Accurate Diabetes Prediction Using a Robust Framework Based on a Hard-Voting Classifier

Article *in* Journal of Information Systems Engineering & Management · January 2025

DOI: 10.52783/jisem.v10i27s.4412

8 authors, including:

Khtam Al-meyah
University of Basrah
7 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Zaid Ameen Abduljabbar
Huazhong University of Science and Technology
189 PUBLICATIONS **1,921** CITATIONS
SEE PROFILE

Vincent O. Nyangaresi
University of Nairobi
194 PUBLICATIONS **4,002** CITATIONS
SEE PROFILE

Ali Hasan Ali
University of Basrah
126 PUBLICATIONS **1,291** CITATIONS
SEE PROFILE

Journal of Information Systems Engineering and Management
2025, 10(27s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Accurate Diabetes Prediction Using a Robust Framework Based on a Hard-Voting Classifier

Mohammed S. Hashim[1], Ghazwan Abdulnabi Al-Ali[1], Khtam AL-Meyah[2], Zaid Ameen Abduljabbar[1,3, 4], Vincent Omollo Nyangaresi[5,6], Ali Hasan Ali[7,8,9], Zaid Alaa Hussien[10], Abdulla J.Y. Aldarwish[1]

1Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Basrah 61004, Iraq.
2Department of Architecture Engineering, College of Engineering, University of Basrah, Basrah, 61004, Iraq.
3Department of Business Management, Al-Imam University College, Balad 34011, Iraq.
4Shenzhen Institute, Huazhong University of Science and Technology, Shenzhen 518000, China.
5Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science and Technology, Bondo 40601, Kenya.
6Department of Applied Electronics, Saveetha School of Engineering, SIMATS, Chennai, Tami lnadu, 602105, India.
7Department of Mathematics, College of Education for Pure Sciences, University of Basrah, Basrah, 61004, Iraq.
8Technical Engineering College, Al-Ayen University, Thi-Qar 64001, Iraq.
9Institute of Mathematics, University of Debrecen, Pf. 400, H-4002 Debrecen, Hungary.
10Information Technology Department, Management Technical College, Southern Technical University, Basrah 61004, Iraq

ARTICLE INFO

Received: 30 Dec 2024

Revised: 14 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

Diabetes is one of the most dangerous diseases that a significant number of people worldwide. An accurate and timely diagnosis of diabetes helps to minimise the overall prevalence of the illness and save the lives of diagnosed individuals. Researchers have proposed a number of diagnostic procedures for the identification of diabetes, but such methods should be improved to guarantee accurate and effective diagnosis. This study aims to develop accurate and timely predictions about diabetes to save the lives of diabetic patients. A three-stage integrated methodology was developed for accurate diagnosis and applied to a Pima Indian Diabetes (PID) Dataset. SMOTE technique was used to balance the dataset and ensure the lack of bias during the training process. The proposed methodology is mainly based on a hard voting classifier that predicts whether a patient will develop diabetes or not. Finally, a set of metrics, namely, accuracy, k-fold cross validation, AUC, precision, recall and f1 score, was used to test the performance of the proposed methodology in diagnosing diabetes. Results showed the superiority of our proposed methodology, with values of 90%, 83.9% with 10-fold cross-validation, 0.901, 0.871, 0.926 and 0.898, respectively.

Keywords: Diabetic Disease, Artificial intelligence, Machine Learning models, Prediction, Hard Voting Classifier, SMOTE, Cross-validation.

INTRODUCTION

Diabetes, which is marked by excessive amounts of sugar in the blood, is one of the most frequent disorders that affect the endocrine glands. It is a persistent condition that manifests itself when the pancreas is unable to generate sufficient insulin and make appropriate use of the insulin that it does produce (Khan, Zeb, Al-Rakhami, Derhab, & Bukhari, 2021). Diabetes can be in a few different forms, the most prevalent of which are type 1 and type 2, as well as pre-diabetes and gestational diabetes (Knight & Nigam, 2017). Many economic and social changes have occurred over the years but have had a major impact on people's health and lives. Changes in lifestyles, poor diet, increased smoking, decreased levels of physical activity and inactivity, Western consumption habits and overweight or obesity are important factors that can play a role in the development of hyperglycaemia, which contributes significantly to the increase in cases of diabetes (Alshammari, Atiyah, Daghistani, & Alshammari, 2020).

The World Health Organisation indicates that diabetes is a public health problem and a very serious and widespread reality. According to statistics conducted by the World Health Organisation since 1990, more than two

hundred million people around the world live with diabetes. This number has quadrupled during the current time, as more than eight hundred million people live with this disease. This terrifying increase is due to poor eating habits, inactivity and increasing rates of obesity. According to the World Health Organisation's expectations, diabetes will be the sixth largest cause of death in 2030 (Abusaib et al., 2020).

Early detection of diabetes has a great impact and importance, as early measures can be taken to avoid increasing the risk and save people's lives by improving the lifestyle of the infected person and modifying eating habits in addition to practicing physical exercises (Kousar, 2019). Artificial intelligence (AI) is very important in the field of health care, including diabetes, because it has created a unique breakthrough in terms of detecting and treating the disease, as it helps the doctor make decisions through accurate diagnosis and early intervention (Contreras & Vehi, 2018). In addition, machine learning (ML) techniques are very important in predicting the disease by examining a set of data that plays an important role in determining whether a person has the disease or not, such as age, smoking status, blood glucose levels, etc. (Ismail & Materwala, 2021).

Studies have encountered difficulties in terms of the precision of the diagnosis. The reason for this might be that the dataset was not sufficiently balanced while the AI models were being trained, or it could be that the AI algorithms are not well matched to this dataset. Both of these possibilities are possible. Therefore, it is necessary in this paper to enhance the accuracy of early diagnosis, which in turn serves to increase the likelihood that the patient will survive, as explained in our contributions below:

- A pre-processing set was carried out to enhance the quality of the data for diabetes prediction, and the analysis of the dataset served as the foundation for this.
- Our efforts concentrate on dataset balancing with SMOTE technology to provide an equitable, unbiased model, hence ensuring accurate learning for machine learning models, which enhances diagnostic precision.
- As a classification model, we proposed the hard voting classifier, which is made up of four separate models for classification purposes. Due to the fact that this classifier is able to merge these four models into a single model that has the power of these models, the accuracy of classification has been dramatically enhanced.
- We assessed a machine learning model's performance and durability using other metrics, including k-fold cross-validation, which had not been used in prior works. These measurements were used to complete the assessment. This process assesses how well the model generalises to a dataset that is not related to it, which is critical for ensuring that the model performs well on previously unseen data.

The other sections of the paper are arranged as follows: the remaining parts of the paper are presented: Section 2 contains all of the pertinent publications that were derived from earlier investigations that were conducted on the prediction of stroke. The methodology that has been proposed is discussed in 3. Section 4 is where the evaluation and presentation of the findings and discussion take place. The conclusions are discussed in Section 5.

RELATED WORKS

Confirming a diagnosis is one of the most important aspects of the healthcare industry, making it one of the most important areas to apply artificial intelligence. As a result, various researchers have utilised artificial intelligence techniques for the early identification of diabetes in order to increase the accuracy and efficiency of categorisation strategies. The Pima Indian Diabetes (PID) Dataset was used in a number of relevant research projects that were presented in this part. These studies utilised artificial intelligence approaches to diagnose diabetes, as explained below.

Tigga and Garg (Tigga & Garg, 2020) proposed the development of a model that would predict the risk of diabetes by utilising machine learning techniques. These techniques include Naive Bayes (NB), K Nearest Neighbour (KNN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM). These techniques are extremely accurate, which is essential in the field of medicine. In addition to performing an online and offline questionnaire that consists of 18 questions relevant to lifestyle, genetics, health, and other topics, these approaches were applied to a trustworthy dataset that was obtained from the Kaggle (PID) Dataset. According to the findings of this research, the random forest method achieved the maximum level of accuracy, which was 75%. Soni and Varma (Simanto et al., 2022) proposed an optimal model design from machine learning techniques (Decision Tree (DT),KNN, Gradient Boosting (GB), LR, SVM and RF.) to predict the risk of diabetes, and applied this model to a

diabetes dataset (PID dataset). According to the findings of this research, the random forest algorithm achieved the maximum performance accuracy, around 77%. Kishore et al. (Naveen Kishore, Rajesh, Vamsi Akki Reddy, Sumedh, & Rajesh Sai Reddy, 2020) worked on early diagnosis of diabetes using artificial intelligence methods. They implemented five algorithms (SVM, KNN, RF, DT, and LR) to the diabetes dataset that was acquired from the Kaggle website (PID) dataset. Based on the outcomes of the investigation, it was determined that the random forest model had the maximum performance accuracy, which was around 74.4%. To project the chance of diabetes development, suggested the use of machine learning classification methods and the implementation of numerous algorithms, namely (SVM, LR, KNN, RF, NB), on a reliable data set for diabetes. They also recommended the use of numerous metrics, including accuracy, recall, precision, error rate, F-measure. The logistic regression approach turned out to have the highest performance accuracy about 79.17%.

Using three distinct classifiers that of multilayer perceptron (MLP), RF, and LR, Butt et al. (Butt et al., 2021) developed an effective model based on machine learning methods. Given its huge relevance in life as it compromises the person's life and fuels other deadly illnesses like heart, kidney, and nerve damage. This helps one forecast the danger of diabetes. They also used linear regression (LR), moving averages (MA), and long- and short-term memory (LSTM). This paper revealed that the long- and short-term memory (LSTM) algorithm had the maximum performance accuracy of 87.26%. Sivaranjani et al. (Sivaranjani, Ananya, Aravinth, & Karthika, 2021) presented the assessment of diabetes risk and the identification of potential infection threats using machine learning methods, specifically employing the SVM and RF techniques on a validated dataset (PIDD) after data preprocessing. They found the characteristics influencing decision-making by the use of forward and backward feature selection. The technique for dimensionality reduction in principal component analysis (PCA) is examined after the identification of certain characteristics. The study's findings showed that the random forest achieved the maximum performance accuracy at 83%. Using LR and DT, Joshi and Dhakal (Joshi & Dhakal, 2021) constructed an optimal model for early prediction of stroke risk. This was done due to the significance of the issue since prompt diagnosis and prediction provide a chance to implement suitable preventative tactics and therapies. A number of characteristics, such as age, glucose levels, and body mass index (BMI), were included in the dataset of Pima Indian women, which they used to apply the model. LR was found to have the best performance accuracy, with an estimated value of 78.26%, according to the findings of this study.

Krishnamoorthi et al. (Krishnamoorthi et al., 2022) suggested an ideal model using machine learning methods (SVM, RF, KNN, LR) based on Decision Tree (DT) to predict diabetes risk and discover disease-related characteristics. This model was applied to the PID dataset, which includes a variety of parameters such as glucose level, insulin, age, gender, and so on. The research found that LR has the greatest performance accuracy (83%). YAKUT (Yakut, 2023) used a PID dataset to develop an optimal model for predicting stroke; he separated the database into two thirds training data and one-third testing data and processed them. After processing the data, it is fed into machine learning models (Gaussian Process Classifier, RF Classifier, and Extra Trees Classifier) that use five fold cross-validation to predict whether or not people have diabetes. In this research, they employed numerous measures to assess the model's performance, including precision, recall, F-score, accuracy, ROC, and AUC. The random forest achieved the highest performance accuracy, around 81.71%. In Tripathi et al. (Maurya & Jain, 2023) paper, they proposed a robust model for diabetes prediction, where they analysed, studied and applied several algorithms on the PID dataset after studying and processing them based on the existing features and characteristics. The model was designed from machine learning algorithms (Soft voting, SVM, KNN, Gradient Boosting (GB), AdaBoost classifier, LR). The finding of this research explains that logistic regression obtained the highest performance accuracy estimated at 84.3%.

Talukder et al. (Talukder et al., 2024) employed machine learning approaches to predict diabetes risk early, using eight machine learning algorithms applied to four distinct datasets. They also preprocessed the data and balanced it by oversampling. The research found that the random forest method attained an accuracy of 86% and 98.48% for datasets 1 and 2, respectively. For datasets 3 and 4, the extreme gradient boosting technique and decision tree achieve 99.27% and 100% accuracy, respectively.

Furthermore, we see that earlier studies have struggled with the accuracy of the diagnosis. This can be because the AI algorithms are not well adapted to this dataset, or it might be because the dataset was not sufficiently balanced during the training of the AI models. In order to save the patient's life, these investigations must increase the precision of early diagnosis. This paper's primary contributions include balancing the dataset to remove bias that

arises during training, which improves the performance of the models, and beginning with a series of preliminary processing, the most crucial of which is the treatment of missing data. Furthermore, we use the hard voting classifier to enhance the prediction outcome, as it can merge many models into a single model that carries the strength of each model. Additionally, we assessed a machine learning model's robustness and performance using novel metrics—like k-fold cross-validation—that had not been covered in earlier research. To ensure that the model performs well on data that has never been seen before, this process aids in assessing how well the model generalises to a dataset that is unrelated to it.

METHODOLOGY

In this paper, we propose an effective methodology for predicting diabetes using a Pima Indian Diabetes dataset (PIDD). This methodology consists of three stages: pre-processing, splitting and prediction. All these successive stages seek to achieve the best results in terms of the accuracy of predicting diabetes. Figure. 1 shows this methodology.

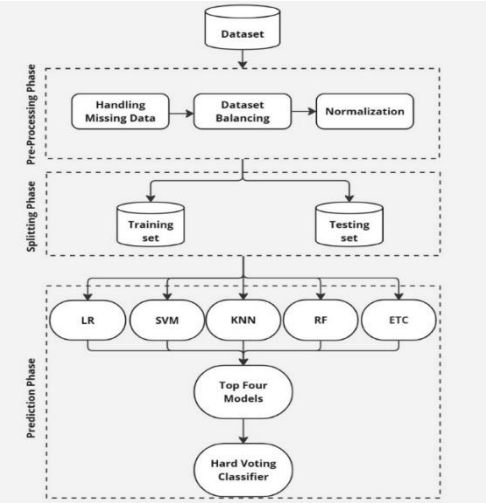


Figure.1. The framework of the proposed methodology.

A. Dataset Description

This research centers on the PID Database as the focal point of our suggested technique. This dataset, gathered by the National Institute of Diabetes and Digestive and Kidney Diseases, is extensively used in machine learning for categorisation purposes. This dataset focuses on a group of Pima Indian women who are at least 21 years old. The dataset comprises 768 records with 8 medical predictor variables (features) that are related to medical and demographic characteristics and 1 target variable (https://www.kaggle.com/uciml/pima-indians- & Diabetes-database, n.d.). All features are numerical values with different ranges. Figure. 2 displays the features of the dataset and some records.

# Pregnancies Number of times pregnant	# Glucose Plasma glucose concentration a 2 hours in an oral glucose tolerance test	# BloodPressure Diastolic blood pressure (mm Hg)	# SkinThickness Triceps skin fold thickness (mm)	# Insulin 2-Hour serum insulin (mu U/ml)	# BMI Body mass index (weight in kg/(height in m)^2)	# DiabetesPedigre... Diabetes pedigree function	# Age Age (years)	# Outcome Class variable 268 of 768 are others are 0
0 17	0 199	0 122	0 99	0 846	0 67.1	0.08 2.42	21 81	0
2	108	80	0	0	27	0.259	52	1
7	136	74	26	135	26	0.647	51	0
5	155	84	44	545	38.7	0.619	34	0
1	119	86	39	220	45.6	0.808	29	1
4	96	56	17	49	20.8	0.34	26	0
5	108	72	43	75	36.1	0.263	33	0
0	78	88	29	40	36.9	0.434	21	0
0	107	62	30	74	36.6	0.757	25	1
2	128	78	37	182	43.3	1.224	31	1
1	128	48	45	194	40.5	0.613	24	1

Figure. 2. The PID dataset.

B. Pre-processing phase

At this stage, we will perform a set of preliminary processing, which are Handling Missing Data, Dataset Balancing and Normalisation. These operations that are performed on the dataset used help in improving the quality of the data by handling missing data and eliminating bias that occurs during the training process and standardizing the features, all of which play a major role in improving the accuracy of classification.

Handling Missing Data

The used dataset (PIDD) for diabetes contains a set of missing data which should be handled because it causes problems during the training process and thus negatively affects the classification process. In this dataset the missing values are expressed as "0" except for the feature "Pregnancies" where the value "0" is within the range. Therefore, in this work we use the mean as a procedure to fill the missing data. Table I shows the features which contain missing values expressed as "0" and their number.

Table 1. The features which contain missing values.

Feature	Number of missing values
Glucose	5
Blood Pressure	35
Skin Thickness	227
Insulin	374
BMI	11

Dataset Balancing

The used dataset for diabetes contains 768 samples. 500 samples represent non-diabetic cases, and 268 samples represent diabetic cases. This discrepancy between the number of diabetic and non-diabetic cases poses a major challenge in the process of predicting the disease. This discrepancy leads to bias by the models used towards non-diabetic cases during the training process because their number is almost double the number of diabetic cases. Therefore, the models will train more on non-diabetic cases, which leads to bias. This bias in training will negatively affect the disease prediction process, and thus we do not get good classification accuracy(Mavrogiorgos, Kiourtis, Mavrogiorgou, Menychtas, & Kyriazis, 2024). Therefore, to solve this problem, we used the SMOTE technique. The SMOTE technique relies mainly on the KNN algorithm, as this technique creates new samples and does not duplicate existing samples. Smoot first finds the nearest neighbours of a sample of its choice and draws longitudinal lines between the nearest neighbours of this sample and the original sample, then creates new samples on the lines connecting them (Letteri, Di Cecco, Dyoub, & Della Penna, 2020), as detailed in Figure. 3. The primary justification for using the SMOTE approach to equilibrate the dataset is that the SMOTE approach equalizes the number of positive samples with those of negative samples to achieve data balance. Also, SMOTE produces synthetic instances instead of replicating existing ones, hence mitigating the danger of overfitting (Hashim & Yassin, 2023).

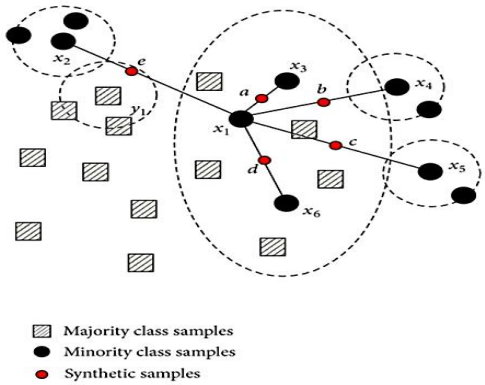
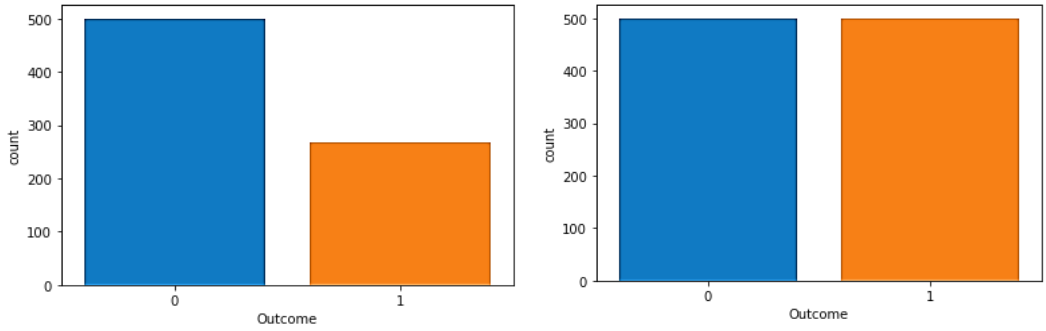


Figure. 3. SMOTE technique (Letteri et al., 2020).

After applying the SMOTE to the diabetes dataset used, the count of samples with the disease becomes equal to the count of samples without the disease and is 500 for both, as shown in Figure.4.



(a) Before (a) After

Figure. 4. The sample counts of dataset a) before balancing, b) after balancing.

Normalisation

We used StandardScaler to normalize the features because to the potential presence of outliers in the dataset, which might cause individual features to behave anomalously if the dataset is not regularly distributed (Sevilla, 2012). Features possess several dimensions and scales, therefore necessitating their scaling for algorithm-dataset modelling. Inconsistencies in the scales of data components complicate dataset modelling. Also, the correlation between misclassification error and accuracy skews the prediction outcomes. Therefore, data scaling is necessary before modelling (Ferreira, Le, & Zincir-Heywood, 2019).

C. Splitting phase

After performing a series of preliminary operations on the diabetes dataset, the data was ready to be passed to the machine learning models for training and testing their performance. Therefore, we followed two methods to divide the data into two parts, the first part for training and the second part for testing. The first division approach used in our research allocates 80% for training and 20% for testing. The second method is referred to as k-fold cross-validation. According to this approach, the dataset is divided into k equal-sized subsets, or "folds." After training on k-1 folds, the model is evaluated on the remaining folds. K-fold cross-validation reduces Overfitting by using different portions of the data for testing in each fold, ensuring that the model is evaluated on various subsets of the data (Gorriz, Segovia, Ramirez, Ortiz, & Suckling, 2024).

D. Prediction phase

In this phase, an in-depth description of the models that were used in our research is provided here as well as the mechanism followed to predict whether a patient has diabetes or not through the data entered for each patient by selecting the best-performing models and passing these models to the voting classifier, as explained in detail below.

Logistic Regression

It is a machine learning and statistical method. Classification tasks are handled by logistic regression (LR), which uses the logistic function to estimate the probability that a given input belongs to a certain class (Biostatistics, 2007). For the purpose of predicting binary values (either 0 or 1), this statistical model is utilised. The dichotomous dependent variable, which is frequently referred to as the response variable, and the independent variable, which is sometimes referred to as the predictor variable, were both subjected to LR in order to assess the relationship between the two variables. It uses the sigmoid function to arrange the prediction to be between (0-1). Furthermore, it addresses multi-class classification using techniques such as One against the Rest (OvR) or multinomial logistic regression (Modhugu & Ponnusamy, 2024).

Support Vector Machine

It is a crucial supervised machine learning technique used in classification and regression tasks, consists of two types: the first is linear, which is employed when the data can be separated linearly and is comparatively easy and computationally efficient, The second kind is non-linear since it involves the use of a specialised function in order to partition data in a space with a high dimension (Bhavsar & Panchal, 2012). The Support vector machine (SVM) algorithm, which stands for support vector machine, is extensively used in a variety of domains, including text classification, picture classification, spam filtering and many more. The fact that it is frugal in memory

consumption and works well in high-dimensional areas are two of its advantageous features (FİDAN, UZUNHİSARCIKLI, & ÇALIKUŞU, 2019).

K-Nearest Neighbours

K-Nearest Neighbours (KNN) is classified as a fundamental machine learning algorithm. The method is considered non-parametric, and its effectiveness depends on the specific instance. This demonstrates that it refrains from making any assumptions regarding the distribution of the fundamental data and instead utilises the data directly during the prediction process (Suyal & Goyal, 2022). It operates in two phases: the training phase and the prediction phase, the anticipated outcome varies based on the task type since classification involves selecting the test point according to the predominant class among the neighbours, while regression analysis is carried out by selecting the test point, which is the mean of the values of its neighbours (Uddin, Haque, Lu, Moni, & Gide, 2022).

Random Forest

Random forest, sometimes known as RF, is a supervised ensemble machine learning technique that is commonly used for classification and regression applications. It integrates multiple decision trees to enhance prediction accuracy and mitigate overfitting (Adetunji et al., 2021). The random forest in its first stage works on collecting bootstrapping, where it randomly creates a number of data sets and trains each tree in the forest on a different data set with the possibility of repetition and replacement. After that, the decision tree is built, and then the results of the trees are collected, where each tree gives a vote to a specific class. The random forest is strong in reducing overfitting and deals with high dimensions (Chen, Wu, Chen, Lu, & Ding, 2022).

Extra Trees Classifier

Extra Trees Classifier (ETC) is an ensemble machine learning technique, categorised under ensemble methods that use decision trees, akin to random forests as it is used for categorisation assignments, recognizing a collection of decision trees that are randomly partitioned and may be trained concurrently since the trees are created separately. It is regarded as one of the most successful algorithms due to its ability to handle high-dimensional data while avoiding overfittin (Ampomah, Qin, & Nyame, 2020).

Hard Voting Classifier

A Hard Voting classifier is a kind of ensemble learning approach that involves the combination of numerous classifiers, with each model casting a "vote" for the class that is expected to be correctly identified. In accordance with Figure. 4, the final output of the ensemble is the category that obtains the majority of votes. This strategy is helpful in circumstances in which a single model may not be dependable enough, so a group of models can provide more accurate predictions than a single model alone (Jabbar, 2024). In it, two or more base models are trained on the same dataset in order to create a hard voting classifier. These models may be of the same or distinct kinds of classifiers, and they may be trained on the same or separate subsets of features. Each of these two possibilities is possible (Shareef & Kurnaz, 2023). The key benefit of using a hard voting classifier is that combining multiple classifiers reduces the risk of poor performance from any single model. Also, individual models tend to overfit; combining them can smooth out overfitting errors, especially if the models are diverse (Shi, Xu, Li, & Li, 2024).

After splitting the dataset into two sets one for training models and the other for evaluating the models' performance we feed the data to the models LR, SVM, KNN, RF, ETC, and others for both training and testing. After testing the performance of the five models mentioned, the poor-performing model is discarded, and the remaining four models are passed to the hard voting classifier to obtain the final prediction of whether the patient will develop diabetes or not, as shown in Figure. 5.

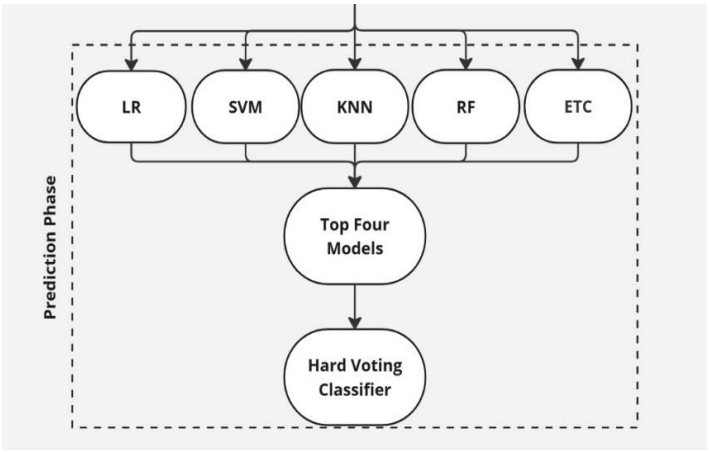


Figure. 5. Predication phase.

RESULTS AND DISCUSSION

In order to demonstrate how effectively the machine learning models perform in these trials, we make use of a number of different metrics. Accuracy, K-fold cross validation, F1 score, precision, recall, area under the curve (AUC) and ROC curves are some examples of these. The performance of the models is shown in the first experiment, which takes place before the process of balancing the data sets. On the other hand, in the second experiment, we provide the findings that were obtained after the balancing process, and then we explain the significance of the balancing process in terms of enhancing the performance of the models. In conclusion, we evaluate the performance of our suggested model by contrasting it with the results of earlier research conducted in the same area and dataset.

A. Experiment (1)

This experiment made use of the dataset that was first collected. A series of early processing operations were carried out on it, with the exception of the balancing procedure, which consisted of filling in missing data and using normalisation principles. Following that, the dataset was partitioned; twenty percent of it was put aside for testing, and 80% was set aside for training. The data that was used to train the models on the dataset was then utilised to test the performance of the models. After testing the performance of the models, the lowest-performing model is dropped, and the remaining models are passed to the hard voting classifier. to test their performance and demonstrate their importance in integrating more than one model and benefiting from the strengths of the models passed to it to obtain the best classification accuracy. The results of the performance of the models that were applied to the test data are shown in Table 2.

Table 2. A performance analysis of the models before the dataset is balanced.

Model	Accurcy (%)	F-1 Score	Precision	Recall
LR	79.87	0.652	0.691	0.617
SVM	80.52	0.651	0.718	0.596
KNN	81.81	0.682	0.732	0.638
RF	81.17	0.695	0.688	0.702
ETC	80.52	0.688	0.673	0.702
Hard Voting Classifier [SVM-KNN-RF-ETC]	**83.12**	0.8	0.683	0.596

Through the results presented in the table above regarding the performance of the models used, we note that the top four models in terms of performance are SVM, KNN, RF and ETC. Therefore, these four models were passed to the hard voting classifier to test its performance and demonstrate its effectiveness in improving the classification

accuracy. The results showed a clear superiority of the hard voting classifier in terms of performance over the other models, as it achieved an accuracy of 83.12%. Here, the importance of using the voting classifier appears, as each model has its strengths and weaknesses, and when combined, they can compensate for each other's errors.

To provide a more reliable measure of model performance than a single train-test split, we use k-fold cross validation to measure the model's performance. Where by averaging results over multiple folds, k-fold cross-validation tends to generalize better than models trained on a single training set. Also, we use AUC to evaluate the model across all possible thresholds, which gives a comprehensive view of the model's performance. Table 3 displays the k-fold cross validation and AUC scores for each model.

Table 3. The k-fold cross validation and AUC scores for models.

Model	10 _ Fold (%)	AUC
LR	77.1%	0.652
SVM	76.04%	0.651
KNN	72.66%	0.682
RF	77.47%	0.695
ETC	76.69%	0.688
Hard Voting Classifier [SVM-KNN-RF-ETC]	**77.82%**	**0.8**

Through this experiment, we found that the hard voting classifier obtained the highest accuracy in predicting diabetes whether the patient has the disease or not, but in this experiment, there remained an important problem that greatly affects the performance of the models, which is the bias problem. Since the number of cases with the disease is not equal to the number of cases without the disease, the bias problem will appear during the model training process, which caused problems and limitations towards the performance of the models, which will be dealt with and addressed in the next experiment.

B. Experiment (2)

This experiment made use of the dataset after balancing it by using SMOTE technique that was first collected. A series of early processing operations were carried out on it, which consisted of filling in missing data and using normalisation principles. Following that, the dataset was partitioned; twenty percent of it was put aside for testing, and 80% was set aside for training. The data that was used to train the models on the dataset was then utilised to test the performance of the models. After testing the performance of the models, the lowest performing model is dropped, and the remaining models are passed to the hard voting classifier to test their performance and demonstrate their importance in integrating more than one model and benefiting from the strengths of the models passed to it to obtain the best classification accuracy. The results of the performance of the models that were applied to the test data are shown in Table 4. The main goal of this experiment is to show how important it is to balance the dataset in order to get rid of the bias that comes from the difference between the number of positive and negative cases during training, where this bias makes disease prediction less accurate.

Table 4. A performance analysis of the models after balancing dataset.

Model	Accurcy (%)	F-1 Score	Precision	Recall
LR	80.5	0.804	0.769	0.842
SVM	85.5	0.854	0.817	0.895
KNN	83.5	0.842	0.772	0.926
RF	86.5	0.864	0.827	0.905
ETC	88.5	0.884	0.846	0.926
Hard Voting Classifier [SVM-KNN-RF-ETC]	**90**	0.898	0.871	0.926

A confusion matrix for each model is included in Figure. 6 for the purpose of discussion and analysis of the findings.

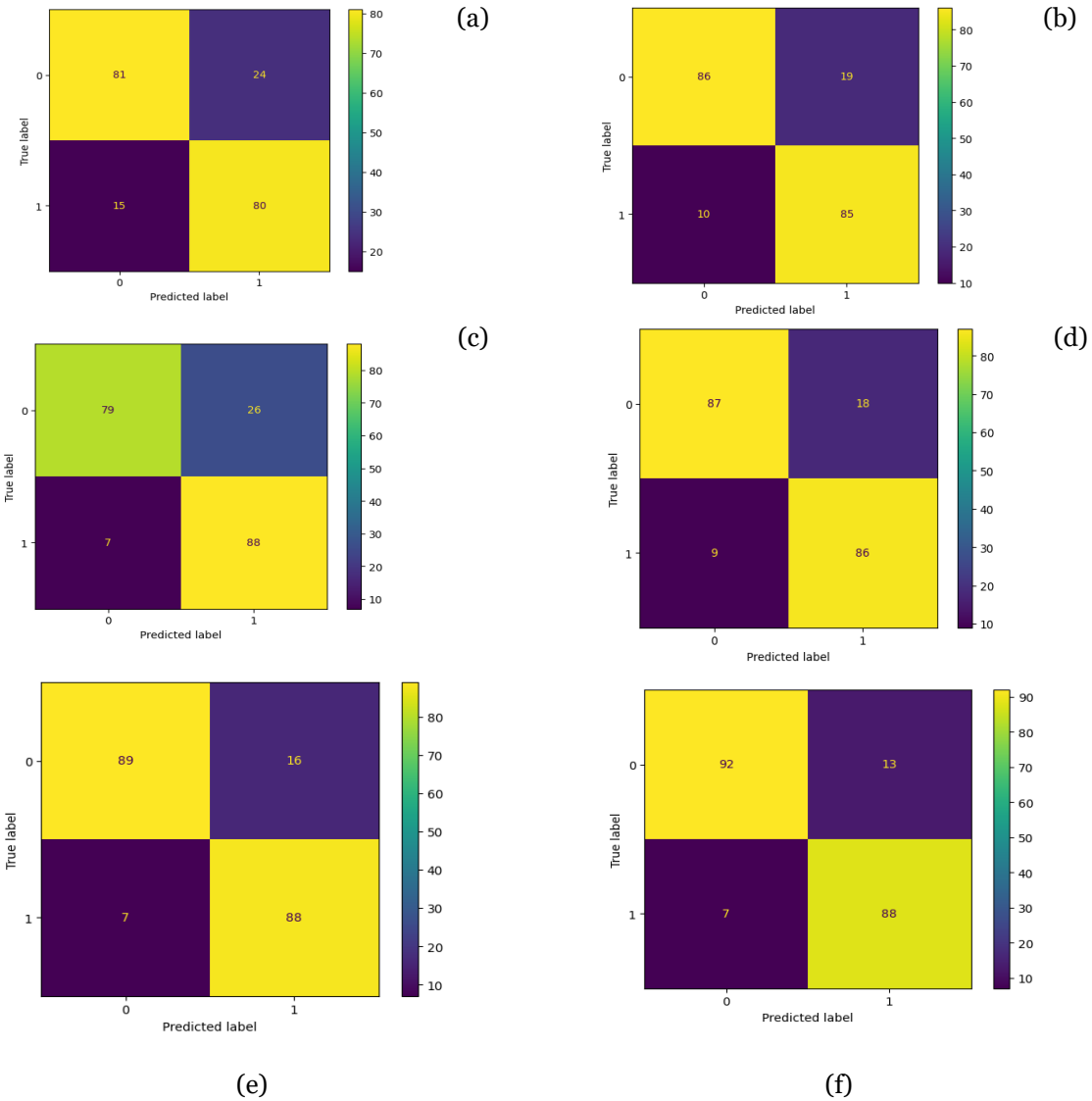


Figure.6. The confusion matrices for a) LR model, b) SVM model, c) KNN model, d) RF model, e) ETC model, f) Hard Voting Classifier.

We make use of k-fold cross validation to evaluate the performance of the model in order to give a more trustworthy measurement of the model's performance than a single train-test split currently offers. The area under the curve (AUC) is another method that we use to assess the model across all of the available thresholds. This provides a complete perspective of the performance of the model. The k-fold cross validation and area under the curve (AUC) scores for each model are shown in Table 5.

Table 5. The k-fold cross validation and AUC scores for models after balancing dataset.

Model	**10 _ Fold (%)**	**AUC**
LR	72.8	0.807
SVM	69.6	0.857
KNN	72.4	0.839
RF	80.8	0.867
ETC	83	0.887
Hard Voting Classifier [SVM-KNN-RF-ETC]	**83.9**	**0.901**

From Table 4, we notice a significant improvement in the performance of the models used. This improvement is due to the balancing of the data set, which eliminated the bias that occurs during the training process towards the majority class. We also notice in the same table that the hard voting classifier gave the highest classification accuracy after excluding the LR model as it is less accurate and passing the remaining four models, as the voting classifier obtained 90%. We also notice from Table 5 the clear superiority of the soft voting classifier in terms of the k-fold value after giving the value of k equal to 10, as it obtained 83.9%. The same applies to the AUC value, as it obtained a value equal to 0.901, as shown in Figure. 7.

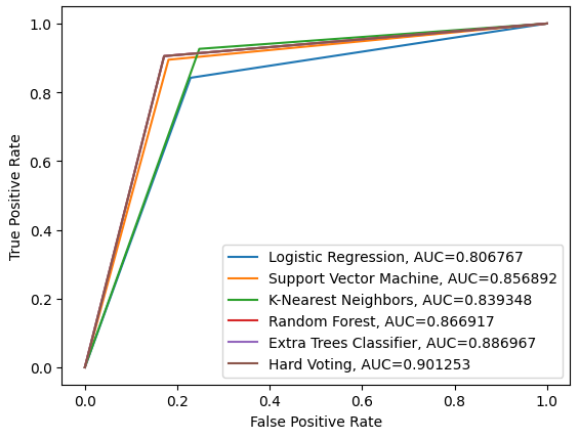


Figure. 7. The ROC curves and AUC for models.

In order to highlight the significance of our findings and the contributions that we have made to both our work and the technique that we have presented; we need to compare the performance outcomes of our proposed approach with performance results from earlier research. As a consequence of this, the comparisons are shown in Table 6.

Table 6. A comparative analysis of performance with analogous studies.

Author	Year	Dataset	Best Model	Accuracy (%)
Tigga (Tigga & Garg, 2020)	2020	PID Dataset	Random Forest	75 %
Soni (Simanto et al., 2022)	2020	PID Dataset	Random Forest	77 %
Kishore (Naveen Kishore et al., 2020)	2020	PID Dataset	Random Forest	74.4 %
Sahoo (Rani, Lamba, Sachdeva, Bathla, & Aledaily, 2020)	2020	PID Dataset	Logistic Regression	79.17 %
Butt (Butt et al., 2021)	2021	PID Dataset	LSTM	87.26 %
Sivaranjani (Sivaranjani et al., 2021)	2021	PID Dataset	Random Forest	83 %
Joshi (Joshi & Dhakal, 2021)	2021	PID Dataset	Logistic Regression	78.26 %
Krishnamoorthi (Krishnamoorthi et al., 2022)	2022	PID Dataset	Logistic Regression	83 %
YAKUT (Yakut, 2023)	2023	PID Dataset	Random Forest	81.71 %
Tripathi (Maurya & Jain, 2023)	2023	PID Dataset	Logistic Regression	84.3 %
Talukder(Talukder et al., 2024)	2024	PID Dataset	Random Forest	86%
The Proposed Methodology	**2024**	**PID Dataset**	**Hard Voting Classifier**	**90 %**

The inherent benefit of our suggested technique is shown in the table that is located above. A number of preliminary treatments that were designed to improve the quality of the data have been implemented, which makes this clear. Additionally, we have effectively picked the ideal classification model, represented by a hard voting classifier, which attained the maximum classification accuracy. The classification accuracy that we attained was

90%, which is much higher than the accuracy that was reached by a group of previous research that used the same dataset.

CONCLUSION

Diabetes, one of the most serious illnesses, affects many individuals worldwide. Accurate and quick diabetes diagnosis reduces its prevalence and saves lives. Despite the fact that researchers worldwide have offered several diagnostic approaches for this condition, the current methods require refinement to provide an accurate and successful diagnosis. This paper aims to generate accurate and timely diabetes forecasts to save diabetic patients' lives. We provide a three-stage comprehensive diabetes diagnostic methodology in this paper. A PID dataset was analysed using this approach. To avoid bias during training, the SMOTE approach was utilised to balance the uneven dataset. This methodology uses a hard voting classifier to predict whether a patient will get diabetes based on his data. Finally, accuracy, k-fold cross validation, AUC, precision, recall and f1 score were utilised to evaluate our diabetes diagnosis technique. Our methodology outperformed existing research with 90%, 83.9% with 10-fold cross-validation, 0.901, 0.871, 0.926 and 0.898.

REFERENCES

[1] Abusaib, M., Ahmed, M., Nwayyir, H. A., Alidrisi, H. A., Al-Abbood, M., Al-Bayati, A., … Mansour, A. (2020). Iraqi Experts Consensus on the Management of Type 2 Diabetes/Prediabetes in Adults. Clinical Medicine Insights: Endocrinology and Diabetes, 13. doi:10.1177/1179551420942232

[2] Adetunji, A. B., Akande, O. N., Ajala, F. A., Oyewo, O., Akande, Y. F., & Oluwadara, G. (2021). House Price Prediction using Random Forest Machine Learning Technique. Procedia Computer Science, 199, 806–813. doi:10.1016/j.procs.2022.01.100

[3] Alshammari, R., Atiyah, N., Daghistani, T., & Alshammari, A. (2020). Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet. Online Journal of Public Health Informatics, 12(1), 1–12. doi:10.5210/ojphi.v12i1.10611

[4] Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. Information (Switzerland), 11(6). doi:10.3390/info11060332

[5] Bhavsar, H., & Panchal, M. H. (2012). A Review on Support Vector Machine for Data Classification. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1(10), 185–189. doi:ISSN 2278 – 1323

[6] Biostatistics, T. (2007). Topics in Biostatistics. (ed Ambrosius, Walter T., Ed.) (Vol. 404). Totowa, New Jersey 07512.

[7] Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for Healthcare Applications. Journal of Healthcare Engineering, 2021. doi:10.1155/2021/9930985

[8] Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. Information Processing and Management, 59(2). doi:10.1016/j.ipm.2021.102798

[9] Contreras, I., & Vehi, J. (2018). Artificial intelligence for diabetes management and decision support: Literature review. Journal of Medical Internet Research, 20(5), 1–21. doi:10.2196/10775

[10] Ferreira, P., Le, D. C., & Zincir-Heywood, N. (2019). Exploring Feature Normalization and Temporal Information for Machine Learning Based Insider Threat Detection. 15th International Conference on Network and Service Management, CNSM 2019, (Cnsm). doi:10.23919/CNSM46954.2019.9012708

[11] FİDAN, U., UZUNHİSARCIKLI, E., & ÇALIKUŞU, İ. (2019). Classification of Dermatological Data with Self Organizing Maps and Support Vector Machine. Afyon Kocatepe University Journal of Sciences and Engineering, 19(3), 894–901. doi:10.35414/akufemubid.591816

[12] Gorriz, J. M., Segovia, F., Ramirez, J., Ortiz, A., & Suckling, J. (2024). Is K-fold cross validation the best model selection method for Machine Learning?

[13] Hashim, M. S., & Yassin, A. A. (2023). Using Pearson Correlation and Mutual Information (PC-MI) to Select Features for Accurate Breast Cancer Diagnosis Based on a Soft Voting Classifier. Iraqi Journal for Electrical and Electronic Engineering, 19(2), 43–53. doi:10.37917/ijeee.19.2.6

[14] https://www.kaggle.com/uciml/pima-indians-, & Diabetes-database. (n.d.). PIDD Dataset.

[15] Ismail, L., & Materwala, H. (2021). IDMPF: intelligent diabetes mellitus prediction framework using machine learning. Applied Computing and Informatics. doi:10.1108/aci-10-2020-0094

[16] Jabbar, H. G. (2024). Advanced Threat Detection Using Soft and Hard Voting Techniques in Ensemble Learning. Journal of Robotics and Control (JRC), 5(4), 1104–1116. doi:10.18196/jrc.v5i4.22005

[17] Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. International Journal of Environmental Research and Public Health, 18(14). doi:10.3390/ijerph18147346
[18] Khan, F. A., Zeb, K., Al-Rakhami, M., Derhab, A., & Bukhari, S. A. C. (2021). Detection and Prediction of Diabetes Using Data Mining: A Comprehensive Review. IEEE Access, 9, 43711–43735. doi:10.1109/ACCESS.2021.3059343
[19] Knight, J., & Nigam, Y. (2017). Diabetes management 1: disease types, symptoms and diagnosis. Nursing Times, 4(113), 4.
[20] Kousar, S. (2019). Type 1 Diabetes: Causes, Symptoms and Treatments, Review with Personal Experience. Current Research in Diabetes & Obesity Journal, 11(4). doi:10.19080/crdoj.2019.11.555817
[21] Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. Journal of Healthcare Engineering, 2022. doi:10.1155/2022/1684017
[22] Letteri, I., Di Cecco, A., Dyoub, A., & Della Penna, G. (2020). A Novel Resampling Technique for Imbalanced Dataset Optimization. ArXiv Preprint ArXiv, 1–23.
[23] Maurya, S., & Jain, A. (2023). Timely Prediction of Diabetes by Means of Machine Learning Practices. Lecture Notes in Networks and Systems (Vol. 421). doi:10.1007/978-981-19-1142-2_19
[24] Mavrogiorgos, K., Kiourtis, A., Mavrogiorgou, A., Menychtas, A., & Kyriazis, D. (2024). Bias in Machine Learning: A Literature Review. Applied Sciences (Switzerland), 14(19). doi:10.3390/app14198860
[25] Modhugu, V. R., & Ponnusamy, S. (2024). Comparative Analysis of Machine Learning Algorithms for Liver Disease Prediction: SVM, Logistic Regression, and Decision Tree. Asian Journal of Research in Computer Science, 17(6), 188–201. doi:10.9734/ajrcos/2024/v17i6467
[26] Naveen Kishore, G., Rajesh, V., Vamsi Akki Reddy, A., Sumedh, K., & Rajesh Sai Reddy, T. (2020). Prediction of diabetes using machine learning classification algorithms. International Journal of Scientific and Technology Research, 9(1), 1805–1808.
[27] Rani, P., Lamba, R., Sachdeva, R. K., Bathla, P., & Aledaily, A. N. (2020). Diabetes Prediction Using Machine Learning Classification Algorithms. International Conference on Smart Computing and Application, ICSCA 2023, (August 2020). doi:10.1109/ICSCA57840.2023.10087827
[28] Sevilla, J. (2012). Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on Nuclear Science, 44.3(July 1997), 1464–1468. doi:10.1109/23.589532
[29] Shareef, A. Q., & Kurnaz, S. (2023). Deep Learning Based COVID-19 Detection via Hard Voting Ensemble Method. Wireless Personal Communications, (0123456789). doi:10.1007/s11277-023-10485-2
[30] Shi, B., Xu, K., Li, M., & Li, Z. (2024). An Ensemble Learning Approach for Effective Prediction of Diabetes Mellitus Using Hard Voting Classifie, (1), 17. doi:10.1117/12.3023400
[31] Simanto, S., Mridha, K., Saha, R., Limbu, M., Ghosh, A., & Shaw, R. N. (2022). Diabetes Prediction Using Machine Learning Techniques. Lecture Notes in Electrical Engineering, 914(09), 317–333. doi:10.1007/978-981-19-2980-9_26
[32] Sivaranjani, S., Ananya, S., Aravinth, J., & Karthika, R. (2021). Diabetes Prediction using Machine Learning Algorithms with Feature Selection and Dimensionality Reduction. 2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021, 141–146. doi:10.1109/ICACCS51430.2021.9441935
[33] Suyal, M., & Goyal, P. (2022). A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning. International Journal of Engineering Trends and Technology, 70(7), 43–48. doi:10.14445/22315381/IJETT-V70I7P205
[34] Talukder, M. A., Islam, M. M., Uddin, M. A., Kazi, M., Khalid, M., Akhter, A., & Ali Moni, M. (2024). Toward reliable diabetes prediction: Innovations in data engineering and machine learning applications. Digital Health, 10. doi:10.1177/20552076241271867
[35] Tigga, N. P., & Garg, S. (2020). Prediction of Type 2 Diabetes using Machine Learning Classification Methods. Procedia Computer Science, 167(2019), 706–716. doi:10.1016/j.procs.2020.03.336
[36] Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports, 12(1), 1–11. doi:10.1038/s41598-022-10358-x
[37] Yakut, Ö. (2023). Diabetes Prediction Using Colab Notebook Based Machine Learning Methods. International Journal of Computational and Experimental Science and Engineering, 9(1), 36–41. doi:10.22399/ijcesen.1185474