the efficacy and accuracy of feature selection for DM prediction, more research and analysis are needed. In order to enhance the models' performance for DM detection, more study is required. This may entail using ensemble models, adjusting parameters, and investigating new designs. Table 1 displays a summary of the relevant work.

Materials and methods

Diabetes dataset

The Kaggle data source provided the dataset for diabetes detection. It comprises a total of 768 samples from female patients [32], all aged 21 years old. Among these samples, 268 belong to patients diagnosed with diabetes, while the remaining 500 are for non-diabetic patients. One notable aspect of the dataset is the presence of multiple zero values for different attributes. For example, there are 227 individuals whose skinfold thickness is zero, 35 individuals with zero diastolic blood pressure, and 27 patients with a BMI of zero. To solve this problem, the existence of zero is handled using a KNN imputer, which has a great deal of predictive strength for identifying the last group (whether diabetic or not). Table 2 provides a description of the dataset and Table 3 shows the correlation coefficients between features of the dataset.

Table 1 An overview of related literature

Classifier	Dataset	Reported accuracy
CIM, SV, ST, LR, XGBoost, SVM, RF, and Ensebmle (CIM, SV, ST)	Private organization HER 2013-2018	81% Ensemble
GNB, SVC, KNN, DT, RF, LR, ADA, ETC, and LTC (ensemble)	PIMA(India)	86.3% LTC (ensemble)
CNN, RNN, and SAN	The dataset of BIDMC	95% CNN
LSTM, MA, LR, RF, MLP, and linear regression	Pima Indian	88.36% LSTM
RF, GBM, and LGBM	PIMA, DMS	99.70% LGBM (DMS)
		93.3% LGBM (Pima)
CNN, NB, DT, Polynomial SVM, Radial SVM, and Linear SVM	Diabetes dataset from UCI	99.90% Linear SVM
RF, ADA, and Bagging	Diabetes dataset from UCI	97% RF
CNN, Dense-NN, CNN-LSTM, BiLSTM, Proposed (CNN-BiLSTM)	PIMA	88.37% CNN-BiLSTM
DNN	Pima	87.46%
LGB, XGB, RF, DT, and NB and proposed (DT+RF+XGB+LGB)	Bangladesh's DDC dataset	73.5% proposed
DT, KNN, RF, SVM, LR, ADA, XGB, Voting, Bagging	PIMA	81% XGB
XGB, CatBoost, LGBM, GBC, ADA	PIMA	92.85% GBC
kNN, DT, RF, ADA, NB, XGB, MLP, weighted ensemble model	PIMA	88.84% Ensemble

Table 2 Details of the PIMA diabetes dataset for India

Dataset property	Detail	Kind of characteristic	Mean ± SD
Age	Years of age	Constant	33.24 ± 11.76
Class (Target)	Diabetes versus normal	Categorized	
Glucose	Glucose in plasma (2 hours)	Continuous	121.67 ± 30.46
Insulin	Serum insulin (μ U/ml) after two hours	Continuous	141.76 ± 89.10
Mass	Body mass index (kg of weight/m2 of height)	Continuous	32.43 ± 6.88
Pressure	Blood pressure diastolic (mm Hg)	Continuous	73.45 ± 12.10
Pedigree	Diabetes family history	Continuous	0.47 ± 0.33
Pregnant	number of pregnancies	Continuous	3.84 ± 3.36
Triceps	Skinfold thickness (mm) of the triceps	Continuous	29.08 ± 8.89

Table 3 Correlation eoefficients between features

	Pregnancies	Glucose	BP	SkinThickness	Insulin	BMI	DPF	Age	Outcome
Pregnancies	1.000	0.129	0.141	-0.082	-0.074	0.018	-0.034	0.544	0.222
Glucose	0.129	1.000	0.153	0.057	0.331	0.221	0.137	0.264	0.467
BP	0.141	0.153	1.000	0.207	0.089	0.282	0.041	0.240	0.065
SkinThickness	-0.082	0.057	0.207	1.000	0.437	0.393	0.184	-0.114	0.075
Insulin	-0.074	0.331	0.089	0.437	1.000	0.198	0.185	-0.042	0.131
BMI	0.018	0.221	0.282	0.393	0.198	1.000	0.141	0.036	0.293
DPF	-0.034	0.137	0.041	0.184	0.185	0.141	1.000	0.034	0.174
Age	0.544	0.264	0.240	-0.114	-0.042	0.036	0.034	1.000	0.238
Outcome	0.222	0.467	0.065	0.075	0.131	0.293	0.174	0.238	1.000

Data preprocessing

Data preprocessing is a crucial step in achieving enhanced ML model performance. It involves the removal of unnecessary or redundant data that holds no meaningful information for the models [33]. By conducting preprocessing, we can improve the efficacy of the models of learning. Preprocessing also contributes to cutting down on computation time. In the present study, several zero values were found in the dataset across various attributes during the data preprocessing stage. For instance, 374 individuals have zero serum insulin, Skinfold thickness is 0 in 227 individuals, 27 patients have zero BMI, 35 people have zero diastolic blood pressure, and so forth. Preprocessing requires the identification and management of these zero values. Values that are zero or null in data characteristics can have a big impact on how well models work. We employed the KNN imputer to solve this problem. Attributes in the dataset that have zeros (missing values) are Diabetes (Target Class), Age, Pregnant, and Pedigree. The attribute Mass has 12 missing values, Glucose has 6, Pressure 36, Tricepts 227, and Insulin 374. The table unequivocally demonstrates the large amount of missing values in the dataset. There are two ways to handle the missing values in the dataset because it is categorical:

KNN imputer: Using the KNN imputer is one method. Utilizing its K closest neighbors' values, this approach estimates the missing figures. The imputer infers and fills in the missing values using the existing data points.

Eliminating missing values: An additional method is to merely eliminate the rows or occurrences of the dataset's missing values.

This entails eliminating from the analysis any samples that have missing values, producing a reduced but full dataset. Each approach has benefits and things to think about. The dataset's characteristics, the amount of missing data, and the analysis's objectives all influence which option is best.

KNN imputer

In the modern world, data is gathered from a variety of sources in order to facilitate analysis, produce insights, and validate theories. Nevertheless, missing information can frequently be found in the data that has been gathered because of problems with extraction or human error during collection [34]. Perhaps the most crucial step in the data preparation process, data imputation takes care of missing variables. This stage is very important because it may affect the model's performance. One commonly used imputation method for imputing missing data is the KNN imputation method, which comes with a sci-kit-learn toolkit. This gives an alternative approach to the more conventional methods. KNN imputer finds out who are the nearest neighbors through the Euclidean distance matrix and, thus, this is a tool for infilling the missing values from the observations. It does this by weighing the non-missing coordinates higher and not taking account of the missing ones in the calculation for the Euclidean distance. One can obtain the Euclidean distance using the formula:

$$D_{xy} = \sqrt{weight * squared\ distance\ from\ present\ coordinates} \quad (1)$$

where weight is the proportion of all coordinates to all coordinates at this time.

Deleting missing values

Eliminating missing values from the data is an alternate approach to handling it. This method removes from the dataset any fields or variables that have missing values. In the second set of tests, this other technique is applied, wherein the missing values are eliminated entirely prior to performing additional analysis.

Models used for diabetes prediction

There are currently numerous machine learning methods available for diabetes prediction. For the same objective, other ML models have already been used, with varying

reported outcomes. The most accurate models are used in this study in light of the outcomes of those models that have been reported. LR, SGD, SVM, XGB, GNB, ETC, and DT are all used in this study. This study's part offers a succinct explanation of various algorithms. The best hyperparameters of all learning models are calculated using a technique of GridSearchCV as shown in Table 4.

XGBoost

For supervised regression models, one popular technique is XGBoost, known for its ability to enhance the accuracy of goal functions and base learners [35]. By merging the results of several different models to produce a single forecast, it makes use of the ensemble learning concept. For this reason, XGBoost is categorized as an ensemble learning method.

Decision tree

A hierarchical tree structure is used by the predictive model DT [35, 36]. In order to anticipate the goal value (found in the tree leaves), it arranges the features (shown as branches). When there is a finite number of possible values for the goal parameter, classification trees are used. In decision tree models, branches represent the feature criteria These result in class labels; the class labels are represented by the leaves themselves. When the target variable accepts continuous values, usually real numbers, regression trees are employed. Knowledge Gain (IG) and the Gini coefficient are the main metrics used in decision trees to choose the root node. IG is used to identify a decision tree's root node.

Gaussian Naive Bayes

The Gaussian Naive Bayes (GNB) model applies Bayes' theorem, using unconditional probability to predict the outcome of an event [37]. In GNB, a sample is split into k categories, represented as $k = \{c1, c2, ..., ck\}$, with c standing for the class. The GNB function is expressed like in this example, where d represents the sample and c the class:

$$P(c|d) = (P(c) \times P(d|c))/P(d) \quad (2)$$

$P(c|d)$ in this equation represents the likelihood of class c given sample d. $P(c_i d)$ represents the likelihood of observing sample d for a given class c, while $P(d)$ is the probability of observing sample d. The prior probability of class c is represented by $P(c)$.

Table 4 Hyperparameters of all learning models and KNN imputation

Model	Hyperparameters
DT	n_estimator=55, max_depth=15, random_state=22
RF	n_estimator=55, max_depth=15, random_state=22
SVC	C= 1.0, kernel= rbf , degree=3, gamma= scale
XGB	n_estimator=55, max_depth=15, random_state=22
KNN	n_neighbors=5, weights= uniform, algorithm=auto
ETC	n_estimator=55, max_depth=30, random_state=22
SGD	penality=l2, loss= log
GNB	alpha= 1.0
LR	penalty=l2 , solver= lbfgs
VC	Voting='soft'

Logistic regression

Logistic regression (LR) is another flexible regression technique. It works by applying Boolean operations on the binary covariates to generate predictors. The method's name refers to the logistic function: it is one of the critical building blocks in this type of analysis [38]. The S-shaped curve of the logistic function, also known as the sigmoid function, maps real-valued integers into values within 0 and 1. LR is preferred in cases where dependent variables are categorical, for it performs well in such cases specifically.

Stochastic gradient classifier

When it comes to multi-class classification issues, SGD is an effective classifier. Its basis is found in the convex loss function-based SVM and LR principles [39]. The one-vs-all (OvA) method is used by SGD to integrate several binary classifiers. The efficiency with which SGD can handle enormous datasets is one of its main features. It does this by processing just one example (in batches of one) per iteration-an extreme method. Regression analysis is the foundation of SGD's simplicity, which makes it simple understandable, and workable. It is crucial to remember that SGD's random sample selection from the batch makes it potentially quite noisy. Furthermore, SGD has a high sensitivity to feature scaling, and precise results necessitate precise hyperparameter calibration.

Random forest

To provide a single result, RF combines the outputs of several decision trees [40]. Decision trees are used as the fundamental method for sampling rows and columns in order to achieve this. Based on the input, how many decision trees there are also referred to as base learners is improved, resulting in increased accuracy and decreased volatility. RF is a well-known and important technique in the field of bagging techniques.

Extra tree classifier

Introduces considerable randomization throughout the node splitting process by randomly selecting the attributes and cut spots [41, 42]. Under extreme circumstances, ETC

can produce totally randomized trees that don't depend on the training sample's output values. The following is the main difference between the popular machine learning model RF and ETC:

- While Random Forest trains its models using bootstrap replicates, or random selections, ETC uses the complete dataset.
- ETC randomly selects the best characteristics and their matching values in order to separate the nodes.
- Owing to these features, the ETC frequently performs better on datasets and is less prone to overfitting.

Support vector classifier

SVC is a well-liked machine learning method that categorizes input points by searching for a hyperplane in N-dimensional space [43]. The hyperplane that increases the margin between classes is the main objective of this approach. The number of features determines the dimensionality, which is denoted by N. It is generally simple to compare two features, however handling many features for classification might be more difficult. SVC improves prediction accuracy by optimizing the margin.

Proposed approach for diabetes detection

The study's dataset came from the reputable platform Kaggle, which houses publicly accessible datasets. Preprocessing steps were implemented to fill in the gaps and boost the effectiveness of the learning models. Specifically, the KNN imputer method was utilized to address missing data. After that, The dataset was divided 70:30 between training and testing sets, with 30% put aside for testing and 70% used for training the models.

The suggested methodology for detecting diabetes uses an ensemble method, combining three algorithms: Random Forest (RF), Extra Trees Classifier (ETC), and XGBoost (XGB). Ensemble models aggregate predictions from several models to improve robustness and accuracy. Every model in the ensemble contributes unique strengths and limitations, and when combined, they frequently produce better outcomes. As seen in Fig. 1, this work suggests utilizing this ensemble learning model to detect diabetes.

The ensemble model is created by combining the predictions of three different machine learning algorithms. Usually, this entails using the same dataset to train several models, and then combining their predictions. In the instance of the XGB+RF+ETC ensemble model, the XGBoost (XGB), Random Forest (RF), and Extra Trees Classifier (ETC) models are each trained independently on the same dataset. Predicted probabilities are produced by each model for every class of the intended variable. The calculated probability from each model is merged to generate a final prediction for every observation in the dataset. Predicted probabilities are often averaged, and each model is given a weight based on how well it performs on a validation set. The ensemble gives higher weight to the models with better performance on the validation set. So, predictions from different models trained on a diabetes dataset can be combined to be more robust with less overfitting. Algorithm 1 explains the details of how the proposed ensemble model works.

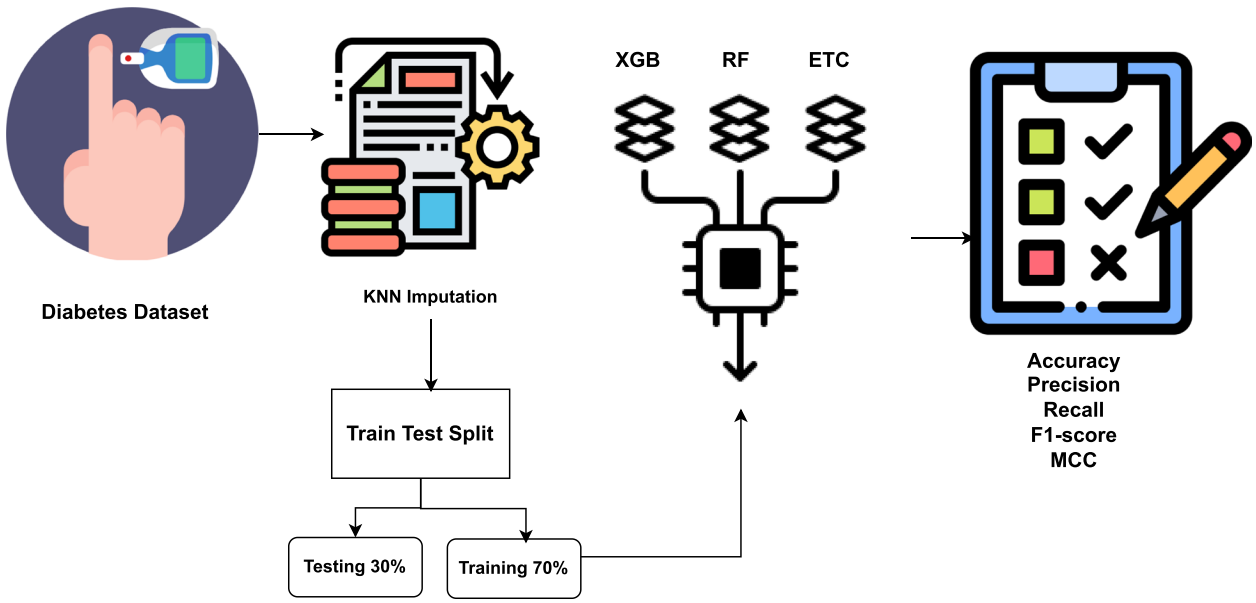

Fig. 1 Workflow diagram of the adopted methodology

Algorithm 1 Ensembling XGB, RF, and ETC for diabetes classification

Model Training:
Train model_XGB on $(X^*_{\text{train}}, y_{\text{train}})$
Train model_RF on $(X^*_{\text{train}}, y_{\text{train}})$
Train model_ETC on $(X^*_{\text{train}}, y_{\text{train}})$
Model Prediction on Validation Set:
Predict probabilities on X_{val} using model_XGB

$$P_{\text{XGB}} \leftarrow \text{model\_XGB.predict\_proba}(X_{\text{val}})[:, 1]$$

Predict probabilities on X_{val} using model_RF

$$P_{\text{RF}} \leftarrow \text{model\_RF.predict\_proba}(X_{\text{val}})[:, 1]$$

Predict probabilities on X_{val} using model_ETC

$$P_{\text{ETC}} \leftarrow \text{model\_ETC.predict\_proba}(X_{\text{val}})[:, 1]$$

Ensemble Predictions:
Combine the predictions using the average method

$$P_{\text{ensemble}} \leftarrow \frac{P_{\text{XGB}} + P_{\text{RF}} + P_{\text{ETC}}}{3}$$

Determine Optimal Threshold:
Find the optimal threshold τ that maximizes the F1 score on P_{ensemble} and y_{val}
Evaluate Ensemble Model on Validation Set:
Generate binary predictions using τ

$$\hat{y}_{\text{val}} \leftarrow (P_{\text{ensemble}} \geq \tau)$$

Calculate accuracy, precision, recall, and F1 score for $\hat{y}_{\text{val}}$
Final Predictions on Test Set:
Predict probabilities on X^*_{test} using model_XGB

$$P_{\text{XGB\_test}} \leftarrow \text{model\_XGB.predict\_proba}(X^*_{\text{test}})[:, 1]$$

Predict probabilities on X^*_{test} using model_RF

$$P_{\text{RF\_test}} \leftarrow \text{model\_RF.predict\_proba}(X^*_{\text{test}})[:, 1]$$

Predict probabilities on X^*_{test} using model_ETC

$$P_{\text{ETC\_test}} \leftarrow \text{model\_ETC.predict\_proba}(X^*_{\text{test}})[:, 1]$$

Combine the predictions using the average method

$$P_{\text{ensemble\_test}} \leftarrow \frac{P_{\text{XGB\_test}} + P_{\text{RF\_test}} + P_{\text{ETC\_test}}}{3}$$

Generate final binary predictions using τ

$$y_{\text{pred}} \leftarrow (P_{\text{ensemble\_test}} \geq \tau)$$

Output Predictions:
Return y_{pred} for X_{test}

Here, the techniques applied to combine the predictions of single individual models, namely Random Forest (RF), Extra Trees Classifier (ETC), and XGBoost (XGB), are presented.

$$\widehat{p} = argmax\{\sum_{i}^{n} XGB_i, \sum_{i}^{n} RF_i, \sum_{i}^{n} ETC_i\}. \quad (3)$$

Each model represents a probability in estimation. The soft voting criterion aggregates the probabilities from each of the three models, giving a final vote of prediction. This would further enable the ensemble model to make use of the combined expertise of $\sum_{i}^{n} XGB_i$, $\sum_{i}^{n} RF_i$, and $\sum_{i}^{n} ETC_i$ for providing more accurate and firm predictions for every test instance.

The most voted class is then selected by averaging the probability score from combined predictions of all the classifiers inside the ensemble model, ash shown in Fig. 2. Then, after averaging out the predicted probabilities from the RF, ETC, and XGB models, the class with the highest probability score will win the final prediction. Such an approach ensures that the ensemble model makes a well-informed decision by leveraging the strengths of each classifier.

Evaluation parameters

Several well-known assessment measures are applied in measuring the act of a model: F1 score, accuracy, precision, and sensitivity. All these metrics are derived from the confusion matrix, which constitutes four classes: FP (false positives), TP (true positives), FN (false negatives), and TN (true negatives). In classification, a true positive (TP) is when the actual label of a record equals the model's prediction for that particular record-that it is diabetic. FP occurs when the label is standard but the model predicts it as diabetic. TN happens when the actual label is standard and the model correctly predicts it. FN occurs when the exact label is diabetic, but the model predicts it to be expected. The effectiveness can be evaluated by calculating the F1 score, recall, precision, and accuracy using these confusion matrix data. The following are the evaluation parameters that are offered:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall(Sensitivity) = \frac{TP}{TP + FN}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

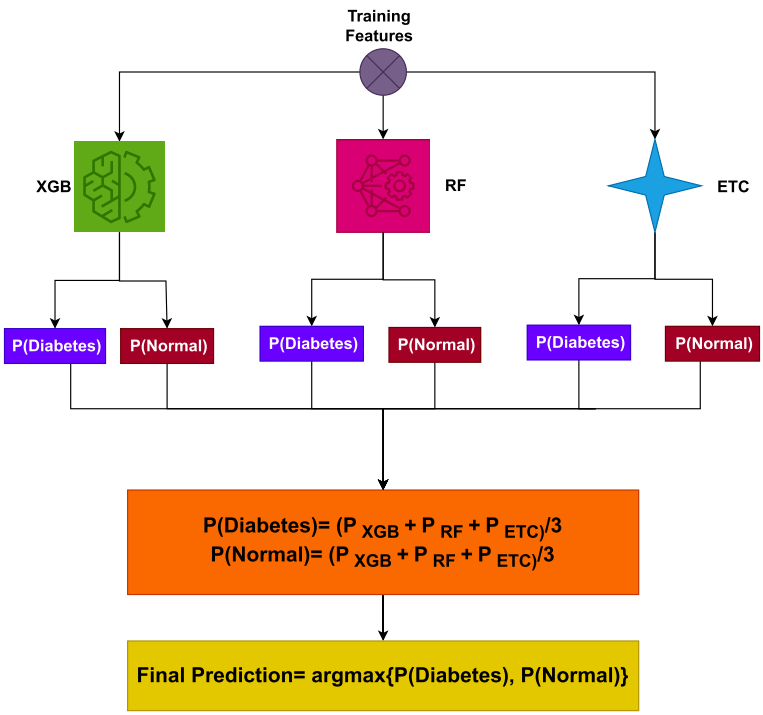

Fig. 2 The proposed voting classifier's architecture

The Matthews Correlation Coefficient (MCC) is a metric used to evaluate the performance of binary classification models [44]. It takes into account true and false positives and negatives and is generally regarded as a balanced measure that can be used even if the classes are of very different sizes. The MCC is essentially a correlation coefficient between the observed and predicted binary classifications; it returns a value between -1 and +1, where:

- +1 indicates a perfect prediction,
- 0 indicates no better than random prediction, and
- -1 indicates total disagreement between prediction and observation.

The MCC is calculated using the formula:

$$\mathrm{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Results and discussion

The findings are elaborated on, and results for the prediction of diabetes mellitus will be articulated in this section. The ML models are written in Python 3.8 and executed with the TensorFlow and Sci-Kit Learn libraries within a Jupyter Notebook environment. Experiments are carried out on a computer with an 11th gen Intel Core i7 processor running at 3.2 GHz under 64-bit Windows 11. Performance is evaluated through the F1 score, recall (sensitivity), precision, and accuracy.

Results of models using missing values

The dataset contains missing values that, from the beginning of the experiments, are addressed prior to applying the ML models to the most recent data. Table 5 shows, in total, a comprehensive snapshot of the performance of the models with regard to their effectiveness in terms of F1 scores, recall, accuracy, and precision.

DT achieved an accuracy of 78.34%, with high precision at 88.61%, recall at 90.55%, and an F1-score of 89.87%. Ensemble methods, such as ETC and RF, showed much better results, with accuracy at 84.18% and 82.75%, respectively. Their values for precision, recall, and F1-score are also identical and stand at 91.45%, showcasing their consistency. The corresponding accuracies were lower in LR and SVC, reducing to 80.55% and 74.87%, respectively, though they still displayed strong precision, recall, and F1 scores. XGB outperformed most models with an accuracy of 84.61% and balanced recall, accuracy, and F1-score values around 91%. The VC combining XGB, RF, and ETC models demonstrated an accuracy of 81.13%, highlighting the effectiveness of ensemble techniques in improving overall performance. Notably, it achieved a remarkably high precision of 94.56%, showing a minimal number of erroneous positives, along with strong recall and F1-score values above 94%. ensemble methods like ETC, RF, and XGB along with the VC combining them showed superior performance across multiple evaluation metrics, indicating their suitability for this particular classification task. However, each model's strengths and weaknesses should be considered depending on the specific requirements and characteristics of the dataset. Figure 3 illustrates the performance of models using the dataset containing the missing values.

Experimental results using data filled using KNN imputer

The KNN imputer is used in the second set of experiments to fill in the dataset's missing values. A few values were discovered to be absent throughout the preprocessing phase; these gaps were filled in using the KNN imputer. In this process, the Euclidean distance measure and the mean of the available data were used. Subsequently, a number of machine learning models that included the KNN imputer were trained and evaluated using the final dataset. Table 6 displays the models'

Table 5 Results of the machine learning models created using the missing variables in the dataset

Model	Accuracy	Precision	Recall(Sensitivity)	F1-Score	MCC
DT	78.34	88.61	90.55	89.87	79.37
ETC	84.18	91.45	91.45	91.45	85.68
GNB	76.48	85.54	86.22	86.09	78.19
LR	74.87	87.64	89.74	88.71	76.67
RF	82.75	90.45	91.85	91.31	85.25
SGD	79.69	87.47	89.98	88.76	81.34
SVC	80.55	87.44	90.44	89.72	83.33
XGB	84.61	91.05	90.99	91.03	86.89
VC(XGB+RF+ETC)	81.13	94.56	94.31	94.43	83.78

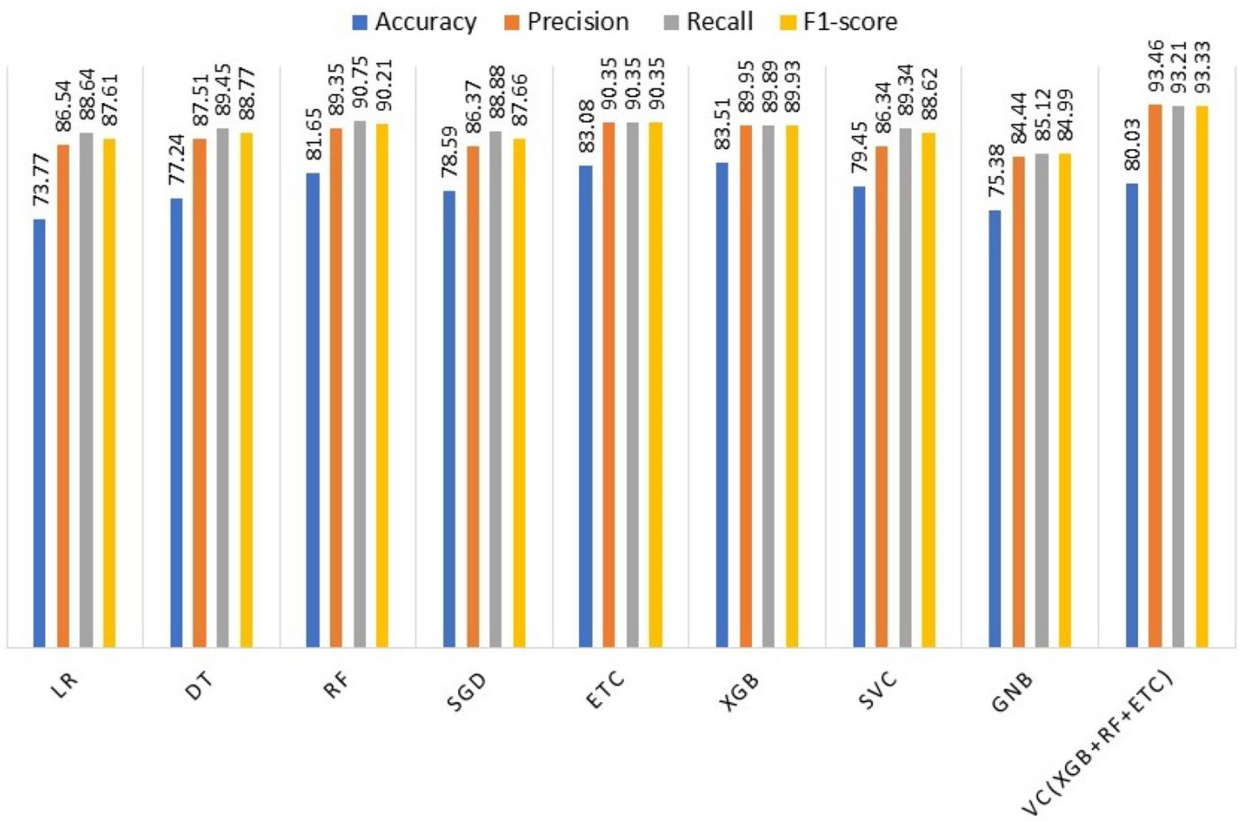

Fig. 3 Outcomes of the machine learning models using the dataset containing missing values

Table 6 Result of the KNN imputer learning models

Model	Accuracy	Precision	Recall(Sensitivity)	F1-Score	MCC
LR	86.52	93.64	96.74	94.61	89.32
DT	89.71	93.61	92.55	93.17	91.52
RF	93.44	96.45	97.98	97.22	95.23
SGD	87.89	92.61	96.93	94.96	88.16
ETC	95.30	97.03	97.43	96.27	96.32
XGB	96.72	97.94	97.35	97.55	97.73
SVC	91.74	95.62	95.53	95.59	93.64
GNB	91.02	91.63	93.30	92.57	92.19
VC(XGB+RF+ETC)	98.59	99.26	99.75	99.45	99.24

performance using the dataset that the KNN imputer enhanced.

Figure 4 illustrates the performance of the models when the data normalized using the KNN imputer is used. The LR model achieved an accuracy of 86.52% with a high precision of 93.64%, recall of 96.74%, and F1-score of 94.61%. DT performed slightly better with an accuracy of 89.71% and a balanced precision-recall tradeoff. RF showed significant improvement with an accuracy of 93.44% and high precision of 96.45%, recall of 97.98%, and F1-score of 97.22%. Compared to other models, ensemble techniques like XGB and ETC performed better, with ETC achieving the highest accuracy of 95.3% and high precision, recall, and F1-score values around 97%. SVC and GNB also demonstrated good performance, with accuracies above 91% and balanced precision, recall, and F1-score values. The VC combining XGB, RF, and ETC models

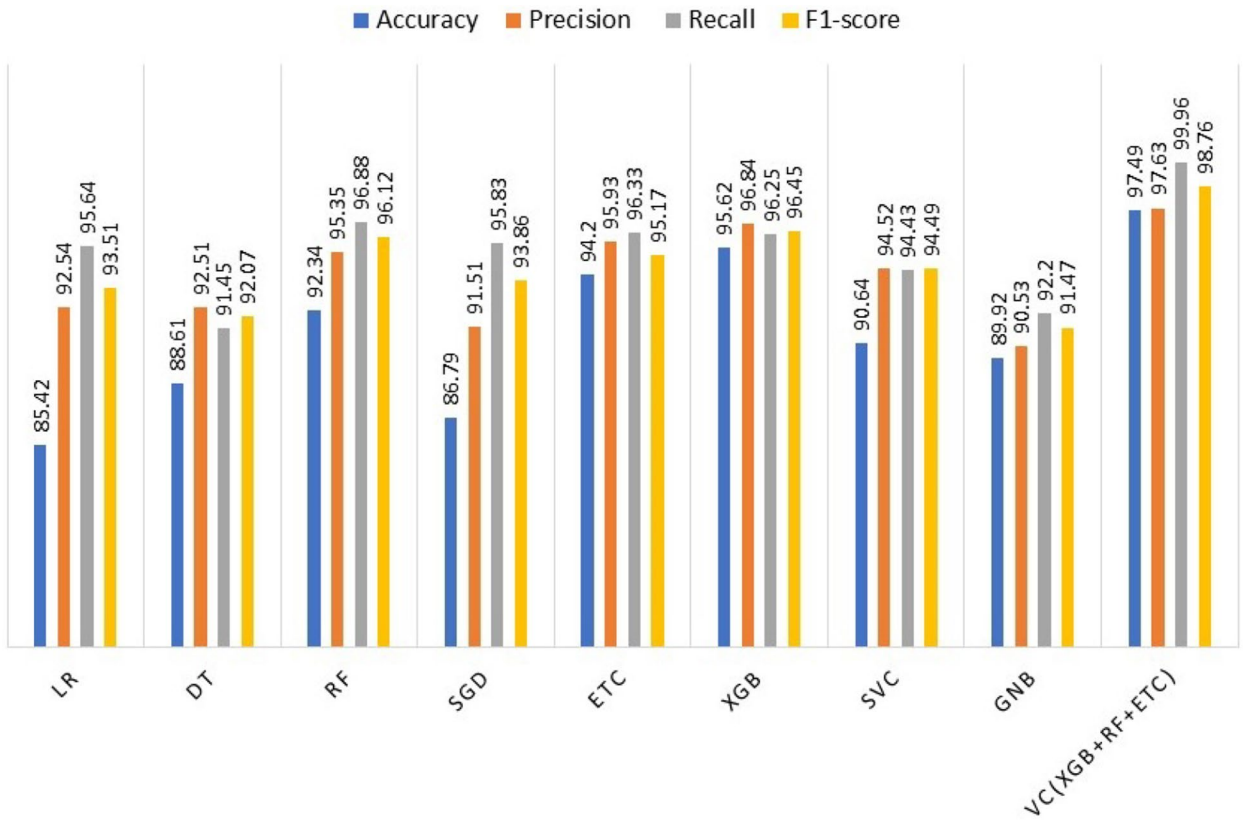

Fig. 4 Results of the KNN imputer machine learning models

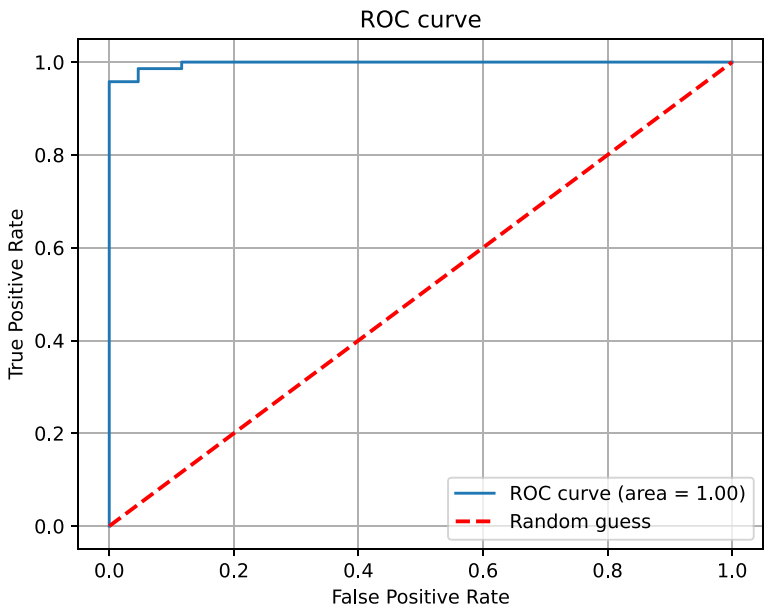

Fig. 5 ROC-AUC curve of the proposed model

showed exceptional performance with an accuracy of 98.59% and extremely high precision of 99.26%, recall of 99.75%, and F1-score of 99.45%. This indicates the effectiveness of ensemble techniques in improving model performance significantly. In summary, ensemble methods like RF, ETC, XGB, and VC showed superior performance across multiple evaluation metrics, highlighting their suitability for this classification task. However, the choice of model should consider the specific requirements and characteristics of the dataset. The ROC-AUC curve of the proposed model is shown in Fig. 5.

Comparison of the two experiments' machine learning models

We compared the performance of machine learning models using the KNN imputer with those that did not in order to evaluate its efficacy. In the second experiment, using the KNN imputer instead of the dataset with removed missing values resulted in a considerable improvement in the machine learning models' performance. Table 7 offers an extensive summary of the results produced by the machine learning models in both scenarios, making a detailed evaluation of their effectiveness possible. This comparison may help to clarify the advantages and results of utilizing the KNN imputer to enhance the models' prediction capabilities.

Results of the k-fold cross validation

The models were validated through k-fold cross-validation. Table 8 shows 5-fold cross-validation measures, which indicate the superiority of the proposed strategy in terms of recall, accuracy, precision, and F1 score over other models. This, combined with the minimal standard deviation, underscores the reliability and consistency of the proposed method. This shows that the proposed approach performs well across different folds, enhancing the method's robustness and reliability.

Table 7 Comparing the machine learning models' accuracy

Model	With KNN	Without KNN
DT	78.34	89.71
ETC	84.18	95.3
GNB	76.48	91.02
LR	74.87	86.52
RF	82.75	93.44
SGD	79.69	87.89
SVC	80.55	91.74
XGB	84.61	96.72
VC(XGB+RF+ETC)	81.13	98.59

Table 8 5-fold cross-validation results for the proposed system

Model	Accuracy	Precision	Recall(Sensitivity)	F1-Score
1st fold	97.62	98.23	99.71	99.22
2nd fold	97.35	98.44	99.84	99.33
3rd fold	98.74	99.77	101.08	100.91
4th fold	98.18	99.88	101.09	100.95
5th fold	98.08	98.25	99.96	99.43
Average	**97.99**	**98.91**	**100.33**	**99.97**

Performance comparison with existing studies

The accuracy of the comprehensive analysis was presented using several relevant research works that compare the performance of the proposed model with the current state-of-the-art models. The selected works were those against which the proposed paradigm was benchmarked to showcase the effectiveness and the differences with previous approaches. Analyzing the differences between the outcomes of the proposed model and the chosen state-of-the-art models provide important information about why the suggested technique performs better in terms of accuracy improvement. For example, an ensemble model named LTC was proposed in [16] and was rated as accurate to 85%. An ensemble of deep learning models was used in [24] to predict diabetes, with a maximum accuracy of 88.37% achieved. Similarly, [21] and [E] used the individual learning model LGBM and XGB to predict diabetes with 92.5% and 80% accuracy respectively. A performance comparison of the suggested model with the previous research is shown in Table 9. The outcomes clearly show that the suggested model performs better than the current models in a variety of performance criteria.

Significance of proposed model

In order to check the significance of the proposed model, we have applied this framework to another

Table 9 Comparison of proposed model with recent studies

Ref.	Year	Technique	Accuracy
[16]	2023	LTC (ensemble)	0.85
[20]	2021	LSTM	0.87
[21]	2022	LGBM	0.93
[24]	2022	CNN-BiLSTM	0.88
[25]	2019	DNN	0.86
[27]	2022	XGB	0.81
[28]	2023	GBC	0.92
[45]	2024	NN	0.79
[46]	2023	XGBoost XGBoost	0.83
[47]	2023	LR	0.84
Proposed	2024	**Ensemble VC(XGB+RF+ETC)**	**0.97**

independent dataset [48]. The Behavioral Risk Factor Surveillance System (BRFSS) is an extensive annual health-related telephone survey administered by the CDC. Since its inception in 1984, BRFSS gathers data from more than 400,000 Americans each year, focusing on various health-related risk behaviors, chronic conditions, and preventive service utilization. For this study, the 2015 dataset from Kaggle in CSV format was utilized. This dataset encompasses responses from 441,455 individuals and includes 330 features. These features comprise direct participant responses to survey questions as well as computed variables derived from individual participant data. The proposed framework gives 98.32% accuracy, 98.79% precision, 98.99% recall, and 98.89% F1-score respectively. These promising results show the stability and significance of the proposed model on all types of datasets to predict diabetes.

Discussion

Diabetes has become a common disease in the world in general and in underdeveloped nations in particular. Early detection and proper treatment care are crucial to reduce its impact. One of the best ways to detect diabetes early is by examining its symptoms where machine learning approaches can become significantly important. Various models are used for performance evaluation in comparison to the proposed approach. Results indicate that DT, LR and GNB show poor results with accuracy of 78.34%, 74.87%, and 76.48%, respectively. Although RF also uses DTs to make a forest, its performance is superior wit a 82.75% accuracy. the DT model is simple and easy to interpret, however, it is prone to overfitting which might lead to poor performance. On the contrary, RF reduces the overfitting and provides better performance. GNB works on the assumption that all features are independent which might not hold true in all practical cases. The assumption can lead to inaccuracies when calculating class probability. In the case of using selective features, it may perform poorly. LR model perform better in the case of linear relationship which is not the case for the current dataset. Although LR is known to perform well with binary outcomes, its performance is affected when using on a large dataset. Compared to DR, LR, and GNB models, boosting and ensemble models like XGB and RF show better results in the case of using the data containing missing values.

On the other hand, when the data, normalized using the KNN imputer is used, the performance of all models is improved. The lowest accuracy of 86.52% is achieved by the LR model. SGD has marginal difference with a 87.89% accuracy. RF, ETC, SVC, XGB, and GNB all obtain accuracy higher than 90%. The KNN imputer significantly improves the performance of the proposed ensemble model achieving an accuracy of 98.59%, precision of 99.26%, recall of 99.75%, F1 score of 99.45%, and MCC of 99.24%.

Limitations of proposed model

Every framework has some limitations, so the proposed model also faces some limitations like the performance of the proposed model highly dependent on the data quality. As KNN imputation and data mining outcomes heavily hinge on the completeness and accuracy of input data. Furthermore, the sensitivity to KNN parameters, such as the number of neighbors (K) and distance metrics, introduces challenges in optimizing imputation quality and prediction accuracy. Moreover, the complexity of data mining techniques used may obscure interpretability, hindering understanding of influential features in diabetes prediction. Scalability concerns arise from resource-intensive computations, potentially restricting the application to larger datasets or real-time settings.

Conclusion

The prevalence of diabetes has sharply increased in recent years, impacting millions of individuals worldwide. Diabetes-related problems could be considerably reduced with an early diagnosis and prompt treatment. It has been discovered that applying machine learning techniques increases detection accuracy in this way. In order to deal with missing values at the data preprocessing level, this research effort suggested a framework that uses an ensemble model for improved classification accuracy and KNN imputer. A high accuracy of 97.49% is shown in experimental findings utilizing the suggested One type of stacked ensemble voting classifier is the XGB + RF + ETC model suggesting that the KNN imputer and ensemble model can yield better results. A comparison with other state-of-the-art models demonstrates the superiority of the proposed model. This project is to investigate the stacking assembly of models for machine learning and deep learning as part of subsequent work. This approach seeks to improve the model's performance and yield more robust and generalized results, particularly on higher-dimensional datasets.

Authors' contributions

A.A. conceived the idea, performed data curation and wrote the original manuscript. A.A.A. conceived the idea, performed formal analysis and wrote the original manuscript. M.U. designed methodology, and performed formal analysis and data curation. E.A.A. designed methodology, dealt with software and performed visualization. S.A. dealt with software, performed visualization and investigation. T.H.K. performed investigation and visualization, acquired funding and I.A. supervised the work, performed validation and edited the manuscript.

Funding
The research is partially funded by Zhejiang Provincial Natural Science Foundation Youth Fund Project (Grant No. LQ23F010004) and the National Natural Science Youth Science Foundation Project (Grant No. 62201508). Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R348), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Availability of data and materials
The dataset used in this study is publicly available at the following links https://www.ncbi.nlm.nih.gov/taxonomy/.

Code availability
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 12 December 2023 Accepted: 28 August 2024
Published online: 27 September 2024

References

1. Gojka D. Diabetes: World Health Organization (WHO). 2019. https://www.who.int/health-topics/diabetes#tab=overview. Accessed 25 May 2023.
2. El-Sappagh S, Ali F, El-Masri S, Kim K, Ali A, Kwak KS. Mobile health technologies for diabetes mellitus: current state and future challenges. IEEE Access. 2018;7:21917–47.
3. Mertz L. Automated insulin delivery: taking the guesswork out of diabetes management. IEEE Pulse. 2018;9(1):8–9.
4. Klein HA, Meininger AR. Self management of medication and diabetes: Cognitive control. IEEE Trans Syst Man Cybern Syst Hum. 2004;34(6):718–25.
5. WHO. Diabetes: World Health Organization (WHO). 2023. https://www.who.int/news-room/fact-sheets/detail/diabetes. Accessed 25 May 2023.
6. Al Jarullah AA. Decision tree discovery for the diagnosis of type II diabetes. In: 2011 International conference on innovations in information technology. IEEE; 2011. pp. 303–7.
7. Kalyankar GD, Poojara SR, Dharwadkar NV. Predictive analysis of diabetic patient data using machine learning and Hadoop. In: 2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC). IEEE; 2017. pp. 619–24.
8. Ahamed BS, Arya MS, Nancy AOV. Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation. Adv Hum Comput Interact. 2022. https://doi.org/10.1155/2022/9220560
9. Wang Y, Wang C, Li K, Song X, Yan X, Yu L, et al. Recent advances of nanomedicine-based strategies in diabetes and complications management: Diagnostics, monitoring, and therapeutics. J Control Release. 2021;330:618–40.
10. Holzer R, Bloch W, Brinkmann C. Continuous glucose monitoring in healthy adults—possible applications in health care, wellness, and sports. Sensors. 2022;22(5):2030.
11. Weinstock RS, Aleppo G, Bailey TS, et al. The Role of Blood Glucose Monitoring in Diabetes Management. Arlington: American Diabetes Association; 2020. Available from: https://www.ncbi.nlm.nih.gov/books/NBK566165/, https://doi.org/10.2337/db2020-31.
12. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023;6(1):120.
13. Perveen S, Shahbaz M, Guergachi A, Keshavjee K. Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci. 2016;82:115–21.
14. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J. 2017;15:104–16.
15. Deberneh HM, Kim I. Prediction of type 2 diabetes based on machine learning algorithm. Int J Environ Res Public Health. 2021;18(6):3317.
16. Rupapara V, Rustam F, Ishaq A, Lee E, Ashraf I. Chi-Square and PCA Based Feature Selection for Diabetes Detection with Ensemble Classifier. Intell Autom Soft Comput. 2023;36(2):1931–49.
17. Saad A, Chen Z, Guo Y, Liu B. Enhanced Deep Learning-based Detection of COVID-19 on Chest X-ray Images. Multimed Tools Appl. 2020;79(35):25665–88. https://doi.org/10.1007/S11042-019-07820-W.
18. Ozturk T, Talo M, Yildirim EA, Baloglu UB, Yildirim O. A deep learning approach for COVID-19 imaging-based detection. Med Hypotheses. 2020;140:109684. https://doi.org/10.1007/s11042-024-18304-x.
19. Deng Y, Lu L, Aponte L, Angelidi AM, Novak V, Karniadakis GE, et al. Deep transfer learning and data augmentation improve glucose levels prediction in type 2 diabetes patients. NPJ Digit Med. 2021;4(1):109.
20. Butt UM, Letchmunan S, Ali M, Hassan FH, Baqir A, Sherazi HHR, et al. Machine learning based diabetes classification and prediction for healthcare applications. J Healthc Eng. 2021;2021.
21. Ahamed BS, Arya MS, Nancy AOV. Diabetes Mellitus Disease Prediction Using Machine Learning Classifiers with Oversampling and Feature Augmentation. Adv Hum Comput Interact. 2022(1):9220560. https://doi.org/10.1155/2022/9220560.
22. Pethunachiyar G. Classification of diabetes patients using kernel based support vector machines. In: 2020 International Conference on Computer Communication and Informatics (ICCCI). New York City: IEEE; 2020. p. 1–4.
23. Laila UE, Mahboob K, Khan AW, Khan F, Taekeun W. An ensemble approach to predict early-stage diabetes risk using machine learning: An empirical study. Sensors. 2022;22(14):5247.
24. Madan P, Singh V, Chaudhari V, Albagory Y, Dumka A, Singh R, et al. An optimization-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment. Appl Sci. 2022;12(8):3989.
25. Kannadasan K, Edla DR, Kuppili V. Type 2 diabetes data classification using stacked autoencoders in deep neural networks. Clin Epidemiol Glob Health. 2019;7(4):530–5.
26. Dutta A, Hasan MK, Ahmad M, Awal MA, Islam MA, Masud M, et al. Early prediction of diabetes using an ensemble of machine learning models. Int J Environ Res Public Health. 2022;19(19):12378.
27. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2022;10(1–2):1–10. https://doi.org/10.1049/htl2.12039.
28. Ganie SM, Pramanik PKD, Bashir Malik M, Mallik S, Qin H. An ensemble learning approach for diabetes prediction using boosting techniques. Front Genet. 2023;14:1252159.
29. Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access. 2020;8:76516–31. https://doi.org/10.1109/access.2020.2989857.
30. Torkey H, Awadallah M, Nour K. Machine Learning Model for Cancer Diagnosis based on RNAseq Microarray. Menoufia J Electron Eng Res. 2021;30(1):5–12. https://doi.org/10.21608/mjeer.2021.146277.
31. Patil S, Sherekar S. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. In: 2015 International Conference on Data Mining and Intelligent Computing (ICDMIC). Mumbai: IEEE; 2015. pp. 1–6. https://doi.org/10.1109/ICDMIC.2015.36.
32. Learning UM. Diabetes: World Health Organization (WHO). 2016. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed 05 May 2023.
33. Hafeez U, Umer M, Hameed A, Mustafa H, Sohaib A, Nappi M, et al. A CNN based coronavirus disease prediction system for chest X-rays. J Ambient Intell Humanized Comput. 2022;1–15.
34. Juna A, Umer M, Sadiq S, Karamti H, Eshmawi A, Mohamed A, et al. Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. Water. 2022;14(17):2592.
35. Zhang Y, Zhang H, Cai J, Yang B. A weighted voting classifier based on differential evolution. Abstr Appl Anal. 2014;2014. Wiley. https://doi.org/10.1155/2014/376950.

36. Brijain M, Patel R, Kushik M, Rana K. A survey on decision tree algorithm for classification. 2014.
37. Karim M, Missen MMS, Umer M, Sadiq S, Mohamed A, Ashraf I. Citation context analysis using combined feature embedding and deep convolutional neural network model. Appl Sci. 2022;12(6):3203.
38. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv (CSUR). 2002;34(1):1–47.
39. Zadrozny B, Elkan C. Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. 2002. New York: Association for Computing Machinery; p. 694–9.
40. Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. Stat Comput. 2017;27(3):659–78.
41. Rustam F, Ashraf I, Mehmood A, Ullah S, Choi GS. Tweets classification on the base of sentiments for US airline companies. Entropy. 2019;21(11):1078.
42. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern. 1991;21(3):660–74.
43. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
44. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. https://doi.org/10.1186/s12864-019-6413-7.
45. Guleria P, Srinivasu PN, Hassaballah M. Diabetes prediction using Shapley additive explanations and DSaaS over machine learning classifiers: a novel healthcare paradigm. Multimedia Tools Appl. 2024;83(14):40677–712.
46. Tasin I, Nabil TU, Islam S, Khan R. Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technol Lett. 2023;10(1–2):1–10.
47. Rastogi R, Bansal M. Diabetes prediction model using data mining techniques. Meas Sensors. 2023;25:100605.
48. Teboul A. Diabetes Health Indicators Dataset. 2023. https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset. Accessed 14 June 2024.

Publisher's Note