Computers in Biology and Medicine 147 (2022) 105757

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM

Mohammed Gollapalli [a,*], Aisha Alansari [b], Heba Alkhorasani [b], Meelaf Alsubaii [b], Rasha Sakloua [b], Reem Alzahrani [b], Mohammed Al-Hariri [c], Maiadah Alfares [c], Dania AlKhafaji [d], Reem Al Argan [d], Waleed Albaker [d]

[a] Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia
[b] Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia
[c] Department of Physiology, College of Medicine, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, 31441, Saudi Arabia
[d] Department of Internal Medicine, College of Medicine, Imam Abdulrahman Bin Faisal University, King Fahad Hospital of the University, Khobar, Saudi Arabia

ARTICLE INFO

Keywords:
Type 1 diabetes
Type 2 diabetes
Pre-diabetes
Machine learning
Stacking
Permutation feature importance

ABSTRACT

Glucose is the primary source of energy for cells, which are the building blocks of life. It is given to the body by insulin that carries out the metabolic tasks that keep people alive. Glucose level imbalance is a sign of diabetes mellitus (DM), a common type of chronic disease. It leads to long-term complications, such as blindness, kidney failure, and heart disease, having a negative impact on one's quality of life. In Saudi Arabia, a ten-fold increase in diabetic cases has been documented within the last three years. DM is broadly categorized as Type 1 Diabetes (T1DM), Type 2 Diabetes (T2DM), and Pre-diabetes. The diagnosis of the correct type is sometimes ambiguous to medical professionals causing difficulties in managing the illness progression. Intensive efforts have been made to predict T2DM. However, there is a lack of studies focusing on accurately identifying T1DM and Pre-diabetes. Therefore, this study aims to utilize Machine Learning (ML) to distinguish and predict the three types of diabetes based on a Saudi Arabian hospital dataset to control their progression. Four different experiments have been conducted to achieve the highest results, where several algorithms were used, including Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (K-NN), Decision Tree (DT), Bagging, and Stacking. In experiments 2, 3, and 4, the Synthetic Minority Oversampling Technique (SMOTE) was applied to balance the dataset. The empirical results demonstrated promising results of the novel Stacking model that combined Bagging K-NN, Bagging DT, and K-NN, with a K-NN meta-classifier attaining an accuracy, weighted recall, weighted precision, and cohen's kappa score of 94.48%, 94.48%, 94.70%, and 0.9172, respectively. Five principal features were identified to significantly affect the model accuracy using the permutation feature importance, namely Education, AntiDiab, Insulin, Nutrition, and Sex.

1. Introduction

Diabetes Mellitus (DM) primarily causes chronic hyperglycemia due to insufficient insulin levels in the bloodstream [1]. Insulin plays an essential role in glucose level reduction in the blood, carbohydrate metabolism, cell proliferation, physical growth, protein, and fat anabolic regulation [2]. Consequently, serious complications are mainly associated with DM affecting patients' quality of life. The chronic complications driven by DM include heart failure, blindness, kidney failure, and cardiovascular diseases [3]. These illnesses induce a high accumulation in mortality rates, strains on personal life, and difficulties in financial situations [4]. DM is broadly categorized as Type 1 Diabetes (T1DM), Type 2 Diabetes (T2DM), and Pre-diabetes. T1DM accounts for approximately 10% of patients under 30 years, while T2DM represents

* Corresponding author.
E-mail addresses: magollapalli@iau.edu.sa (M. Gollapalli), 2180004329@iau.edu.sa (A. Alansari), 2180007152@iau.edu.sa (H. Alkhorasani), 2180001038@iau.edu.sa (M. Alsubaii), 2180002151@iau.edu.sa (R. Sakloua), 2180005319@iau.edu.sa (R. Alzahrani), mtalhariri@iau.edu.sa (M. Al-Hariri), mnalfares@iau.edu.sa (M. Alfares), dmkhafaji@iau.edu.sa (D. AlKhafaji), Rjalarqan@iau.edu.sa (R. Al Argan), wialbakr@iau.edu.sa (W. Albaker).

https://doi.org/10.1016/j.compbiomed.2022.105757
Received 19 April 2022; Received in revised form 27 May 2022; Accepted 18 June 2022
Available online 20 June 2022

about 90% of diabetic individuals above 30 years old [5]. There are also 318 million adults worldwide suffering from Pre-diabetes having an asymptomatic nature [1]. To distinguish these types, doctors analyze results from agreed-upon tests to prescribe the appropriate treatment methods based on the detected form. However, the accurate diagnosis of the correct type is sometimes ambiguous to medical professionals causing difficulties in managing the illness [6]. The global prevalence of diabetes is steadily boosting, especially in middle-income nations [7]. Therefore, we conducted this study to predict diabetes types using computational intelligence techniques to support doctors in providing the most appropriate treatment strategy.

Diabetes prevention and treatment are challenging due to the lack of adequate policies to assemble supportive environments for healthy behavior and the absence of good health care in multiple settings. By 2030, the Sustainable Development Community (SDC) aims to reduce premature mortality for several disorders, including DM [7]. Therefore, researchers are constantly studying different aspects of DM, particularly in Saudi Arabia. In terms of diabetes rates, Saudi Arabia has the second-highest rate in the Middle East and seventh in the globe, according to the World Health Organization (WHO) [8]. An approximation of a ten-fold diabetic case boost was also recorded within the last three years in Saudi Arabia [8]. Machine Learning (ML), which falls within data mining, has significant potential in supporting health care decision-making and automating numerous mundane tasks. Furthermore, the implementation of ML does not follow any explicit or comprehensive framework, enabling researchers to expand and improve previous work [9]. The literature review demonstrated the focus of researchers on a single type of DM for all the constructed models. It also mostly showed unsatisfactory performance in detecting DM, particularly in Saudi Arabia. Hence, we aim to focus our work on investigating the three mentioned types of DM in the Saudi Arabian region by using a dataset obtained from King Fahad University Hospital.

In this research, four different sets of experiments were conducted as part of the study utilizing Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Decision Tree (DT), Bagging, and Stacking. K-NN has the benefits of simplicity and non-parametric behavior [10], while DT advantages include flexibility, unambiguity, robustness, and others [11]. Generalizing high-dimensional data is a ground-breaking feature of SVM, known for its superiority in classification problems [12]. Furthermore, ensemble techniques that include Bagging and Stacking propose diversity, stability, and outstanding performance illustrated in many recent studies [13]. Empirical results indicated that Stacking trained using the upsampled data obtained the highest accuracy, weighted recall, weighted precision, and cohen's kappa score of 94.48%, 94.48%, 94.70%, and 0.9172, respectively. We combined the K-NN, Bagging DT, and Bagging K-NN models with the K-NN as the meta classifier into a single robust model. Afterward, the permutation feature importance tool was used to extract the most crucial features contributing to the model accuracy [14]. Five principal risk factors were identified from the Stacking model, namely Education, AntiDiab, Insulin, Nutrition, and Sex.

The remaining sections are organized as follows. The second section contains the literature review, while the third section comprises the technical description of the four deployed ML classifiers. Section four represents the empirical study, including dataset description, statistical analysis, experimental setup, performance measures, optimization strategy, proposed Bagging models, and proposed Stacking models. The fifth section demonstrates the experiment results and discussion, whereas the last section contains our conclusions and future work recommendations.

2. Review of related literature

Several researchers utilized the Pima Indian dataset provided by the National Institute of Diabetes, located at John Hopkins University. The dataset consisted of 8 features and 768 instances, where 268 patients were diagnosed with Type 2 Diabetes (T2DM), and the others were non-T2DM patients. D. Joshi and K. Dakhal [15] utilized the Pima Indian dataset to predict T2DM using the R-statistical program. The authors first analyzed and pre-processed the dataset, in which the analysis showed that five features were essential, namely, body mass index (BMI), glucose, age, diabetes pedigree function, and pregnancy. The LR and DT classifiers were then utilized, achieving an accuracy of 78.26% and 74.48%, respectively.

Furthermore, S. Sivaranjani et al. [16] used the same Pima dataset to train two ML algorithms: Random Forest (RF) and SVM. After data pre-processing, the step forward and step backward feature selection techniques were applied, and their effects were compared. Afterward, the Principle Component Analysis (PCA) method was performed for dimensionality reduction. However, it did not impact the performance due to the dataset size. Later, the two models' performance was compared with four features, where RF with step backward feature elimination achieved the highest accuracy of 83%, sensitivity of 83%, and specificity of 82%.

Similarly, Kumari et al. [17] explored the effect of utilizing an ensemble soft voting classifier on the same dataset. The proposed ensemble soft voting classifier combined three binary weak classifiers, including RF, LR, and Naïve Bayes (NB). The proposed classifier's performance was then compared with state-of-the-art techniques and base classifiers, including Adaptive Boosting (AdaBoost), LR, SVM, RF, NB, Bagging, GradientBoost, eXtreme Gradient Boosting (XGBoost), and CatBoost (CAT). The empirical results showed that the proposed ensemble technique outperformed the base classifiers with an accuracy of 79.04%, precision of 73.48%, recall of 71.45%, and f1-score of 80.6%.

More recently, Kumar Kalagotla et al. [18] focused on diabetes detection by employing ensemble classifiers trained using the PIMA clinical dataset. Outliers were analyzed using the interquartile range (IQR) technique, and inconsistent data were imputed by computing the mean values, then transformed using Min-Max scaling. Hence, the dataset evolved into 439 nondiabetic and 200 diabetic patients with nine features selected through the correlation method. Multilayer Perceptron (MLP), SVM, LR, and Adaboost were examined. Still, the Staking technique of (MLP, SVM, and LR) achieved the highest accuracy, precision, recall, and F-score of 78.2%, 72.2%, 54.4%, and 59.4%, respectively.

In another study, Rajendra and Latifi [19] combined the PIMA diabetes dataset and the Vanderbilt dataset to build diabetes prediction models. The main models were built using LR and two ensemble techniques; max-voting and Stacking. Additional classifiers were then added to the ensemble models, including SVM, DT, K-NN, and NB. In both datasets, the ensemble model suppressed the LR model with accuracies of 78% and 93%, respectively.

Comparatively, Khaleel and Al-Bakry [20] employed and evaluated three ML classifiers using the aforementioned Pima dataset, in which the results revealed that LR achieved the highest accuracy of 94%. Vidya J et al. [21] achieved even better results by imputing the missing values and removing redundant data. Additionally, Vidya J et al. utilized 10-fold cross-validation to evaluate six ML algorithms. The results indicated that RF attained the best performance with an accuracy of 97.2% and a Reciever Operator Characteristics Curve (ROC) of 0.963.

In another study, Xiong et al. [22] employed an ensemble-based technique to estimate the T2DM's risk in urban Chinese society. The dataset was collected from patients in Nanjing, including 3845 confirmed cases with T2DM and 8000 nondiabetic patients. Five machine learning algorithms, namely MLP, Adaboost, RF, SVM, and Gradient Tree Boosting (GTB), were trained individually, then combined in an ensemble. The empirical results revealed that the combination of the weak classifiers in an ensemble-based model achieved better results with an accuracy of 91%, sensitivity of 83%, specificity of 95%, precision of 88%, and Area Under the Curve (AUC) of 97%.

Conversely, Semerdijan and Frank [23] utilized the National Health and Nutrition Examination Survey (NHANES) to predict T2DM using an

ensemble model. Five base classifiers, including K-NN, LR, SVM, RF, and Gradient Boosting, were combined, achieving an AUC of 0.834, F-score value of 0.78, recall of 0.82, and precision of 0.78. However, the results indicated that Gradient Boosting outperformed the proposed ensemble model with an AUC of 0.84, F-score value of 0.81, recall of 0.82, and precision of 0.80.

Farooq Ahmad et al. [24] investigated the effect of Health-related attributes on T2DM prediction using ML approaches. The dataset consisted of 3000 patient records with 16 clinical attributes from various Saudi hospitals. Many pre-processing techniques were applied, in which the number of instances decreased to 162. Moreover, the dataset was divided into two sets based on the diabetes-specific tests (Hb-A1c, FPG). Feature permutation and hierarchical clustering were applied for feature selection on both datasets. Afterward, LR, RF, DT, Ensemble Majority Voting (EMV), and SVM algorithms were utilized for modeling. These models were evaluated twice, with nine and eight features, using 10-folds cross-validations. The results of the first dataset indicated that SVM outperformed the others with an accuracy of 82.10% with both nine and eight attributes, whereas the results of the second dataset showed that RF got the highest accuracy of 88.27% with nine features and 87.65% with eight features.

Likewise, Hassan Syed and Khan [25] built an application for the early classification of patients with a high risk of developing Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia. The dataset was obtained from King Abdulaziz University (KAU) cross-sectional diabetes survey containing 990 diabetic instances and 3906 non-diabetic instances. The significant features were extracted using Pearson's Chi-Squared test and binary LR. The pre-processed dataset was split using a ratio of 80:20 for training and testing sets and then balanced using SMOTE. Nine binary classification algorithms were employed, where Decision Forest (DF) outperformed the other models, attaining a respective accuracy, recall, precision, F1 score, and AUC of 0.833 ± 0.018, 0.896 ± 0.0131, 0.800 ± 0.0137, 0.8453 ± 0.0268, and 0.8801 ± 0.016 after the tuning.

On the other hand, Beom Choi et al. [26] developed a screening model to predict Pre-diabetes using ANN and SVM. They collected a dataset from the Korean National Health and Nutrition Examination Survey (KNHANES) containing 9251 samples; the samples from 2010 were used for training and internal validation, whereas samples from 2011 were used for external validation. Moreover, they applied backward LR for feature selection and then optimal hyper-parameters were identified using 10-folds cross-validations. SVM outperformed the other techniques with an accuracy of 66.1%

The reviewed literature indicated that most of the studies focused on predicting T2DM, where Khaleel and Al-Bakry [20] and Vidya J et al. [21] attained the highest prediction scores of 94% and 97.2%, respectively. However, to the best of our knowledge, there is no research conducted on classifying T1DM using ensemble techniques, although it leads to long-term effects. On the other hand, only one study focused on predicting Pre-diabetes, achieving an accuracy of 66.1%. The accurate detection of Pre-diabetes plays a significant role in enhancing the patients' lifestyle to avoid T2DM progression. Therefore, considering T1DM and Pre-diabetes is necessary to improve the patients' quality of life. Apart from that, it is revealed that the studies conducted in Saudi Arabia did not achieve significant results, whereby the study undertaken by Farooq Ahmad et al. [24] achieved the highest accuracy of 88.27%, and the study conducted by Hassan Syed and Khan [25] attained the highest accuracy of 0.833 ± 0.018. Accordingly, this study aims to explore a Saudi Arabian hospital's diabetes dataset and build models that classify diabetes types, including Pre-diabetes, T1DM, and T2DM, to enhance the DM patients' quality of life by assisting doctors in effectively controlling the status of the DM type.

3. Methodologies

3.1. Support Vector Machine (SVM)

Support vector machine (SVM) is a trendy statistical-based supervised machine learning algorithm employed in regression and classification tasks [27]. Cortes and Vapnik introduced it in 1995 to increase the separation of classes and decrease the prediction error. SVM is known to work for linear and non-linear data and very effectively handles the curse of dimensionality issues [28]. In particular, it is efficient with high dimensional feature spaces and small datasets. In the case of linear data, SVM separates training samples into different classes by finding a hyperplane with the maximum margin. Moreover, it determines the largest distance between the nearest points to the margin edge called support vectors and the hyperplane with n-1 dimensions [29]. Equation (1) represents the mathematical formula for maximizing the margin, where w denotes the weight vector, x the input vector, and b the bias [30].

$$\text{minimize} = \frac{1}{2}\left|\left|w\right|\right|^2 \tag{1}$$

$$\textit{subject to } y_i(\langle w \cdot x\rangle + b) > 0$$

SVM utilizes a kernel-based technique to deal with non-linear data, finding the best hyperplane to linearly separate data using some kernel functions and the kernel trick [31]. The list of Kernel functions searched in this study to find the optimal is listed below [30].

Equation (2) represents the Linear Kernel function, where c is a constant number.

$$K(x_i, x_j) = x_i^T x_j \tag{2}$$

Equation (3) represents the Polynomial Kernel function, where d is the degree of Polynomial, γ is the slope, and r is a constant term.

$$K(x_i, x_j) = \left(\gamma x_i^T x_j + r\right)^d, \gamma > 0 \tag{3}$$

Equation (4) represents the RBF Kernel function, where γ is the Gamma and $\exp(-\gamma\left|\left|x_i - x_j^2\right|\right|)$ is the Euclidean distance between two points x_i and x_j.

$$K(x_i, x_j) = \exp\left(-\gamma\left|\left|x_i - x_j^2\right|\right|\right), \gamma > 0 \tag{4}$$

Equation (5) represents the Sigmoid Kernel function, where γ is the slope and r is a constant term.

$$K(x_i, x_j) = tanh\left(\gamma x_i^T x_j + r\right) \tag{5}$$

3.2. K-Nearest Neighbor (K-NN)

The K-Nearest Neighbors (K-NN) is a non-parametric supervised machine learning algorithm developed in the first 1950s, which Thomas Cover subsequently expanded later [32]. K-NN is considered a lazy learner technique as it uses the whole dataset to classify the unlabeled data points by categorizing them to the nearest class depending on the distance measurement. The distance measures searched in this study to find the optimal results are listed below. Equation (6), Equation (7), and Equation (8) represent the formulas for calculating the Euclidean distance, Minkowski distance, and Manhattan distance, respectively, where k denotes the total number of neighbors and p is any real value [33].

Euclidean distance:

$$d(x, y) = \sqrt{\sum_{i=1}^{k}(x_i - y_i)^2} \tag{6}$$

Minkowski distance:

$$d(x,y) = \left(\sum_{i=1}^{k} (|x_i - y_i|)^p \right)^{1/p} \quad (7)$$

Manhattan distance:

$$d(x,y) = \sum_{i=1}^{k} |x_i - y_i| \quad (8)$$

K-NN starts by searching through the entire training dataset for (K) neighbors with the minimum distance between the data points and the target point. Then the majority voting amongst the data points in the neighborhood is used to classify the new data [34].

3.3. Decision Tree

Decision Tree (DT) is a widely used supervised machine learning algorithm in both classification and regression problems. The idea of a DT existed since the late 1950s but gained popularity in 1986 when Quinlan proposed a concept of trees with multiple answers [35]. It is known for its tree-like structure that could be easily interpreted through tree visualization. DT consists of internal nodes and leaf nodes. The internal nodes denote a test over an attribute and contain several branches indicating the test outcome, whereas the leaf nodes denote the resulting class. DT is constructed through a recursive divide-and-conquer in a top-down process where the best attributes are chosen based on a heuristic or statistical manner [36]. In each iteration, the information gain is calculated to determine whether a feature should be added or not. Equation (9) represents the mathematical formula for calculating the information gain, where $-\sum_{i=1}^{c} p_i log_2(p_i)$ calculates the entropy, c denotes the classes present in the dataset, and Pi represents the instance (i) that has the highest probability. Moreover, S represents the sample set, S_v compromises the elements holding v in attribute A, V(A) includes attributes A values, and E is the entropy [37].

$$IG(S,A) = -\sum_{i=1}^{c} p_i log_2(p_i) - \sum_{v \in V(A)}^{C} \frac{S_v}{S} E(S_v) \quad (9)$$

3.4. Bagging

Bagging is an ensemble method from supervised machine learning algorithms utilized for classification and regression tasks. Breiman introduced it in 1996 to improve accuracy and reduce overfitting [38]. Also known as Bootstrap Aggregating, it employs bootstrapping sampling to generate diverse subsets of the training data at random with a replacement approach, which lessens the variance. These subsets would be trained in parallel through multiple weak classifiers that could be of the same or different types. Afterward, the prediction of each classifier is aggregated and then their average or majority is calculated based on the task type [39]. Measuring the average of the models' outcomes in regression cases is known as Soft voting, whereas in classification cases, it is called Hard voting, which focuses on the class with majority votes [40]. In our study, the Hard voting was utilized, as represented in Equation (10), where $argmax_k$ is responsible for the class attaining the majority vote, k is the number of classes, and $\widehat{f}_{bag}(x)$ is the bagged estimate function [41].

$$\widehat{G}_{bag}(x) = argmax_k \widehat{f}_{bag}(x) \quad (10)$$

3.5. Stacking

Stacking is an ensemble framework that uses a meta-model where a new classifier combines several individual base learning predictions to predict the target variable. Wolpert introduced it in 1992 to reduce bias and variance comprehensively, which implies boosting the predictive accuracy [42]. The stacking structure consists of two main layers. The first layer comprises multiple basic learning algorithms, whereas the second layer is the meta-learner, the combiner. First, the base level data is separated using k-fold cross-validation; k-1 is used to train the base models, while the remaining one-fold is used for validation. The basic learning algorithms generate a probability distribution to forecast each class, as shown in Equation (11), where c denotes the class value and $P^M(c_i|x)$ is the likelihood that x belongs to Ref. c_i[43]. Consequently, the meta-model training data is derived using the predictions of the base learning models. Besides, the average of each fold's predictions on test data is used to create the meta-model test data [18,44]. The meta-leaner learns each base model's strengths and flaws and logically blends their predictions to generate the best performance.

$$P^M(x) = \left(P^M(c_1|x),\ P^M(c_2|x),\ \ldots..,\ P^M(c_m|x)\right) \quad (11)$$

4. Empirical studies

4.1. Study data

The Saudi DM types dataset was collected from King Fahad University Hospital (KFUH), Eastern Province, Khobar, Saudi Arabia. It contains 10 unique features from 897 admitted patients between 2018 and 2020, where 731 patients were Pre-diabetic, 89 were diagnosed with T1DM, and 77 with T2DM. The patients' private details, including their national IDs and phone numbers, were omitted to maintain their privacy. Table 1 outlines the description of the database attributes used in this study. The proposal has been reviewed and approved by the local ethical committee at the university (IRB-2022-09-254).

4.2. Statistical analysis

The statistical analysis offers vital tools for visualizing and comprehending a data pattern to improve the data pre-processing and modeling process. Table 2 provides the statistical analysis of the numerical attributes presented in the Saudi Diabetes Types dataset, including the missing values, central tendency measures, standard deviation, minimum, maximum, as well as first and third quartiles. Moreover, Table 3 outlines a brief statistical description of the nominal features.

As illustrated in Table 2, there is a significant difference between the minimum values and the first quartile of the A1c, Tg, and LDL attribute levels which demonstrates the presence of outliers. Furthermore, the considerable discrepancy between the maximum values and the third quartile of the Tg and LDL indicates the presence of an outlier. However, these outliers were not treated since they belong to valid patients. Therefore, both inliers and outliers were considered while training the algorithms. Boxplots were constructed to further visualize the presence of existing outliers, as illustrated in Fig. 1.

Table 1
Features' description.

Feature	Description
Sex	Male or Female.
A1c	The glycated hemoglobin calculates how much sugar is attached to the blood's hemoglobin protein.
Tg	The level of triglycerides in patients' blood.
LDL	The low-density lipoprotein, which is the amount of bad cholesterol.
Albumin	The amount of protein made by the liver.
AntiDiab	Counteracting diabetes: denoting an oral medication that reduces blood sugar.
Insulin	An external injectable insulin form to compensate insulin in the body, which is a hormone made by the islet cells of the pancreas, is taken or not.
Injectable	A medication called liraglutide is taken or not.
Nutrition	Whether the followed diet is healthy or not.
Education	Educated to take care of themselves or not.
Outcome	The target class including three types: Pre-diabetes, T1DM, and T2DM.

Table 2
Numerical attributes statistical description.

Feature	count	mean	std	min	25%	50%	75%	max
A1c	881	8.85	2.22	0	7.2	8.5	10.2	17.1
Tg	897	116.30	87.43	18	61	96	144	844
LDL	897	108.09	38.34	16	81	103	131	324
Albumin	775	3.79	0.47	1.6	3.5	3.8	4.1	4.9

Table 3
Nominal attributes statistical analysis.

Feature	Count	Values
Sex	897	Female (517), Male (380)
Education	897	Yes (277), No (620)
AntiDiab	897	Yes (305), No (592)
Insulin	897	Yes (426), No (471)
Injectable	897	Yes (68), No (829)
Nutrition	897	Yes (212), No (685)
Outcome	897	Pre-diabetes (731), T1DM (89), T2DM (77)

Fig. 2 illustrates the feature correlation heatmap. The map indicates that the most relevant features for classifying DM types (outcomes) were Education and Nutrition. However, it is also revealed that the overall correlation between the variables and the target class is weak.

4.3. Experimental setup

In this experiment, the Jupyter Notebook was utilized to build the DM types classification models using HP Spectre x360, Intel(R) Core (TM) i7-1065G7 CPU, and 16 GB RAM. Before training the models, the dataset was pre-processed by encoding the categorical variables using the Label Encoding method and imputing the missing values using the K-Nearest Neighbor (K-NN) imputer, with a nearest-neighbor value of 3. Moreover, the features were scaled using the MinMax scaler provided by the sklearn library. After cleaning the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was applied to equalize the number of instances that belong to each class using the imblearn library, in which each class compromised of 731 records. Subsequently, stratified 10-folds cross-validation was utilized in the process of training and evaluating three ML algorithms, namely, Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Decision Tree (DT), using the original and upsampled data. To optimize the models' performances, GridSearchCV with stratified 10-folds cross-validation was used. Afterward, the performance of the proposed algorithms trained using the original and upsampled data was compared. It was concluded that the performance of the models trained on the upsampled data was better. Accordingly, three Bagging ensemble models combined with each weak classifier individually with their optimal hyper-parameter were developed using the upsampled data. The Bagging ensemble was also optimized using the GridSearchCV. A novel Stacking ensemble model was then established by combining the best weak classifier (K-NN) and the best two Bagging ensembles (Bagging KNN and Bagging DT). The K-NN algorithm was used as a meta-classifier and was optimized using the GridSearchCV technique. Finally, the "feature_importance_permutation" function from the mlxtend library was used to produce this measurement with a "num_rounds" of 10 and a Seed of 1. A constant number of 42 was set to the random state variable for all the processes. Fig. 3 illustrates a summary of the proposed framework.

4.4. Performance measure

Four performance measures were utilized to evaluate the models' performance proposed in this paper, including precision, recall, accuracy, and cohen's kappa score. The recall and precision weighted averages were used due to the multiclassification task. The weighted average multiplies the prediction values by the contributed samples for each class and then combines them over the entire number of samples in the data. The weighted recall is the sum of each class's recall value multiplied by the class's weight over the total number of samples. Similarly, the weighted precision adds the multiplication of each class's precision value with a specific number of instances and then divides them by the total samples. Cohen's kappa score measures the agreement between the predicted and actual labels, where p_0 represents the accuracy of the models and p_e denotes the agreement between the predicted and actual labels. Equation (12), Equation (13), Equation (14), and Equation (15) represent the formulas we employed for calculating the precision, recall, and accuracy, respectively.

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

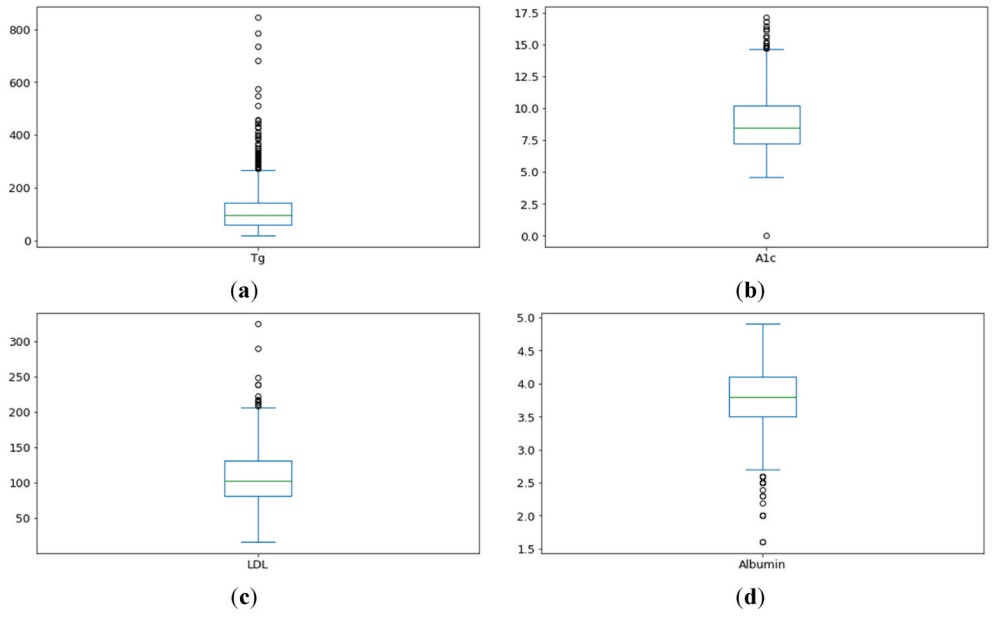


Fig. 1. (a) Tg Boxplot, (b) A1c Boxplot, (c) LDL Boxplot, (d) Albumin Boxplot.

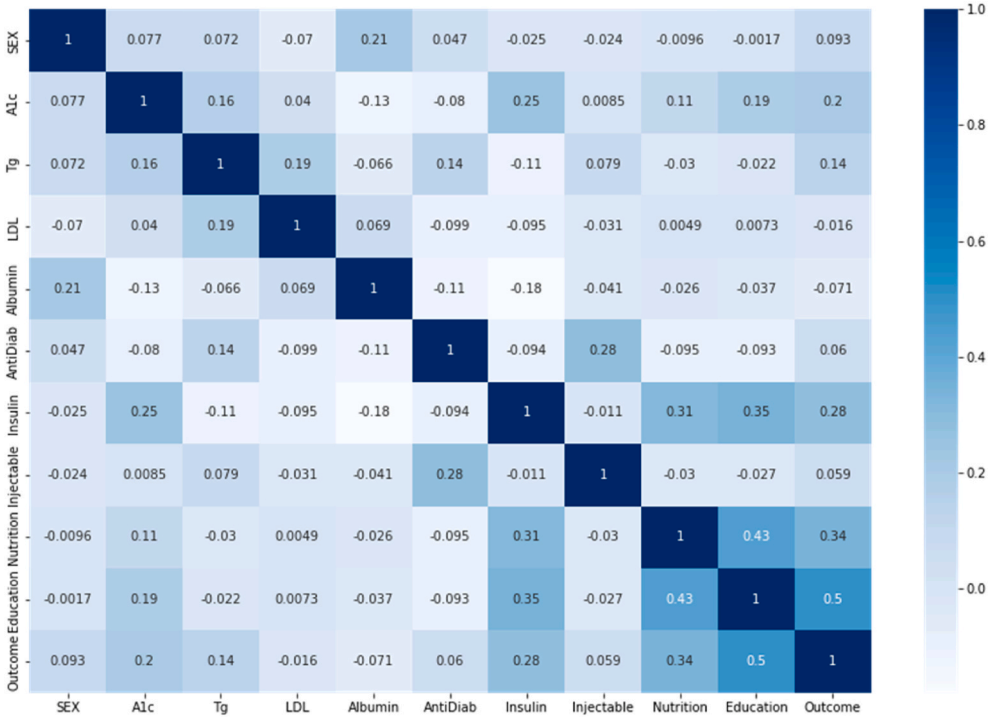


Fig. 2. Feature correlation heatmap.

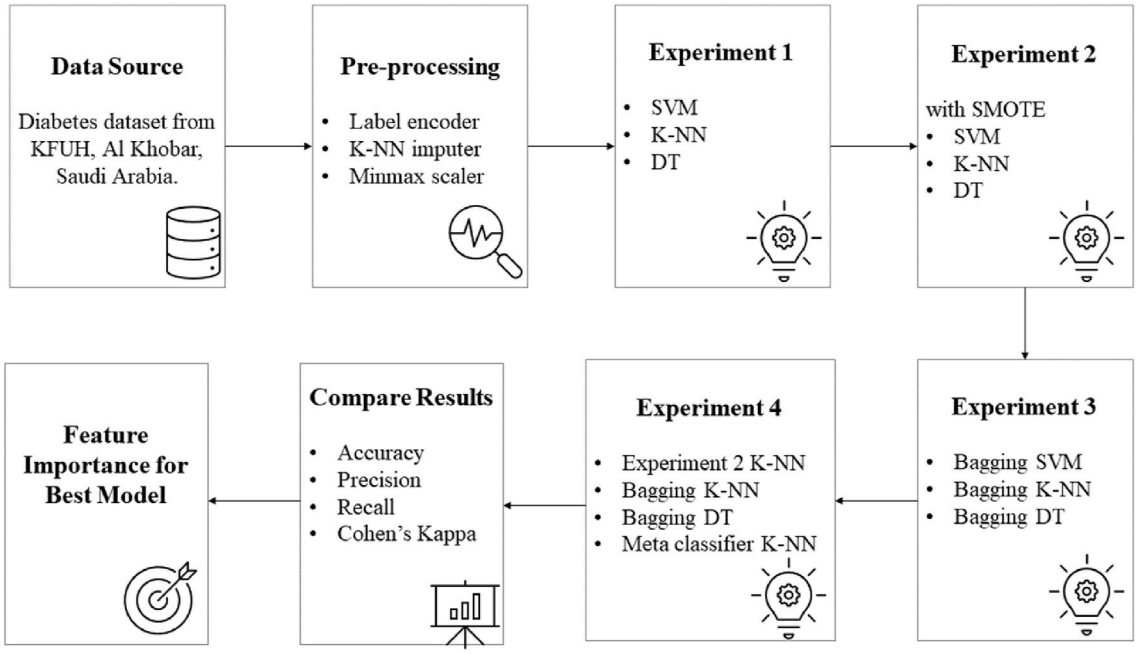


Fig. 3. Study's framework.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$Cohen's\ Kappa = \frac{p_0 - p_e}{1 - p_e} \tag{15}$$

4.5. Optimization strategy

The GridSearchCV hyper-parameter tuning method was utilized to find the optimal hyperparameters producing the highest accuracy, which specifies a grid search of selected hyper-parameters with specific values. The technique tries every possible combination based on the grid search and returns the best combination that achieves the best results using stratified 10-folds cross-validation. This strategy contributes to enhancing the model performance resulting in better outcomes. Below is the hyper-parameter grid identified for each of the experimented algorithms.

- SVM: Cost {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20}, Gamma {1, 0.1, 0.01, 0.001, 0.0001, scale, auto}, and Kernel {RBF, Sigmoid, Linear, Poly}.

- K-NN: Metric {Minkowski, Euclidean, Manhattan} and N_neighbors {1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39}.
- DT: Max_depth {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 18, 20, 23, 25, 27, 30, 31, none}, Criterion {Gini, Entropy}, and Splitter {Best, Random}.

4.5.1. Before SMOTE

Fig. 4 shows the performance of each of the Kernels with different Cost and Gamma values. As can be seen, the auto and 0.1 Gamma values attained the best score using the Polynomial, RBF, and Sigmoid kernels. Additionally, the Gamma value of 1 reached a significant score with the RBF kernel. Conversely, It was revealed that the change in Gamma did not affect the performance while using the Linear Kernel, since all of the Gamma values yielded the same score. Overall, the Polynomial Kernel achieved the highest score of 85.84% when using a Gamma value of 0.1 and a Cost of 8.

Fig. 5 illustrates the performance of K-NN when using different Metrics and N_neighbors. It was indicated from the figure that the Minkowski and Euclidean distances achieved similar scores, whereas the Manhattan metric showed another trend. Overall, the highest accuracy of 85.95% was attained by the Euclidean and Minkowski Metrics when the N_neighbors hyper-parameter was set to 11.

Fig. 6 shows the impact of utilizing different Splitter and Criterion types in specific Max_depth values. Overall, Best and Random Splitters' performance increased when the Max_depth values ranged between 1 and 4. The performance significantly dropped using the Best Splitter and showed an unstable performance using the Random Splitter. Generally, the Entropy Criterion achieved the highest performance score of 85.95% using Random Splitter and a Max_depth of 4.

4.5.2. After SMOTE

Fig. 7 illustrates the performance of SVM using four Kernel functions with different Cost and Gamma values. It is indicated that the Gamma value 1 performed the best when used with the Polynomial and RBF Kernels. However, it attained the lowest performance when utilized with the Sigmoid Kernel. Apart from that, it is revealed that the change in Gamma value did not affect the performance of SVM when using the Linear kernel. Overall, the RBF kernel with Gamma value 1 and Cost value 20 produced the highest accuracy of 90.83%.

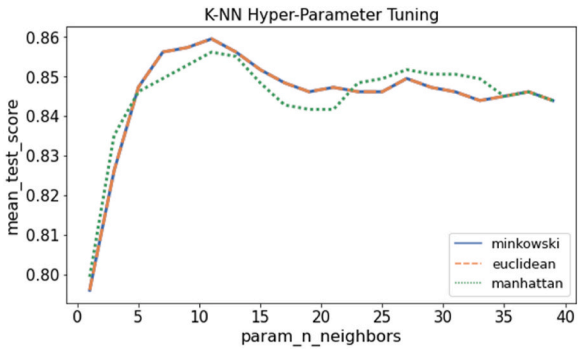


Fig. 5. KNN metric parameter.

Fig. 8 illustrates the performance of Minkowski, Euclidean, and Manhattan Metrics with different numbers of N_neighbors. It is revealed that both Euclidean and Minkowski had similar performances, whereas the Manhattan Metric performed differently. Overall, the increment in N_neighbors affected K-NN's performance negatively. It is concluded that K-NN attained the highest performance value of 93.11% using the Manhattan Metric and N_neighbors of 1.

Fig. 9 displays the performance of DT when using a different Splitter, Criterion, and Max_depth. It is indicated that the increase in Max_depth showed an overall increasing pattern. DT achieved the best performance with an accuracy of 90.56% while using the Gini Criterion, with Max_depth of 15, and with Random Splitter.

A summarization of the optimal hyperparameters of each experimented model is shown in Table 4.

4.6. Proposed bagging models

Recently, ensemble techniques have gained widespread popularity among researchers for their ability to enhance the performance of traditional algorithms. In the third experiment, we built a Bagging model for each of the experimented standard algorithms, including SVM, K-NN, and DT, with their optimal hyper-parameters obtained from

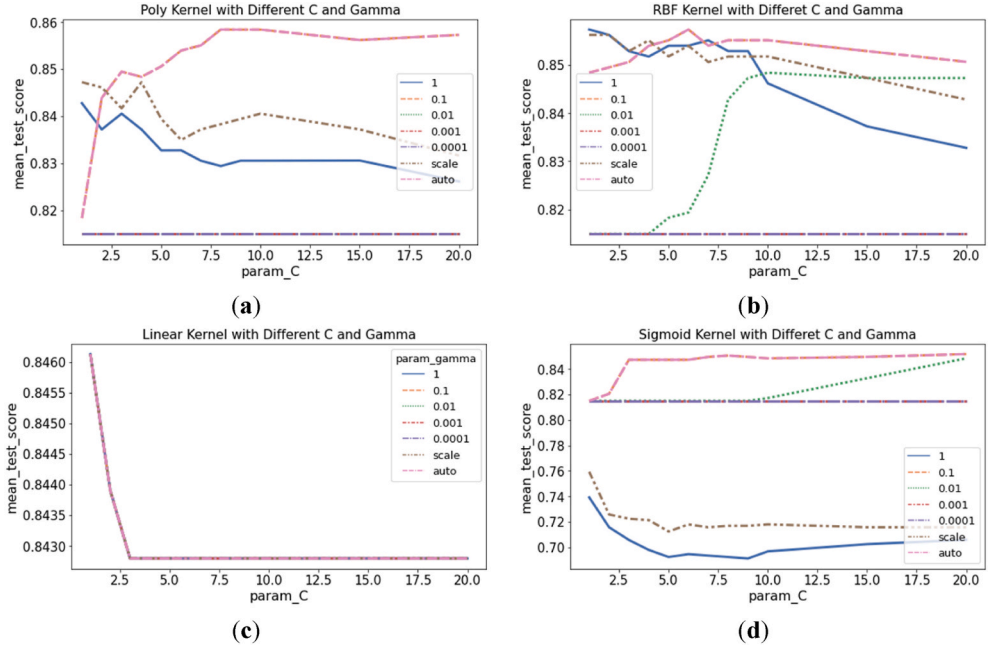


Fig. 4. (a) Poly kernel; (b) RBF kernel (c) Linear kernel; (d) Sigmoid kernel.

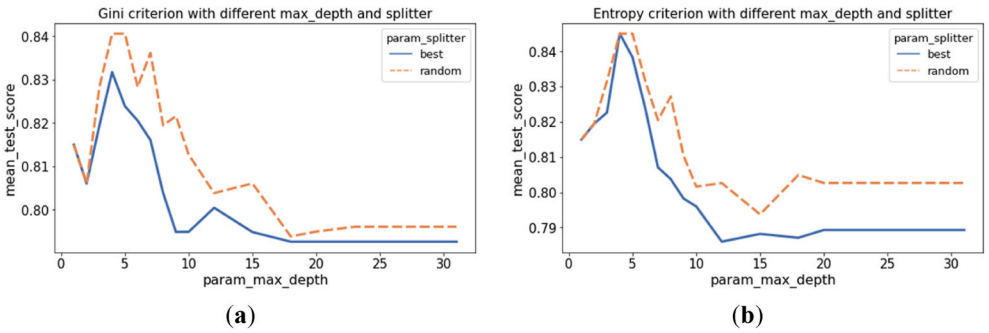


Fig. 6. (a) Gini criterion (b) Entropy criterion for DT model.

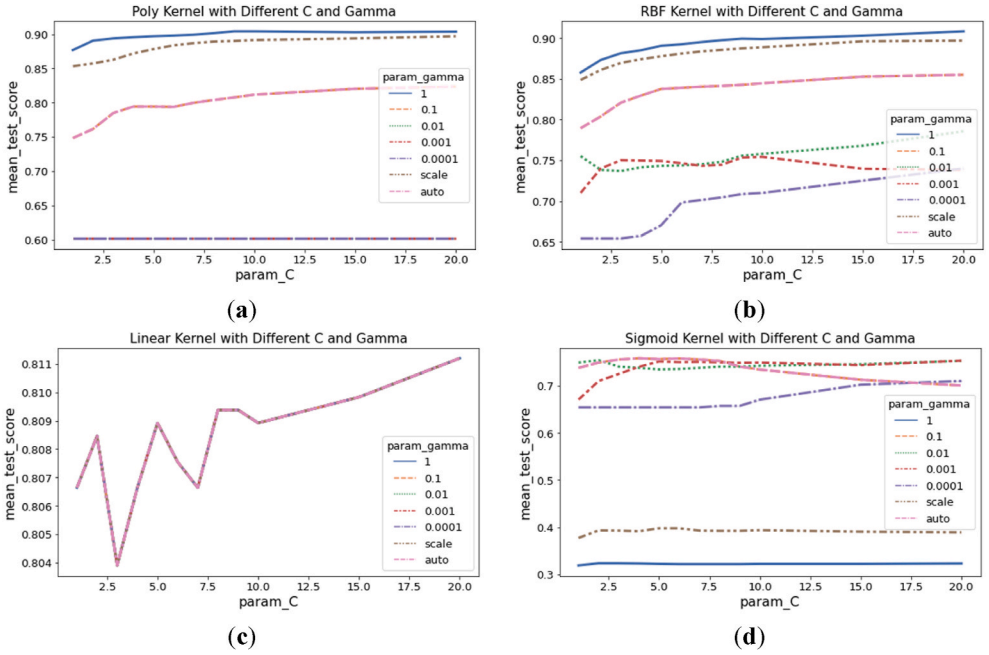


Fig. 7. (a) Poly kernel; (b) RBF kernel (c) Linear kernel; (d) Sigmoid kernel.

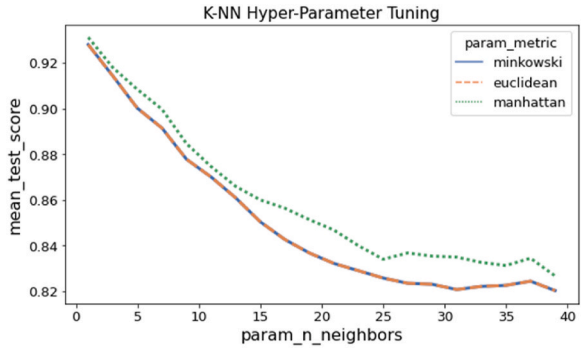


Fig. 8. KNN metric parameter.

the GridSearchCV technique employed using the upsampled data. Later, GridSearchCV was performed to optimize the Bagging models. Below is the hyper-parameter grid for the Bagging algorithms.

- n_estimators {10, 20, 40, 60, 80, 100, 150, 200, 250, 300, 350, 400, 450, 500}
- max_features {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}.
- max_samples {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1}.

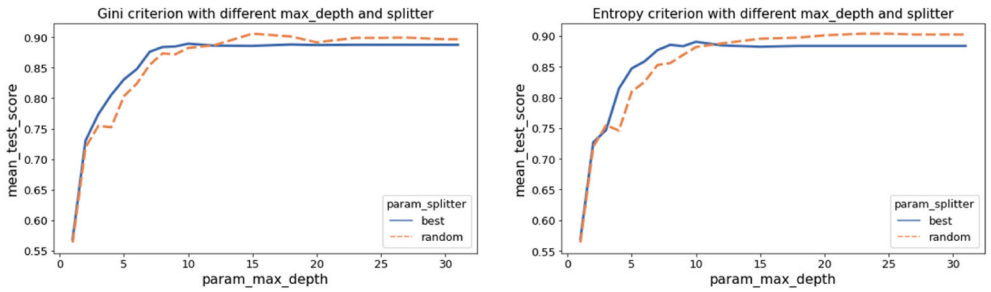
Table 5 outlines the Bagging models with their optimal hyper-parameters.

4.7. Proposed stacking model

Stacking differs from Bagging in that it integrates heterogeneous algorithms into a single model using a meta-classifier, fusing their prediction skills into a single model. In the fourth experiment, we built a Stacking model by combining the most effective three models obtained from the previous experiments with their optimal hyper-parameters. The models include the experiment's 2 K-NN model, bagging K-NN, and bagging DT. The meta-classifier selected was the best-formed traditional model, which is the experiment 2 K-NN model. GridSearchCV was utilized to optimize the meta-classifier in order to attain the highest possible accuracy. Below is the hyper-parameter grid for the Stacking algorithm.

- Metric {minkowski, euclidean, manhattan}.
- N_neighbors {1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 23, 25, 27, 29, 31, 33, 35, 37, 39}.

Table 6 outlines the Stacking model with its optimal hyper-parameter.



(a) (b)

Fig. 9. (a) Gini criterion (b) Entropy criterion for DT model.

Table 4
The optimal hyper-parameters for each classifier.

Experiment	Classifier	Hyper-parameter	Values	Accuracy
Experiment 1 (Performance without sampling)	SVM	Cost	8	85.84%
		Gamma	0.1	
		Kernel	Poly	
	K-NN	N_neighbors	11	85.95%
		Metric	Minkowski	
	DT	Splitter	Best	84.50%
		Criterion	Entropy	
		Max_depth	4	
Experiment 2 (Performance with upsampling)	SVM	Cost	20	90.83%
		Gamma	1	
		Kernel	RBF	
	K-NN	N_neighbors	1	93.11%
		Metric	Manhattan	
	DT	Splitter	Random	90.56%
		Criterion	Gini	
		Max_depth	15	

Table 5
Bagging models optimal hyper-parameters.

Bagging Model	Hyper-parameters	Values	Accuracy
SVM Bagging	n_estimators	10	90.70%
	max_features	0.9	
	max_samples	0.9	
K-NN Bagging	n_estimators	350	94.34%
	max_features	0.7	
	max_samples	0.8	
DT Bagging	n_estimators	100	94.12%
	max_features	0.8	
	max_samples	0.8	

5. Empirical results

Table 7 shows a comparison of the suggested classifier's performance in terms of the previously described performance measures for experiments 1 and 2 following hyper-parameter tuning. As can be seen from the table, in experiment 1, K-NN achieved the highest accuracy rate of 85.95%, followed by SVM, almost attaining the same accuracy with a slight difference of only 0.2%. On the other hand, DT obtained the lowest accuracy rate of 84.50%. It is also indicated that the highest weighted precision rate of 84.95% was achieved by K-NN, followed by SVM with a difference of only 0.17%, whereas the lowest weighted precision rate of 81.17% was attained by DT. Similarly, the highest weighted recall rate of 85.95% was achieved by K-NN, followed by SVM with a slight difference of 0.11%, while DT obtained the lowest weighted recall rate of 84.50%.

Table 6
Stacking optimal hyper-parameters.

Stacking Model	Hyper-parameters	Values	Accuracy
Stacking (bagging K-NN, bagging DT, K-NN)	Metric	Manhattan	94.48%
	N_neighbors	1	

Table 7
Classification performance of the final models.

Experiment	Classifier	Accuracy	Weighted precision	Weighted recall	Cohen's kappa
Experiment 1 (Performance without sampling)	SVM	85.84%	84.78%	85.84%	**0**.5092
	K-NN	85.95%	84.95%	85.95%	0.4769
	DT	84.50%	81.17%	84.50%	0.4519
Experiment 2 (Performance with upsampling)	SVM	90.83%	91.37%	90.83%	0.8625
	K-NN	93.11%	93.38%	93.11%	0.8967
	DT	90.56%	90.77%	90.56%	0.8584
Experiment 3 (Bagging ensemble with upsampling)	Bagging SVM	90.70%	91.19%	90.70%	0.8605
	Bagging K-NN	94.34%	94.59%	94.34%	0.9152
	Bagging DT	94.12%	94.42%	94.12%	0.9118
Experiment 4 (Stacking with upsampling)	Stacking	94.48%	94.70%	94.48%	0.9172

In experiment 2, it is revealed that K-NN achieved the highest accuracy rate of 93.11%. On the other hand, DT obtained the lowest accuracy rate of 90.56%. It is also indicated that the highest weighted precision rate of 93.38% was achieved by K-NN, whereas the lowest weighted precision rate of 90.77% was attained by DT. Similarly, the highest weighted recall rate of 93.11% was achieved by K-NN, while DT obtained the lowest weighted recall rate of 90.56%.

In general, it is denoted that SMOTE considerably improved the models' performance by increasing the accuracy of SVM, K-NN, and DT by 4.99%, 7.16%, and 6.06%, respectively. Moreover, in terms of the cohen's kappa score, it is indicated that the agreement improved remarkably after performing SMOTE with a difference of 0.3533, 0.4198, and 0.4065 for SVM, K-NN, and DT, respectively. The reason behind the considerable improvement is the ability of SMOTE to generate new instances synthetically that differ moderately from the original records in order to balance the classes through upsampling the minority class. Accordingly, it enhances the generalization capabilities of the models, leading to an improvement in the overall performance.

Furthermore, it was indicated that the boosting ensemble improved the performance of both K-NN moderately and DT significantly in terms of all the performance measures. However, it slightly degraded the performance of SVM. Bagging K-NN achieved the highest accuracy rate of 94.34%, followed by bagging DT, almost attaining the same accuracy with a slight difference of only 0.22%. On the other hand, bagging SVM obtained the lowest accuracy rate of 90.70%. It is also indicated that the

highest weighted precision rate of 94.59% was achieved by bagging K-NN, followed by bagging DT with a difference of only 0.17%, whereas the lowest weighted precision rate of 91.19% was attained by bagging SVM. Similarly, the highest recall rate of 94.34% was achieved by bagging K-NN, followed by bagging DT with a slight difference of 0.22%, while bagging SVM obtained the lowest weighted recall rate of 90.70%. Moreover, the highest cohen's kappa score of 0.9152 was achieved by bagging K-NN, followed by bagging DT with an inconsiderable difference of 0.0034, whereas bagging SVM attained the lowest cohen's kappa score of 0.8605. Overall, It is concluded that the proposed stacking ensemble in experiment 4 achieved the highest performance with an accuracy of 94.48%, weighted precision of 94.70%, weighted recall of 94.48%, and cohen's kappa score of 0.9172.

The failure to accurately recognize the disease leads to severe consequences that affect the patients' survival while significantly costing the hospital. Misdiagnosis worsens the patient's condition leading to the progression of the disease to the later stages while reducing the patient's chances of complete recovery. Moreover, misdiagnosed patients may take legal action against the hospital resulting in costly compensation. Not to mention other serious consequences such as psychological outcomes and negatively impacting the public confidence in the diagnosis [45]. In this study, the misdiagnosis of T1DM results in a life-threatening complication called Diabetic Ketoacidosis (DKA), which is the leading cause of death in children with T1DM, as well as severe long-term consequences [46]. Misdiagnosis of T2DM, on the other hand, results in premature death due to significant consequences such as kidney failure, cerebrovascular illness, and cardiovascular disease. Moreover, permanent disability, blindness, and foot ulcers could also be listed as other diabetes complications [47]. Hence, obtaining the highest number of accurate predictions is essential when selecting the best-performing classifier. Fig. 10 illustrates the confusion matrix of the proposed stacking model. It is revealed that the model succeeded in reliably classifying 96.9% of the T2DM cases with 23 false-negative and 96.6% of the T1DM cases with 25 false-negatives. Additionally, the model accurately classified the Pre-diabetic patients with 85 false-negatives. Accordingly, it was concluded that the proposed model attained promising results when classifying the most critical DM types, T2DM and T1DM, whereas it achieved an acceptable result for classifying Pre-diabetes with a percentage of 88.3%.

Permutation feature importance is an effective method for explaining black box models that detects and ranks features based on their predictive power during or after training. Each predictor's score is assigned based on how well it could improve the predictions, allowing attribute interpretation based on relative predictive power. Fig. 11 shows the values obtained from the stacking model's permutation feature significance. Education, AntiDiab, and Insulin were the features that had the greatest impact on model accuracy, as seen in the graph. Table 8 shows a more detailed extracted knowledge from the proposed technique. It is concluded that patients with Pre-diabetes are more likely to be males, uneducated, unhealthy, and taking Insulin and AntiDiab

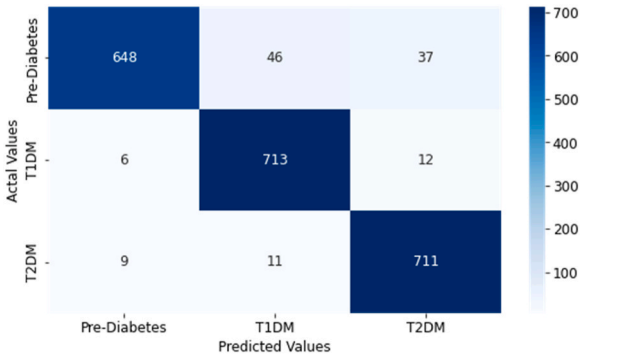


Fig. 10. Stacking confusion matrix.

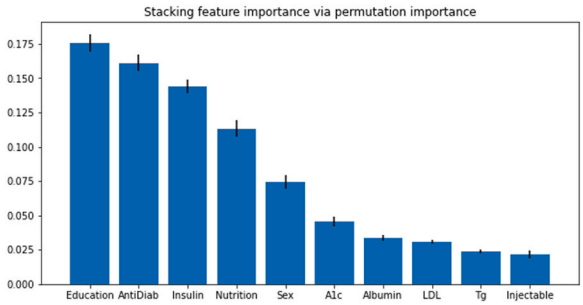


Fig. 11. Stacking permutation feature importance.

medications. In contrast, T1DM and T2DM cases are accompanied by Education and non-Insulin consumers.

6. Discussion

Recently, the overuse of trending technologies has transformed humans' lifestyles and imposed unhealthy habits. Practices that push people towards a sedentary lifestyle and detract from physical activity cause many chronic non-communicable disorders, leading to life-threatening consequences. Diabetes mellitus (DM) is a non-communicable disease with one of the world's highest yearly fatality rates [48]. Therefore, employing the power of ML technology and identifying the essential traits that cause DM would immensely assist in reducing the prevalence of the disease. This could also help control the progression of DM types and prevent the anticipated complications associated with incorrect practices patients might adopt. Some studies, such as Habibi et al. [49], claimed that diabetes could be predicted using the evolving ML techniques without acquiring any laboratory tests. However, results from their study indicated the model's weakness due to omitting laboratory test variables and including only basic features. The model's validity is also controversial when investigating different regions, populations, and lifestyles. This fact encourages researchers to expand the experimentation of previous publications for perceiving new patterns in diverse nations and exploring markers of various types using laboratory tests.

In this study, we performed four distinct experiments to predict the

Table 8
Extracted Knowledge from the stacking model.

Attribute	Attribute Values	Pre-diabetes (731)	Type 1 (89)	Type 2 (77)
Education	0	594	10	16
	1	137	79	61
AntiDiab	0	478	83	31
	1	253	6	46
Insulin	0	444	2	25
	1	287	87	52
Nutrition	0	622	19	44
	1	109	70	33
Sex	0	435	49	33
	1	296	40	44
A1c	(0–7.7]	276	15	19
	(7.7–9.6]	248	18	25
	(9.6–17.1]	191	56	33
Albumin	(1.6–3.7]	260	32	41
	(3.7–4.0]	175	13	20
	(4.0–4.9]	202	20	12
LDL	(16.0–89.0]	258	18	29
	(89.0–120.0]	227	46	22
	(120.0–324.0]	246	25	26
Tg	(18.0–73.0]	248	44	12
	(73.0–123.0]	243	25	31
	(123.0–844.0]	240	20	34
Injectable	0	677	88	64
	1	54	1	13

most common types of DM using a Saudi Arabian hospital dataset. The narrow prerequisite for our model comprised three basic features (Sex, Nutrition, Education), four laboratory tests (A1c, Tg, LDL, Albumin), and three external medications (AntiDiab, Insulin, Injectable). To build the predictive models, we utilized SVM, K-NN, and DT classifiers, along with Bagging and Stacking ensemble-based approaches. Before assembling the models, nominal variables were encoded using the Label encoder, missing values were imputed using K-NN imputer, and features were scaled using the MinMax scaler. Experiments 1 and 2 presented the robustness of K-NN over SVM and DT when dealing with imbalanced and upsampled datasets. As the outcomes obtained from using upsampled dataset were dramatically higher, the ensemble models were developed using the SMOTE dataset. Similarly, Bagging K-NN in experiment 3 outperformed the other Bagging models. The performance of Bagging DT achieved asymptotic results to Bagging K-NN with a difference of 0.22%. Subsequently, the Stacking model incorporated the prior optimal three models, K-NN, Bagging K-NN, and Bagging DT, as conducted in experiment 4.

Despite the challenges encountered in distinguishing a single DM type in the reviewed studies, our novel ensemble model outperformed almost all the selected benchmarks and accurately identified three DM forms. The highest performance was accomplished by the novel Stacking model, with an accuracy of 94.48%, precision of 94.70%, recall of 94.48%, and cohen's kappa score of 0.9172. If the model is used in a platform that serves medical doctors, it is critical to keep false-negative and false-positive misdiagnosis rates to a minimum. The reduction of false-negatives would save patient's lives, whereas the management of false-positive rates would improve the service's integrity and reliability. Predicting DM types using our approach helps slow the course of DM types by identifying the best treatment plan for each patient, preventing complications that could lead to death.

Moreover, our proposed model could be utilized to extract the most relevant risk factor associated with diabetes. Permutation feature importance is a recent technique proposed for assessing the impact of each feature on the model performance through randomly shuffling each attribute [24]. Acknowledging the importance of the dataset variables for the model may broaden the detected features in diabetes diagnosis. Hence, identifying the contribution of all the attributes through analyzing their importance is also significant in this study. According to the performed feature importance permutation, Education, AntiDiab, and Insulin were the most influential features affecting the model's performance. As mentioned in the aforementioned sections, DM is associated with unhealthy habits and lifestyles. Hypothetically, diabetes-educated individuals are more likely to retain appropriate routines due to their awareness of the disease and healthy diets. Several studies, such as the one conducted by Ref. [50], outlined the effect of education on maintaining healthy behavior, adequate self-care, and reducing DM consequences. On the other hand, the least influential markers were Injectable, Tg, LDL, Albumin, and A1c. Nutrition and Sex closely contributed to the model's performance in predicting DM types. Due to the unsatisfactory performance presented by the previous studies that primarily utilized traditional risk factors, the research conducted in this study using the proposed blended (basic, laboratory, and medication) features could assist doctors in diagnosing high-risk individuals with any of the investigated DM types.

The proposed ensemble approach has clearly shown the potential value in identifying the three types of diabetes based on simple, clinical, and demographical traits that physicians might not routinely incorporate into diabetes diagnosis. According to the findings presented in Table 8, patients with T1DM and T2DM are more likely to be educated. As a result, we anticipate that they will experience a variety of symptoms that will compel them to learn the proper methods for managing their conditions. T1DM patients were biased towards nutrient and non-AntiDiab takers. Other patterns can also be observed from the least important attributes.

7. Conclusion and recommendation

Diabetes has been steadily increasing due to sedentary lifestyles and unhealthy diets. Identifying diabetic patients before they develop significant symptoms is for preventing and managing the impact of the various types of this disease. Consequently, this study aimed to distinguish and predict three types of diabetes, namely T1DM, T2DM, and Prediabetes, using computational intelligence techniques. A total of four experiments were conducted in this study utilizing Support Vector Machines (SVMs), K-Nearest Neighbor (K-NN), Decision Trees (DT), Bagging, and Stacking. The dataset was officially obtained from the King Fahad University Hospital (KFUH), including 10 unique attributes and 897 instances. Experiments 1 and 2 used the three basic classifiers, while experiments 3 and 4 employed the two ensemble approaches. As the accuracy of the model significantly improved in experiment 2 due to applying SMOTE to the dataset, the consequent models were utilized using the upsampled dataset. Empirical results demonstrated the prevalence of the stacking model over the others with accuracy, recall, and precision of 94.48%, 94.48%, and 94.70%, respectively. After performing the permutation feature importance analysis, it appeared that Education, AntiDiab, Insulin, Nutrition, and Sex are the most important features affecting the model's ability to predict significantly. Accurate diabetes type prediction would aid in the development of interventional programs and behavioral changes that could prevent or delay the progression of DM, which causes severe complications. Currently, the model could be used in an application or website to help Saudi doctors diagnose diabetes. It is concluded that the findings of this study would also improve the social, economic, and medical conditions of diabetic patients. Unusual detected features may also aid the healthcare sector in determining an individual's medical status.

As part of our recommendations for future work, we look forward to incorporating more classifiers into the suggested innovative stacking model, collecting more clinical data for T1DM and T2DM, and examining the value of other features. Furthermore, future attempts could include researching the most popular types of nutrition by region in order to improve the model's generalization capacity based on patient's lifestyle. In addition, future studies could focus on other types of diabetes, such as gestational diabetes, which develops during pregnancy and puts babies at risk for disorders like hypoglycemia and obesity.

Declaration of competing interest

The authors declare that there is no conflict of interest in this entire research.

References

[1] H. Sone, Diabetes mellitus, in: R.S. Vasan, D.B. Sawyer (Eds.), Encyclopedia of Cardiovascular Research and Medicine, Elsevier, Oxford, 2018, pp. 9–16, https://doi.org/10.1016/B978-0-12-809657-4.99593-0.
[2] T. Andoh, Subchapter 19A - insulin, 157-e19A-3, in: Y. Takei, H. Ando, K. Tsutsui (Eds.), Handbook of Hormones, Academic Press, San Diego, 2016, https://doi.org/10.1016/B978-0-12-801028-0.00148-3.
[3] J. Hippisley-Cox, C. Coupland, Diabetes treatments and risk of amputation, blindness, severe kidney failure, hyperglycaemia, and hypoglycaemia: open cohort study in primary care, Mar, BMJ 352 (2016), i1450, https://doi.org/10.1136/bmj.i1450.
[4] A.N. Baanders, J.W.M. Heijmans, The impact of chronic diseases: the partner's perspective, Fam. Community Health 30 (4) (2007) 305–317.
[5] C.V.A. Collares, et al., Transcriptome meta-analysis of peripheral lymphomononuclear cells indicates that gestational diabetes is closer to type 1 diabetes than to type 2 diabetes mellitus, Sep, Mol. Biol. Rep. 40 (9) (2013) 5351–5358, https://doi.org/10.1007/s11033-013-2635-y.
[6] A.E. Butler, D. Misselbrook, Distinguishing between type 1 and type 2 diabetes, Aug, BMJ 370 (2020), m2998, https://doi.org/10.1136/bmj.m2998.
[7] World Health Organization, Global Report on Diabetes, World Health Organization, Geneva, 2016. Accessed: Apr. 11, 2022. [Online]. Available: https://apps.who.int/iris/handle/10665/204871.
[8] M.A. Al Dawish, et al., Diabetes mellitus in Saudi Arabia: a review of the recent literature, Curr. Diabetes Rev. 12 (4) (2016) 359–368, https://doi.org/10.2174/1573399811666150724095130.

[9] A.A. Verma, et al., Implementing machine learning in medicine, Aug, CMAJ (Can. Med. Assoc. J.) 193 (34) (2021) E1351–E1357, https://doi.org/10.1503/cmaj.202434.

[10] P. Nadkarni, in: P. Nadkarni (Ed.), Chapter 10 - Core Technologies: Data Mining and 'Big Data,'" in *Clinical Research Computing*, Academic Press, 2016, pp. 187–204, https://doi.org/10.1016/B978-0-12-803130-8.00010-5.

[11] Y. Song, Y. Lu, Decision tree methods: applications for classification and prediction, Apr, Shanghai Arch Psychiatry 27 (2) (2015) 130–135, https://doi.org/10.11919/j.issn.1002-0829.215044.

[12] Y.-P.P. Chen, E.P. Ivanova, F. Wang, P. Carloni, 9.15 - bioinformatics, in: H.-W. (Ben) Liu, L. Mander (Eds.), Comprehensive Natural Products II, Elsevier, Oxford, 2010, pp. 569–593, https://doi.org/10.1016/B978-008045382-8.00729-2.

[13] Y. Yang, Chapter 4 - ensemble learning, in: Y. Yang (Ed.), Temporal Data Mining via Unsupervised Ensemble Learning, Elsevier, 2017, pp. 35–56, https://doi.org/10.1016/B978-0-12-811654-8.00004-X.

[14] 4.2. Permutation feature importance — scikit-learn 1.0.2 documentation." https://scikit-learn.org/stable/modules/permutation_importance.html (accessed Apr. 18, 2022).

[15] R.D. Joshi, C.K. Dhakal, Predicting type 2 diabetes using logistic regression and machine learning approaches, Int. J. Environ. Res. Publ. Health 18 (14) (Jan. 2021), https://doi.org/10.3390/ijerph18147346. Art. no. 14.

[16] S. Sivaranjani, S. Ananya, J. Aravinth, R. Karthika, Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction, Mar, in: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS), vol. 1, 2021, pp. 141–146, https://doi.org/10.1109/ICACCS51430.2021.9441935.

[17] S. Kumari, D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, International Journal of Cognitive Computing in Engineering 2 (Jun. 2021) 40–46, https://doi.org/10.1016/j.ijcce.2021.01.001.

[18] S.K. Kalagotla, S.V. Gangashetty, K. Giridhar, A novel stacking technique for prediction of diabetes, Aug, Comput. Biol. Med. 135 (2021), 104554, https://doi.org/10.1016/j.compbiomed.2021.104554.

[19] P. Rajendra, S. Latifi, Prediction of diabetes using logistic regression and ensemble techniques, Jan, Computer Methods and Programs in Biomedicine Update 1 (2021), 100032, https://doi.org/10.1016/j.cmpbup.2021.100032.

[20] F. Alaa Khaleel, A.M. Al-Bakry, Diagnosis of diabetes using machine learning algorithms, *Mater. Today: Proceedings*, Jul (2021), https://doi.org/10.1016/j.matpr.2021.07.196.

[21] J. Vidya, S.T. Jain, S. Boosi, H.C. Bhanujyothi, C. Tukkoji, Prognosis of diabetes mellitus using machine learning techniques, Art. no. 5, Turkish Journal of Computer and Mathematics Education (TURCOMAT) 12 (5) (Apr. 2021), https://doi.org/10.17762/turcomat.v12i5.1491.

[22] X. Xiong, R. Zhang, Y. Bi, W. Zhou, Y. Yu, D. Zhu, Machine learning models in type 2 diabetes risk prediction: results from a cross-sectional retrospective study in Chinese adults, CURR MED SCI 39 (4) (Aug. 2019) 582–588, https://doi.org/10.1007/s11596-019-2077-4.

[23] J. Semerdjian, S. Frank, An Ensemble Classifier for Predicting the Onset of Type II Diabetes, arXiv:1708.07480 [stat], Aug. 2017. Accessed: Apr. 11, 2022. [Online]. Available: http://arxiv.org/abs/1708.07480.

[24] H.F. Ahmad, H. Mukhtar, H. Alaqail, M. Seliaman, A. Alhumam, Investigating health-related features and their impact on the prediction of diabetes using machine learning, Appl. Sci. 11 (3) (Jan. 2021), https://doi.org/10.3390/app11031173. Art. no. 3.

[25] A.H. Syed, T. Khan, Machine learning-based application for predicting risk of type 2 diabetes mellitus (T2DM) in Saudi Arabia: a retrospective cross-sectional study, IEEE Access 8 (2020) 199539–199561, https://doi.org/10.1109/ACCESS.2020.3035026.

[26] S.B. Choi, et al., Screening for prediabetes using machine learning models, 2014, Comput. Math. Methods Med. (Jul. 2014), e618976, https://doi.org/10.1155/2014/618976.

[27] S.K. Satapathy, S. Dehuri, A.K. Jagadev, S. Mishra, Chapter 1 - introduction, in: S. K. Satapathy, S. Dehuri, A.K. Jagadev, S. Mishra (Eds.), EEG Brain Signal Classification for Epileptic Seizure Disorder Detection, Academic Press, 2019, pp. 1–25, https://doi.org/10.1016/B978-0-12-817426-5.00001-6.

[28] I. Zoppis, G. Mauri, R. Dondi, Kernel methods: support vector machines, in: S. Ranganathan, M. Gribskov, K. Nakai, C. Schönbach (Eds.), Encyclopedia of Bioinformatics and Computational Biology, Academic Press, Oxford, 2019, pp. 503–510, https://doi.org/10.1016/B978-0-12-809633-8.20342-7.

[29] Y. Xia, Chapter Eleven - correlation and association analyses in microbiome study integrating multiomics in health and disease, in: J. Sun (Ed.), Progress in Molecular Biology and Translational Science, vol. 171, Academic Press, 2020, pp. 309–491, https://doi.org/10.1016/bs.pmbts.2020.04.003.

[30] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (3) (Aug. 2004) 199–222, https://doi.org/10.1023/B:STCO.0000035301.49549.88.

[31] K. Roy, S. Kar, R.N. Das, Chapter 6 - selected statistical methods in QSAR, in: K. Roy, S. Kar, R.N. Das (Eds.), Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment, Academic Press, Boston, 2015, pp. 191–229, https://doi.org/10.1016/B978-0-12-801505-6.00006-5.

[32] Y. Song, J. Huang, D. Zhou, H. Zha, C.L. Giles, IKNN: informative K-nearest neighbor pattern classification, in: Knowledge Discovery in Databases: PKDD 2007, 2007, pp. 248–264, https://doi.org/10.1007/978-3-540-74976-9_25. Berlin, Heidelberg.

[33] R.C. Neath, M.S. Johnson, Discrimination and classification, in: P. Peterson, E. Baker, B. McGaw (Eds.), International Encyclopedia of Education, third ed., Elsevier, Oxford, 2010, pp. 135–141, https://doi.org/10.1016/B978-0-08-044894-7.01312-9.

[34] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, KNN model-based approach in classification, in: On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, Berlin, Heidelberg, 2003, pp. 986–996.

[35] S.L. Salzberg, C4.5: programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers, inc, Sep. 1994, Mach. Learn. 16 (3) (1993) 235–240, https://doi.org/10.1007/BF00993309.

[36] G. Stein, B. Chen, A.S. Wu, K.A. Hua, Decision tree classifier for network intrusion detection with GA-based feature selection, in: Proceedings of the 43rd Annual Southeast Regional Conference -, vol. 2, Mar. 2005, pp. 136–141, https://doi.org/10.1145/1167253.1167288. New York, NY, USA.

[37] D. Bienvenido-Huertas, J.E. Nieto-Julián, J.J. Moyano, J.M. Macías-Bernal, J. Castro, Implementing artificial intelligence in H-BIM using the J48 algorithm to manage historic buildings, Int. J. Architect. Herit. 14 (8) (Sep. 2020) 1148–1160, https://doi.org/10.1080/15583058.2019.1589602.

[38] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Chapter 12 - ensemble learning, in: I. H. Witten, E. Frank, M.A. Hall, C.J. Pal (Eds.), Data Mining, fourth ed., Morgan Kaufmann, 2017, pp. 479–501, https://doi.org/10.1016/B978-0-12-804291-5.00012-X.

[39] S. Simske, Chapter 1 - introduction, overview, and applications, in: S. Simske (Ed.), Meta-Analytics, Morgan Kaufmann, 2019, pp. 1–98, https://doi.org/10.1016/B978-0-12-814623-1.00001-0.

[40] D. Talia, P. Trunfio, F. Marozzo, Chapter 1 - introduction to data mining, in: D. Talia, P. Trunfio, F. Marozzo (Eds.), Data Analysis in the Cloud, Elsevier, Boston, 2016, pp. 1–25, https://doi.org/10.1016/B978-0-12-802881-0.00001-9.

[41] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, NY, 2009, https://doi.org/10.1007/978-0-387-84858-7.

[42] D.H. Wolpert, Stacked generalization, Neural Network. 5 (2) (Jan. 1992) 241–259, https://doi.org/10.1016/S0893-6080(05)80023-1.

[43] T. Rahman, et al., QCovSML: a reliable COVID-19 detection system using CBC biomarkers by a stacking machine learning model, Comput. Biol. Med. 143 (Apr. 2022), 105284, https://doi.org/10.1016/j.compbiomed.2022.105284.

[44] V. Chaurasia, S. Pal, Stacking-based ensemble framework and feature selection technique for the detection of breast cancer, SN COMPUT. SCI. 2 (2) (Feb. 2021) 67, https://doi.org/10.1007/s42979-021-00465-3.

[45] M. Petticrew, A. Sowden, D. Lister-Sharp, FALSE-NEGATIVE results in screening programs: medical, psychological, and other implications, Int. J. Technol. Assess. Health Care 17 (2) (Apr. 2001) 164–170, https://doi.org/10.1017/S0266462300105021.

[46] C. Muñoz, et al., Misdiagnosis and diabetic Ketoacidosis at diagnosis of type 1 diabetes: patient and caregiver perspectives, Clin. Diabetes 37 (3) (Jul. 2019) 276–281, https://doi.org/10.2337/cd18-0088.

[47] Z. Liu, C. Fu, W. Wang, B. Xu, Prevalence of chronic complications of type 2 diabetes mellitus in outpatients - a cross-sectional hospital based survey in urban China, Health Qual. Life Outcome 8 (1) (Jun. 2010) 62, https://doi.org/10.1186/1477-7525-8-62.

[48] S.A. Tabish, Lifestyle diseases: consequences, characteristics, causes and control, Jul, Journal of Cardiology & Current Research 9 (2017).

[49] S. Habibi, M. Ahmadi, S. Alizadeh, Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining, Global J. Health Sci. 7 (5) (Sep. 2015) 304–310, https://doi.org/10.5539/gjhs.v7n5p304.

[50] S.A. Mazzuca, et al., The diabetes education study: a controlled trial of the effects of diabetes patient education, Diabetes Care 9 (1) (Jan. 1986) 1–10, https://doi.org/10.2337/diacare.9.1.1.