Archives of Computational Methods in Engineering (2022) 29:313–333
https://doi.org/10.1007/s11831-021-09582-x

ORIGINAL PAPER

Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation

Leila Ismail[1,4] · Huned Materwala[1,4] · Maryam Tayefi[2] · Phuong Ngo[2] · Achim P. Karduck[3]

Received: 13 September 2020 / Accepted: 26 March 2021 / Published online: 15 April 2021

Abstract

Diabetes, one of the top 10 causes of death worldwide, is associated with the interaction between lifestyle, psychosocial, medical conditions, demographic, and genetic risk factors. Predicting type 2 diabetes is important for providing prognosis or diagnosis support to allied health professionals, and aiding in the development of an efficient and effective prevention plan. Several works proposed machine-learning algorithms to predict type 2 diabetes. However, each work uses different datasets and evaluation metrics for algorithms' evaluation, making it difficult to compare among them. In this paper, we provide a taxonomy of diabetes risk factors and evaluate 35 different machine learning algorithms (with and without features selection) for diabetes type 2 prediction using a unified setup, to achieve an objective comparison. We use 3 real-life diabetes datasets and 9 feature selection algorithms for the evaluation. We compare the accuracy, F-measure, and execution time for model building and validation of the algorithms under study on diabetic and non-diabetic individuals. The performance analysis of the models is elaborated in the article.

Keywords Artificial intelligence · Diabetes mellitus type 2 · Diagnosis · Machine learning · Prognosis · Risk factors

1 Introduction

Diabetes Mellitus, commonly referred to as diabetes, is a chronic disease that affects how the body turns food into energy [1]. It is one of the top 10 causes of death worldwide with 4.2 million deaths in 2019 [2]. There are three main types of diabetes: type 1, type 2, and gestational diabetes [1]. Type 1 diabetes is thought to be caused by an autoimmune reaction where the body's immune system affects the insulin-producing beta-cells. Type 2 diabetes is caused by inadequate production of insulin and the inability of the body cells to respond to insulin properly. Gestational diabetes affecting pregnant women mostly during 6 and 9 months of pregnancy is caused by the hormone produced by the placenta leading to insulin resistance. The proportion of people with type 2 diabetes are increasing compared to ones with type 1 and gestational diabetes [2]. In 2019, more than 1.1 million children and adolescents were suffering from type 1 diabetes, while 374 million people were at increased risk of type 2 diabetes. Consequently, in this paper, we focus on type 2 diabetes.

The prevalence of type 2 diabetes in an individual is found to be associated with the interactions of several risk factors that are related to lifestyle, psychosocial, medical conditions, demographic, and genetic (Hereditary) [3–5]. Diagnosis of type 2 diabetes if not performed at an early stage can lead to several serious life-threatening complications [6]. Machine learning-based decision support systems for the prediction of chronic diseases [7–12] have thus gained a lot of attention for better prognosis/diagnosis support to health professionals and public health [13]. Several research efforts have been proposed in the literature for using machine learning classification algorithms to predict the prevalence of type 2 diabetes based on different risk factors [14–45]. These algorithms are either tree-based [46, 47] that construct a classification tree with the dataset features as the nodes and the class labels as the leaves, probability-based [48] that make use of a probability distribution function over the class labels for a given observation, lazy

✉ Leila Ismail
leila@uaeu.ac.ae

1 Intelligent Distributed Computing and Systems Research Laboratory, Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, 15551 Al Ain, Abu Dhabi, United Arab Emirates

2 Norwegian Centre for E-Health Research, Tromsø, Norway

3 Faculty of Informatics, Furtwangen University, Furtwangen, Germany

4 National Water and Energy Center, United Arab Emirates University, Al Ain, Abu Dhabi, United Arab Emirates

approach-based [49, 50] that store the dataset in the memory without developing a model and use the dataset to classify a new observation using a distance function, function-based [51–53] that use a regularization function that aims to minimize the model prediction error, rule-based [54–57] that use IF–THEN statements extracted from decision trees, classifier ensemble-based [58–60] that use a set of classifiers whose individual decisions are combined by using a voting mechanism, clustering and classifier ensemble-based [61] that first perform clustering on the dataset to remove outliers and then apply a classification algorithm, or meta heuristic-based [62–64] that combine the output of different classification algorithms with each modelled using a random sample of the dataset. The work in the literature evaluates and compares different classification algorithms using heterogeneous datasets and evaluation metrics. However, we are unaware of any objective comparison of these algorithms using unified datasets and evaluation metrics. This paper aims to address this void.

The key research contributions are as follows.

- We classify type 2 diabetes risk factors based on their common characteristics to analyze which categories are more significant than others in predicting type 2 diabetes.
- We evaluate the performance of 35 different algorithms in terms of accuracy, F-measure, and execution time in a unified setup using 3 real-life diabetes datasets with and without feature selection.

2 Taxonomy of Type 2 Diabetes Risk Factors

In this section, we present a taxonomy of type 2 diabetes risk factors (Fig. 1). We classify them into five categories: (1) lifestyle, (2) medical condition, (3) hereditary, (4) psychosocial, and (5) demographic. The purpose of this classification is to analyze which category of the risk factors significantly contributes to the prediction of type 2 diabetes. Lifestyle factors refer to the ones that are highly influenced by the lifestyle and environment of an individual. The medical condition-based factors are related to the presence of certain diseases in an individual such as serum uric acid, obesity, hypertension, cardiovascular disease. A serum uric acid level of more than 370 µmol/l is considered high in an individual and associated with the prevalence of type 2 diabetes [65]. Obesity is defined as the excessive amount of fat accumulation in an individual and characterized by a Body Mass Index (BMI) higher than 30 [66]. High blood pressure is the medical condition in which the blood pressure in the arteries remains persistently elevated and characterized by systolic pressure 140–159 mmHg or diastolic pressure 90–99 mmHg [66]. Cardiovascular disease refers to the conditions affecting heart or blood vessels such as abnormal heart rhythms, heart attack, heart failure, and stroke [67]. The risk factors that are passed on from one generation to another fall under the hereditary category. The psychosocial factors include illness related to the mental health of an individual, and the demographic risk factor refers to the characteristics of an individual.

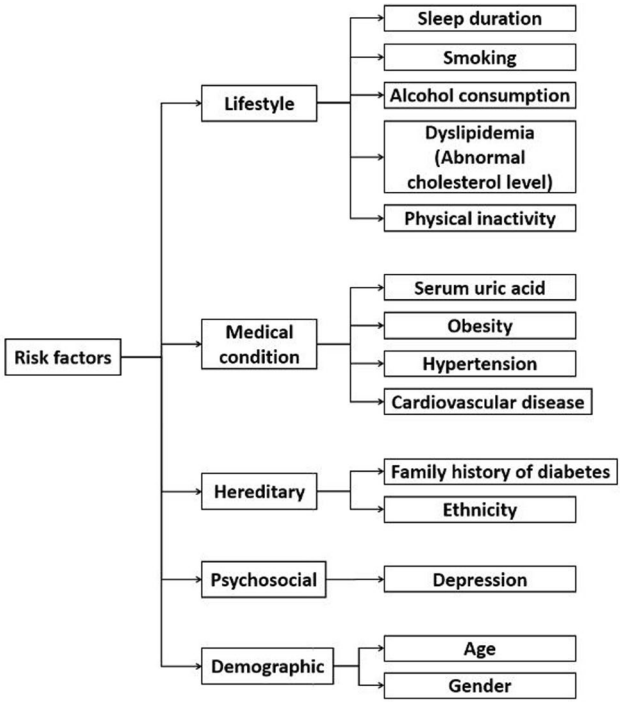

Fig. 1 Taxonomy of risk factors for type 2 diabetes

3 Machine Learning Classification Algorithms

In this section, we describe the implementation of the machine learning algorithms under study for the prediction of type 2 diabetes.

3.1 Decision Tree (DT)

It constructs a tree structure to define the sequences of decisions and outcomes [46] and to use it for prediction. At each node of the tree, the algorithm selects the branch having the maximum information gain (Eq. 1).

$$InfoGain_R = H_{diabetes} - H_{diabetes|R} \tag{1}$$

where R is the risk factor, $H_{diabetes}$ represents the base entropy (Eq. 2), and $H_{diabetes|R}$ represents the conditional entropy (Eq. 3).

$$H_{diabetes} = \sum_{\forall diabetese\{diabetes,non-diabetes\}} P(diabetes)\log_2 P(diabetes) \tag{2}$$

where P(diabetes) is the probability of the number of observations in the diabetes class compared to the total number of observations.

$$H_{diabetes|R} = \sum_{r} P(r)H(diabetes|R=r) = \sum_{\forall r \epsilon R} P(r) \sum_{\forall diabetese \epsilon \{diabetes, non-diabetes\}} P(diabetes|r)\log_2 P(diabetes|r) \quad (3)$$

where R is a risk factor and r is the set of its values from all the observations in the dataset.

A one-level Decision tree, i.e., a tree where the root is immediately connected to the leaf nodes, is known as Decision Stump (DS) [47] that makes the prediction based only on the risk factor with the highest information gain.

3.2 Bayesian Network (BN)

It is a graphical representation of probabilistic relationships among a set of risk factors and diabetes prevalence [48]. BN is a set of edges (E) and risk factors ($R = \{R_1, R_2, \ldots, R_n\}$), which forms a directed acyclic graph (DAG) G=(R, E) that encodes a joint probability distribution over R. Each node is represented using a conditional probability distribution (CPD) as stated in Eq. (4).

$$CPD = P(R_i|Pa(R_i)) \quad (4)$$

where $Pa(R_i)$ indicates the parent of R_i in G.

In a Bayesian network, a joint distribution of risk factors is obtained by multiplying the CPD for each risk factor as stated in Eq. (5).

$$P(R_1, R_2, \ldots, R_n) = \prod_{i=1}^{n} CPD(R_i) \quad (5)$$

3.3 Naïve Bayes (NB)

It is primarily based on Bayes' theorem that gives the relationship between the probabilities of two risk factors and their conditional probabilities [46]. Given that a risk factor A already exists in an individual, the conditional probability of a risk a factor C occurring is given by Eq. (6).

$$P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{P(A|C) \cdot P(C)}{P(A)} \quad (6)$$

For a record with a set of risk factors ($R_1, R_2, \ldots, R_n$), the goal is to predict diabetes/non-diabetes class C_i from the set of classes, $C = \{C_{diabetes}, C_{non\text{-}diabetes}\}$ that maximizes the conditional probability $P(C_i|R_1, R_2, \ldots, R_n)$. The general form of Bayes' theorem for assigning diabetes or non-diabetes class to observation with multiple risk factors is given by Eq. (7).

$$P(C_i|R_1, R_2, \ldots, R_n) = \frac{P(R_1, R_2, \ldots, R_n|C_i) \cdot P(C_i)}{P(R_1, R_2, \ldots, R_n)} \quad (7)$$

The NB algorithm extends the Bayes' theorem mentioned in Eq. (7) using two assumptions: 1) each risk factor is conditionally independent of every other factor given a class C_i as shown in Eq. (8), and 2) ignoring the term $P(R_1, R_2, \ldots, R_n)$ in Eq. (7). Consequently, the probability of C_i given the probabilities of all the risk factors, $P(C_i|R_1, R_2, \ldots, R_n)$ can be calculated using Eq. (9).

$$P(R_1, R_2, \ldots, R_n|C_i) = P(R_1|C_i) \cdot P(R_2|C_i) \ldots P(R_n|C_i) = \prod_{j=1}^{n} P(R_j|C_i) \quad (8)$$

$$P(C_i|R_1, R_2, \ldots, R_n) = P(C_i) \cdot \prod_{j=1}^{n} P(R_j|C_i) \quad (9)$$

3.4 K Nearest Neighbors (K-NN)

It stores the dataset and classifies a new observation based on how likely it is to be a member of diabetes or non-diabetes class [49]. The algorithm calculates the distance of the new observation from all the existing ones in the dataset. It then assigns the new observation to a class that appears the maximum number of times in a group of k (positive integer parameter) neighbors.

3.5 K Star

It uses an entropic measure based on the probability of transforming one observation into another by randomly choosing between all possible transformations [50]. A new observation is assigned to the same diabetes or non-diabetes class as that of one in the dataset having the shortest distance from the new observation. The distance between observations in K star is indicated by the complexity of transforming one observation into another. Let I be a set of observations and T a finite set of transformations on I. Let P be the set of all prefix codes from T* which are terminated by σ. σ is a member of T which maps observations to themselves, i.e., $(\sigma(a) = a)$. Members of T* define a transformation on I as shown in Eq. (10).

$$\bar{t}(a) = t_n(t_{n-1}(\ldots t_1(a)\ldots)), where\ \bar{t} = t_1, \ldots, t_n \quad (10)$$

A probability function p is defined on T* that satisfies the conditions stated in Eq. (11).

$$0 \leq \frac{p(\bar{t}u)}{p(\bar{t})} \leq 1, \sum_{u} p(\bar{t}u) = p(\bar{t}),\ and\ p(\Lambda) = 1 \tag{11}$$

The function P* is defined as the probability of all paths from observation a to observation b (Eq. 12).

$$P^*(b|a) = \sum_{\bar{t}\epsilon P:\bar{t}(a)=b} p(\bar{t}) \tag{12}$$

The K star function is defined using Eq. (13).

$$K^*(b|a) = -\log_2 P^*(b|a) \tag{13}$$

3.6 Logistic Regression (LR)

It predicts the probability that a given observation belongs to diabetes or non-diabetes class using a sigmoid function [51] as stated in Eq. (14).

$$p(diabetes) = \frac{e^{\beta_0+\sum_{i=1}^{n}\beta_i R_i}}{1+e^{\beta_0+\sum_{i=1}^{n}\beta_i R_i}} \tag{14}$$

where p(diabetes) represents the probability of having diabetes, R is the set of risk factors, and β_0 and β_i are the regression coefficients representing the intercept and the slope respectively. The values of regression coefficients are calculated using maximum likelihood estimation such that the value of Eq. (15) is the maximum.

$$l(\beta_0, \ldots, \beta_i) = \prod_{i:y_i=1} P(diabetes_i) \prod_{i':y_i'=0} (1 - p(non{-}diabetes_i) \tag{15}$$

3.7 Support Vector Machine (SVM)

It aims to create a decision boundary known as a hyperplane that can separate n-dimensional instance space into diabetes and non-diabetes classes [52]. The hyperplane is created using the extreme points (support vectors) of the dataset. The generation of hyperplane is an iterative process to find the maximum possible margin between the support vectors of the opposite classes. Let r_i and y_i represent the risk factors and classes in the dataset and there exists a hyperplane that separates diabetes and non-diabetes classes as stated in Eq. (16).

$$\begin{aligned} &w^T r + b = 0 \\ &\text{s.t., } w^T r^{(i)} + b > 0, if\ y^{(i)} = +1\ and\ w^T r^{(i)} + b < 0,\ if\ y^{(i)} = -1 \end{aligned} \tag{16}$$

where w is the normal of the hyperplane and b is the bias.

The minimization problem to obtain the optimal hyperplane that maximizes the margin can be formulated using Eq. (17).

$$Minimize\ \emptyset(W) = \frac{1}{2}||W||^2,\ such\ that\ y_i(W \cdot r_i + b) \geq 1 \tag{17}$$

In the case of a non-linear dataset, SVM uses a kernel trick to transform the input space to a higher dimensional space to generate a hyperplane. SVM uses different kernel functions such as linear, polynomial, Radial Basis Function (RBF), and sigmoid.

3.8 Artificial Neural Networks (ANN)

It is an iterative process that consists of networks of neurons based on the neural structure of the human brain [53]. In this paper, we use Multilayer Perceptron (MLP) which is a feedforward neural network that utilizes backpropagation for training. An MLP consists of three layers: risk factors (input), hidden, and diabetes-and-non-diabetes classes (output). Except for the risk factors nodes, each node in the network is a neuron that uses a non-linear activation function. In this paper, we use the sigmoid activation function to develop a non-linear relationship between the risk factors and the diabetes-and-non-diabetes classes. MLP is an iterative process and after each iteration, the algorithm compares the result of the output layer with actual diabetes or non-diabetes class labels and calculates an error term for each node. These error terms are then used to adjust the weights in the hidden layers such that the prediction accuracy increases in the next iteration. The output of each hidden layer is calculated using Eq. (18).

$$a = \emptyset(Wi + b) \tag{18}$$

where W is the weight matrix for each risk factor, i is the input vector consisting of the risk factors, b is the bias vector, Ø(.) is the sigmoid activation function and a is the vector output consisting of diabetes and non-diabetes class labels.

3.9 Zero Rule (ZeroR)

It is a frequency-based algorithm [54]. It labels all observations based on the majority (diabetes or non-diabetes). It does not require a model development for diabetes prediction.

3.10 One Rule (OneR)

It is a frequency-based algorithm that generates one rule for each risk factor in the dataset to predict diabetes and then selects the rule with the smallest error [55]. The algorithm first constructs a frequency table for each risk factor against diabetes and non-diabetes classes to create the rules for the risk factors. The error for each rule is then calculated by dividing the number of the observations in the minority class

(diabetes or non-diabetes) by the total number of observations as stated in Eq. (19).

$$Error = \frac{Number\ of\ observations\ in\ the\ minority\ class}{(Total\ number\ of\ observations)} \quad (19)$$

3.11 JRip

It is based on association rules with reduced error pruning by implementing Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [56]. An initial set of rules is generated for the risk factors that cover all the observations in the diabetes class. This overlarge set of rules is then iteratively simplified by performing pruning operations. At each stage of simplification, a rule is removed such that there is the highest error reduction. The iteration stops when further pruning operations would increase the error.

3.12 Decision Table (DTable)

It is a set of If-Then rules that are formulated in terms of a candidate decision table [57]. A set of rules for the association of risk factors and prevalence of diabetes are stored in a

Table 1 Advantages and disadvantages of classification algorithms

Algorithm	Advantages	Disadvantages
DT [46]	Suitable for datasets having missing values and data scaling and normalization is not required Easy to implement and interpret	Sensitive to change The probability of overfitting is high
BN [48]	Suitable for datasets having missing values	Not suitable with small datasets Computationally expensive
NB [46]	Suitable for datasets with missing values and is scalable Easy to implement	Suffers from the issue of zero frequency
K-NN [49]	Suitable for datasets having outliers Easy to implement	Determining the value of k is challenging High computation cost
K star [50]	Suitable for datasets having outliers Easy to implement	Not suitable for large datasets High computation cost
LR [51]	Suitable for large datasets Easy to interpret	Not suitable for linear data and datasets having a smaller number of observations than features
SVM [52]	Suitable for high dimensional and non-linear datasets	Feature scaling is required The output is difficult to interpret Selection of kernel is difficult
ANN [53]	Suitable for high dimensional datasets having a greater number of observations and can handle missing values	High computational cost Complex process
ZeroR [54]	Easy to understand Used as a baseline benchmark	No prediction involved
OneR [55]	Easy to understand Used as a baseline benchmark	Only suitable for datasets having categorical features Not suitable for linear data
JRip [56]	Suitable for non-linear data Easy to implement and interpret	Only suitable for datasets having categorical features Suffers from overfitting
DTable [57]	Suitable for dynamic datasets Easy to implement and interpret	Prone to overfitting Complex for high dimensional datasets
RT [58]	Suitable for datasets having missing values and data scaling and normalization is not required Easy to implement and interpret	Sensitive to changes in the dataset The probability of overfitting is high
RF [59]	Suitable for high dimensional datasets and can handle missing values	Difficult to implement Complex algorithm
REPTree [60]	Suitable for large datasets Easy to interpret compared to a decision tree	Sensitive to changes in the dataset Prone to overfitting
K-means [61]	Suitable for large datasets Simple to implement and interpret	Difficult to predict the value of k Sensitive to outliers
Bagging [62]	Suitable for high dimensional datasets and can handle missing values Reduces data overfitting	Model is biased Computationally expensive
Boosting [63]	Reduces data overfitting Easy to interpret	Not suitable for large datasets Sensitive to outliers
Stacking [64]	Reduces overfitting	Memory intensive

candidate decision table. The table is then pruned by removing redundant rules and the rules having a confidence value less than 1. The confidence for a risk factor rule is calculated as stated in Eq. (20).

$$Confidence = \frac{Total\ number\ of\ observations\ having\ same\ risk\ factors\ for\ the\ diabetes\ class}{Total\ number\ of\ observations\ having\ same\ risk\ factors} \quad (20)$$

3.13 Random Tree (RT)

It is similar to a decision tree where the algorithm generates a decision tree by selecting the risk factors with high information gain. However, instead of considering all the risk factors for selection, RT uses a set of randomly selected factors [58].

3.14 Random Forest (RF)

It is a set of decision trees constructed using randomly selected samples of the dataset [59]. It performs voting on the output of each decision tree and classifies an observation into diabetes or non-diabetes depending on the majority of the decision trees' output.

3.15 Reduced Error Pruning Tree (REPTree)

It builds several decision trees iteratively using different risk factors and selects the tree with the least prediction error. It builds a tree for the risk factors with the risk factor having the highest information gain as the root similar to DT. The algorithm then prunes the tree using reduced error pruning. This is by removing the subtree rooted at each risk factor and making that risk factor as a leaf node by assigning the majority diabetes or non-diabetes class. If the error rate of the new tree is less than or equal to the original tree, then pruning is done [60]. The algorithm iterates until further pruning reduces the classification performance.

3.16 K-means

It is a clustering technique rather than a classification [61]. It is an iterative algorithm that divides the given dataset into diabetes and non-diabetes clusters. The observations having similar risk factors values are placed in the same cluster. K-means is used along with a classification algorithm

Table 2 List of feature selection algorithms used in the experiments

Feature selection algorithm	Description
Correlation-based Feature Selection (CFS) Subset Evaluator (CSE) [71]	Evaluates the rank of a subset of features by considering the features that are highly correlated with the class labels and less correlated with other features
Classifier Attribute Evaluator (CAE) [72]	Evaluates the rank of a feature using a user-specified classifier. The weight of each feature is determined by the performance degradation of the classifier when evaluated without that feature
Correlation Attribute Evaluator (CAE) [54]	Evaluates Pearson's correlation of each feature and the class labels and select the features that have a moderate positive correlation or negative correlation
Gain Ratio Attribute Evaluator (GRAE) [73]	Evaluates the weight of each feature by calculating its gain ratio for each class. The ratio is calculated by dividing the difference between the class entropies over the entropy of the feature
OneR Attribute Evaluator (OAE) [73]	Evaluates the weight of a feature by using the OneR classifier. The classifier is applied to each feature and the ones having the best classification performance are selected
Principal Component (PC) [73]	Combines the subset of features such that the removed features have less variance on the performance of the classifier
Relief Attribute Evaluator (ReAE) [73]	Evaluates the weight of a feature by iteratively sampling an observation and considering the value of the given feature for the nearest observation of the same and different class
Symmetrical Uncertainty Attribute Evaluator (SUAE) [73]	Evaluates the weight of each feature by measuring the symmetrical uncertainty of the feature. The symmetrical uncertainty is calculated as: 2*(the difference between the entropy of the class and that related to the feature) divided by the sum of the entropies of the class and the feature
Information Gain Attribute Evaluator (IGAE) [73]	Evaluates the weight of each feature by calculating the information gain as the difference between the entropy of the class and that related to the feature

Table 3 Specification of pre-processed datasets used in the experiments

Dataset	#features	Features	#observations	#positive diabetes class	#negative diabetes class
PIMA Indian	9	*Numerical*—pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, age *Binary*—diabetes prevalence	537	179	358
UCI	12	*Categorical*—race, gender *Numerical*—age *Binary*—Alcohol, blood pressure, cholesterol, heart disease, obesity, pregnancy, uric acid, blurred vision, diabetes prevalence	65,840	51,034	14,806
MIMIC III	4	*Categorical*—ethnicity, gender *Numerical*—age *Binary*—diabetes prevalence	39,698	1,242	38,456

to increase the classification performance. The algorithm performs K-means and removes the incorrectly clustered observations, i.e., individuals with diabetes classified as non-diabetes and vice-versa. The dataset having correctly clustered observations is then used by a classifier to build a classification model.

3.17 Bagging

It is an ensemble meta-estimator that uses another classification algorithm as a parameter [62]. It first randomly generates multiple samples of the observations from the dataset with replacement. The base classifier is then applied to each random subset of the dataset for the diabetes prediction. The algorithm then aggregates the classifier outputs for each subset and selects the final output based on voting.

3.18 Boosting

It is an ensemble meta-estimator that aims to boost the classification performance for diabetes prediction by creating a strong classifier from several weak classifiers [63]. It builds a model from the dataset to predict diabetes and then creates a second model that attempts to correct the errors from the first model. Models are added until the training set is predicted as expected or the maximum number of models are added. The models when added to the other models are weighted based on their weakness. The weakness is identified by the classifier's error rate as stated in Eq. (21). After each model is added, the data weights are readjusted, known as re-weighting.

$$Error = \frac{\sum_{j=1}^{n} W_j I\left(C\left(R_j\right) \neq Y_j\right)}{\sum_{j=1}^{n} W_j}, I(x) = \begin{cases} 1, & if\ r\ is\ true \\ 0, & if\ r\ is\ false \end{cases} \tag{21}$$

where C is the classifier, R is the matrix containing the risk factors, Y is a vector containing diabetes and non-diabetes class labels and W is the assigned weight.

3.19 Stacking

It is an ensemble meta-estimator in which a classification model is developed using the observations to predict diabetes or non-diabetes classes. The output of this classifier is then used as an input to develop another classification model for diabetes prediction [64].

Table 1 shows the advantages and disadvantages of the classification algorithms used in the literature for the prediction of type 2 diabetes. It shows which algorithm is suitable for which dataset type.

4 Performance Analysis

In this section, we analyze and compare the performance of the studied algorithms in terms of accuracy, F-measure, and execution time.

4.1 Datasets

We use Weka 3.8 [54] for the implementation and evaluation of the studied algorithms. We evaluate the performance of the algorithms with and without feature selection using three datasets, i.e., PIMA Indian [68], UCI [69], and MIMIC III [70]. PIMA Indian dataset is from the National Institute of Diabetes and Digestive and Kidney Diseases and is used to predict whether or not a patient has diabetes, based on diagnostic measurements. UCI dataset includes patient and hospital outcomes from 130 US hospitals between 1999–2008. MIMIC III contains information of over 40,000 patients who

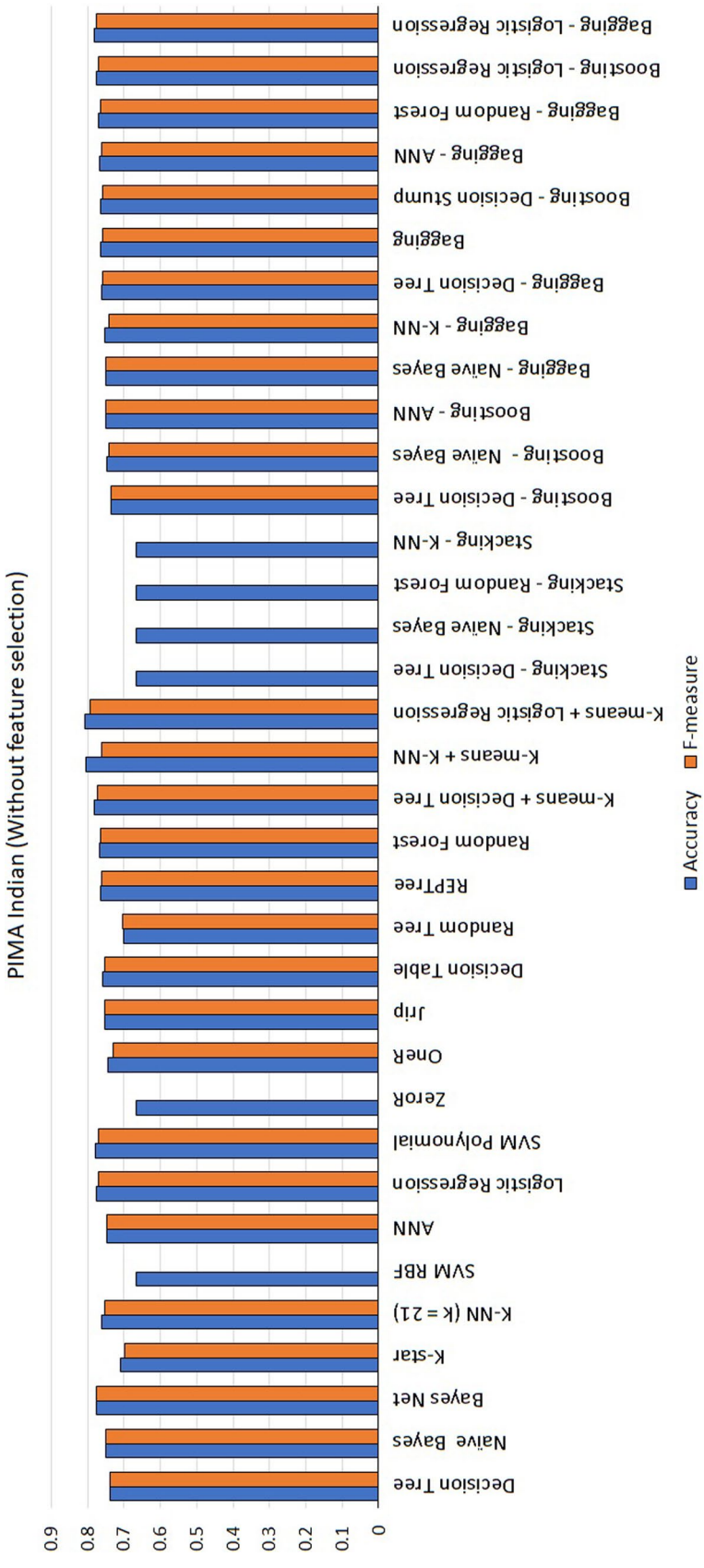

Fig. 2 Accuracy and F-measure of the algorithms for the PIMA Indian Dataset without feature selection

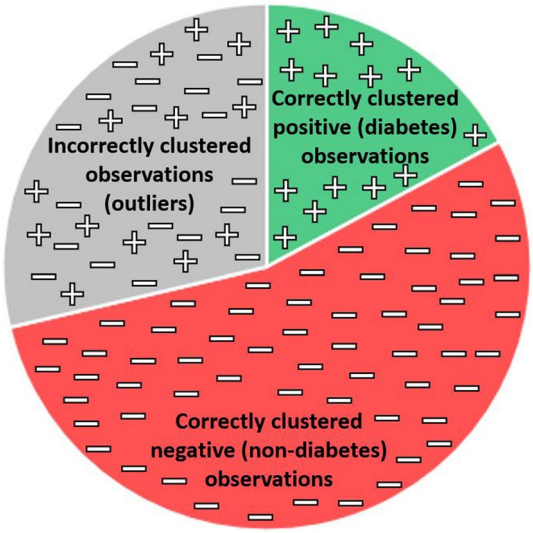

Fig. 3 Removal of incorrectly clustered diabetes observations by K-means clustering on PIMA Indian Dataset

stayed in critical care units of the Berth Israel Deaconess Medical Center between 2001 and 2012.

4.2 Feature Selection Algorithms

Table 2 shows the feature selection algorithms used.

4.3 Experiments

The UCI dataset includes race, gender, age, diagnosis 1, diagnosis 2, diagnosis 3, and diabetes medication columns. Diagnosis 1, 2, and 3 represent the results of primary, secondary, and additional secondary diagnoses respectively. We create a class label for diabetes based on the diabetes medication column. For diagnosis 1, 2, and 3 columns, we extract the values of diseases that are risk factors for type 2 diabetes such as obesity, hypertension, and cardiovascular disease, and use them as binary features in the dataset. For the MIMIC III dataset, we consider the available diabetes risk factor features that are ethnicity, gender, age, and diabetes. Table 3 shows the total number of features, observations, positive and negative diabetes classes for each dataset. It shows that the number of observations in the positive class (negative class) for the UCI (MIMIC III) dataset is very high compared to that in the negative (positive) class, leading to imbalanced datasets [74].

We evaluate the algorithms under study with and without features selection using the tenfold cross-validation method [75] where the dataset is divided into k partitions. One partition is for testing data and *k-1* partitions for training with replacement. This is repeated until each partition is used for training and testing. The resultant model is then obtained by averaging the result of each iteration. For K-NN, we run the algorithm for different values of 'k' from 1 to square root of the number of observations and select the value of 'k' that gives the highest accuracy. For SVM, we use the polynomial and RBF kernels. Each algorithm is executed 3 times on each dataset (we call it a run) and the average for accuracy, F-measure, and execution time is calculated over each run. For the feature selection, we execute the selection algorithms (Table 2) on each dataset. We calculate the frequency of the features selected by each algorithm and select the features which appear in more than 50% of the algorithms. The accuracy and the F-measure are calculated using Eq. (22) and Eq. (23) respectively. The execution time is calculated by adding the model building and validation times.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{22}$$

where *TP* is True Positive, *TN* is True Negative, *FP* is False Positive, and *FN* is False Negative. TP (TN) represents the number of observations in the positive (negative) class that are classified as positive (negative), and FP (FN) represents the number of observations in the negative (positive) class that are classified as positive (negative).

$$F{-}measure = \frac{2(Recall * Precision)}{(Recall + Precision)} \tag{23}$$

where recall and precision for the positive (negative) class are calculated using Eqs. (24) and (25) respectively.

$$Recall = \frac{TP(TN)}{TP(TN) + FN(FP)} \tag{24}$$

$$Precision = \frac{TP(TN)}{TP(TN) + FP(FN)} \tag{25}$$

4.4 Experimental Results Analysis

In this section, we analyze our experimental results and give insights and conclusions. In particular, we reveal the reasons behind the performance of these algorithms.

4.4.1 Classification Algorithms without Feature Selection

Figure 2 shows the accuracy and F-measure of the algorithms for the PIMA Indian dataset without feature selection. The accuracy of K-means + Logistic Regression (LR) algorithm is the highest among all the studied ones. This is because the algorithm removes the incorrectly clustered observations (outliers) from the dataset using K-means clustering (Fig. 3) before developing a classification model. The observations with the positive (negative) class label placed in the negative (positive) cluster are

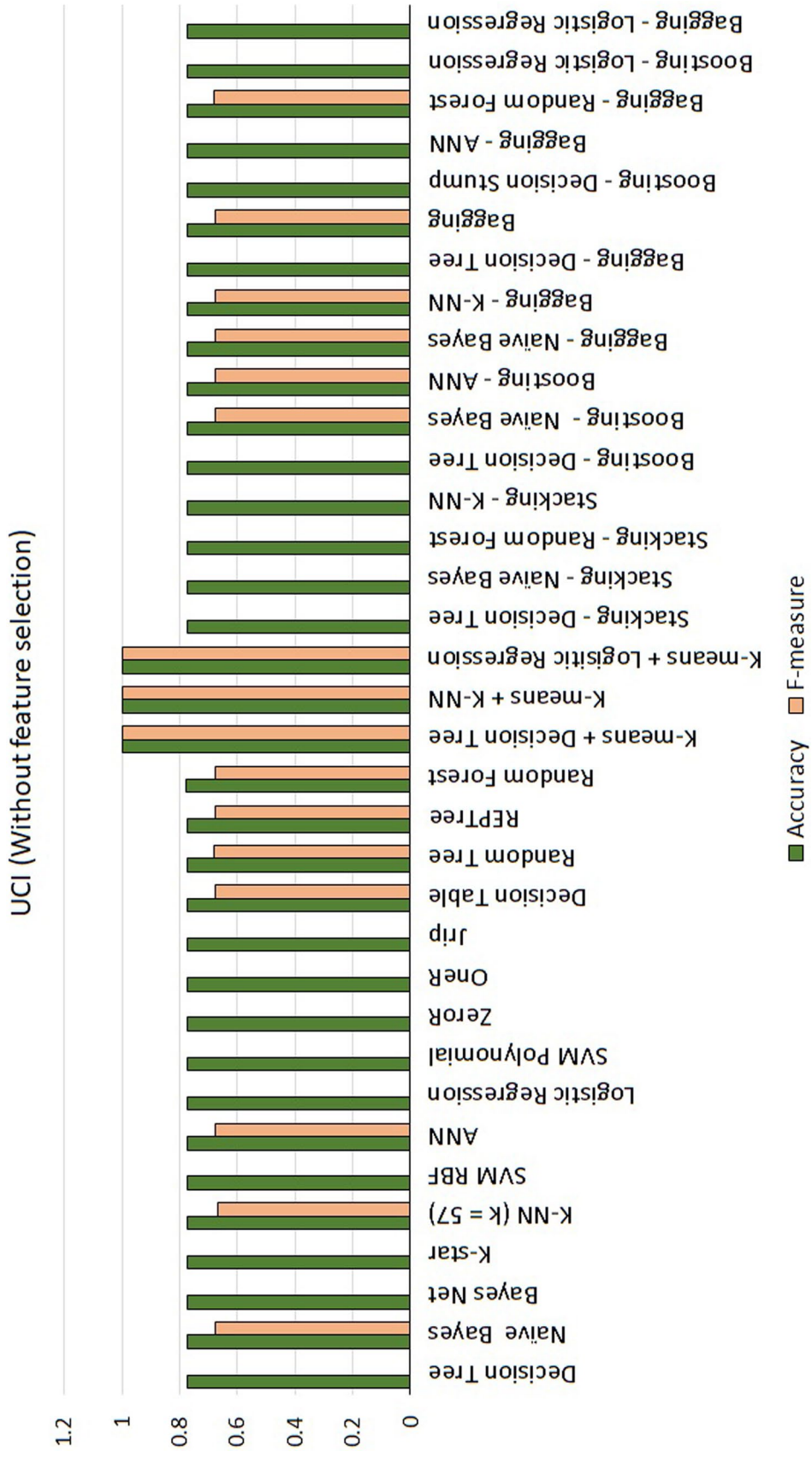

Fig. 4 Accuracy and F-measure of the algorithms for the UCI Dataset without feature selection

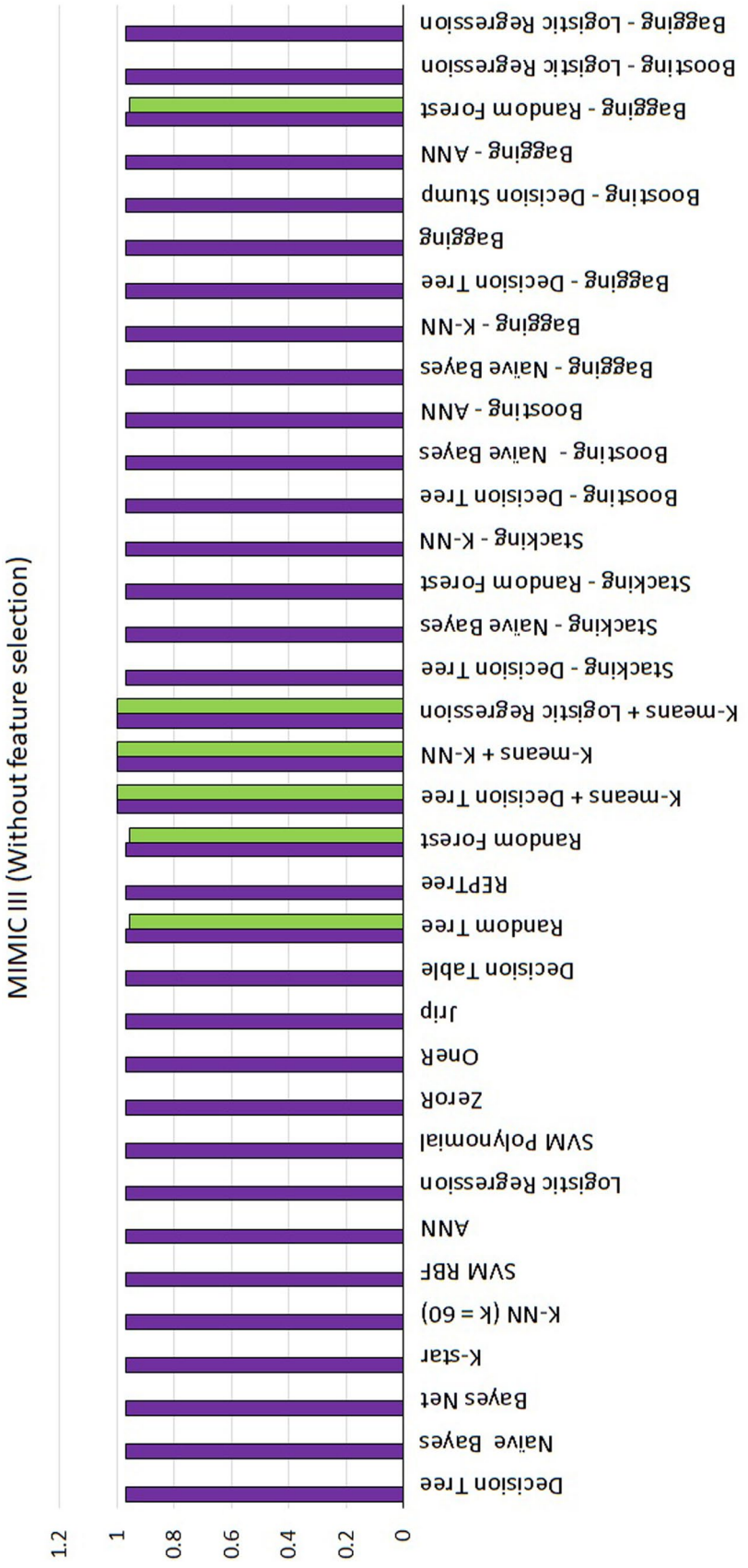

Fig. 5 Accuracy and F-measure of the algorithms for the MIMIC III Dataset without feature selection

Table 4 Execution time for model building and validation without feature selection

Algorithms	The execution time of model building and validation (min)		
	PIMA Indian	UCI	MIMIC III
DT	0	0.267	0.034
NB	0	0.016	0
BN	0	0.05	0
K star	0.016	103.116	1.216
K-NN	0	7.4	1.466
SVM-RBF	0.016	708.85	17.166
ANN	0.05	1454.383	48.5
LR	0	0.166	0.3
SVM-polynomial	0	48.634	2.8
ZeroR	0	0	0
OneR	0	0.2	0
JRip	0	0.1	0.0167
DTable	0	0.5	0.083
RT	0	0.066	0.016
REPTree	0	0.116	0.016
RF	0.016	4.3	0.716
K-means + DT	0	0.016	0
K-means + K-NN	0	1.334	0.284
K-means + LR	0	0.116	0.05
Stacking-DT	0	0.033	0
Stacking-NB	0	0.033	0.016
Stacking-RF	0	0.516	0.25
Stacking-K-NN	0	5.916	2.184
Boosting-DT	0.016	3.25	0.3
Boosting-NB	0	1	0.084
Boosting-ANN	0.5	56.5	108.984
Boosting-DS	0	0.35	0.05
Boosting-LR	0	0.85	0.183
Bagging-NB	0.016	0.183	0.033
Bagging-K-NN	0	66.883	15.816
Bagging-DT	0.016	2.334	0.15
Bagging-REPTree	0	1.167	0.167
Bagging-ANN	0.6	1584.5	478.483
Bagging-RF	0.234	39.933	7.334
Bagging-LR	0.016	1.483	3.55

known as the outliers. The LR is then applied to the dataset without outliers resulting in higher accuracy. However, the removal of outliers before implementing LR is not justified for the diabetes dataset. For instance, if a diabetic patient having all the risk factors is clustered in the non-diabetes cluster, then the patient will be removed by the k-means as an outlier. Consequently, the algorithm will be trained using the dataset with those patients removed. This will classify the potential individual at risk of developing diabetes as non-diabetic. On the other hand, the LR model, when used with the Bagging heuristic (Bagging-LR), performs better without removing the outliers. This is thanks to the logistic function used by the meta-heuristic algorithm which can linearly separate diabetes and non-diabetes binary classes for the interdependent features such as cholesterol and obesity. Figure 2 shows that there is no F-measure value for SVM with RBF kernel, ZeroR, and stacking with DT, NB, RF, and K-NN. This is because these algorithms predict the negative (non-diabetes) class which is the majority in the dataset and the positive (diabetes) class is never predicted.

Figures 4 and 5 show the accuracy and F-measure of the algorithms without feature selection for the UCI and MIMIC III datasets respectively. RF has the highest accuracy, whereas most of the studied algorithms are unable to detect the minority class as the datasets are imbalanced. RF performs better with an imbalanced dataset because while constructing an individual decision tree, the algorithm bootstraps a sample from the minority class and the same size of the sample with replacement from the majority class. Consequently, each decision tree algorithm in RF is applied to a balanced subset of the dataset leading to a highly accurate classifier with both classes of diabetes being detected. Most of the algorithms in Figs. 4 and 5 have no F-measure value. This is because the UCI and MIMIC III datasets are highly imbalanced, and the algorithms are not able to predict the minority class.

Table 4 shows the execution times for model building and validation of the algorithms under study for each dataset. The execution times of the algorithms for the PIMA Indian

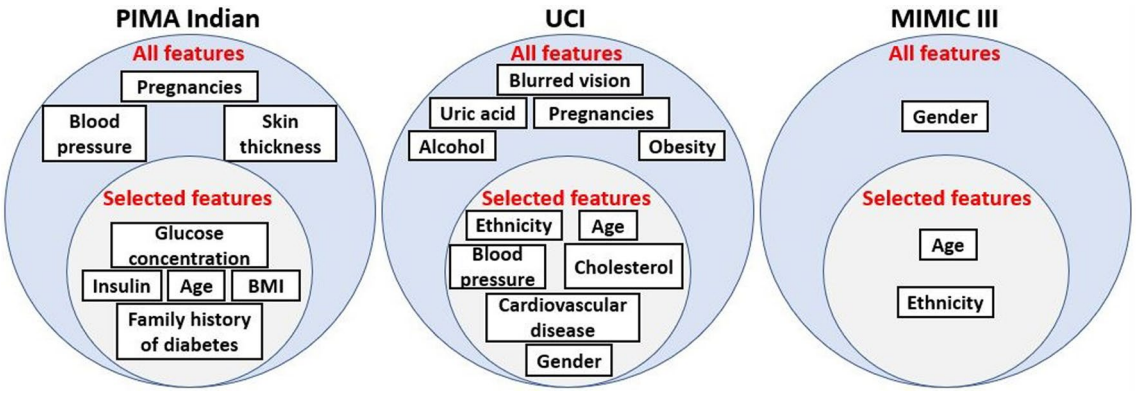

Fig. 6 Selected features for each dataset using feature selection algorithms

dataset are negligible while, for the UCI are the highest. This is because time is a function of the number of features and observations.

4.4.2 Classification Algorithms with Feature Selection

Figure 6 shows the results of the feature selection algorithms. The results reveal that the following risk factors have a significant impact on the prediction of type 2 diabetes: age (demographic category), ethnicity and family history of diabetes (hereditary category), hypertension, obesity, and cardiovascular disease (medical conditions category), and cholesterol (lifestyle category). Figures 7, 8 and 9 show the performance of the studied algorithms for each dataset with feature selection. The relative performance of the algorithms remains the same as that without feature selection. The most accurate algorithm for PIMA Indian is Bagging-LR while that for UCI and MIMIC III datasets is RF. Table 5 shows the execution time for model building and validation of each algorithm under study. The relative performance is similar to that without feature selection (Table 4). However, the execution times for algorithms with feature selection are less than the ones without feature selection in all the datasets under experiment. This is because of the reduced number of features. The execution time for PIMA Indian is reduced up to 2.5 times and for UCI is 1.23 times. There is no significant reduction in execution times for the MIMIC III dataset as only one feature is removed.

5 Related Work

In the last decade, there have been many research efforts by academic and industrial researchers to predict type 2 diabetes using machine learning algorithms [14–45]. However, the algorithms in these works are compared using different datasets and evaluation metrics (Table 6), making an objective comparison difficult. In this paper, we evaluate and compare 35 algorithms in a unified setup.

6 Conclusions

Urbanization, uneven diet, and changing lifestyles have led to the increase in type 2 diabetes globally. Many works in the literature have compared several algorithms to accurately predict type 2 diabetes. Those algorithms were evaluated using different datasets and evaluation metrics, making it difficult to compare their relative performance.

In this paper, we evaluate these algorithms using three diabetes datasets in a unified setup and compare their performance in terms of accuracy, F-measure, and execution time. Our experimental results show that the Bagging-LR algorithm is the most accurate for a balanced dataset with and without feature selection while, for an imbalanced dataset, RF is the most accurate. In addition, we classify type 2 diabetes risk factors to analyze the most significant categories for diabetes prediction.

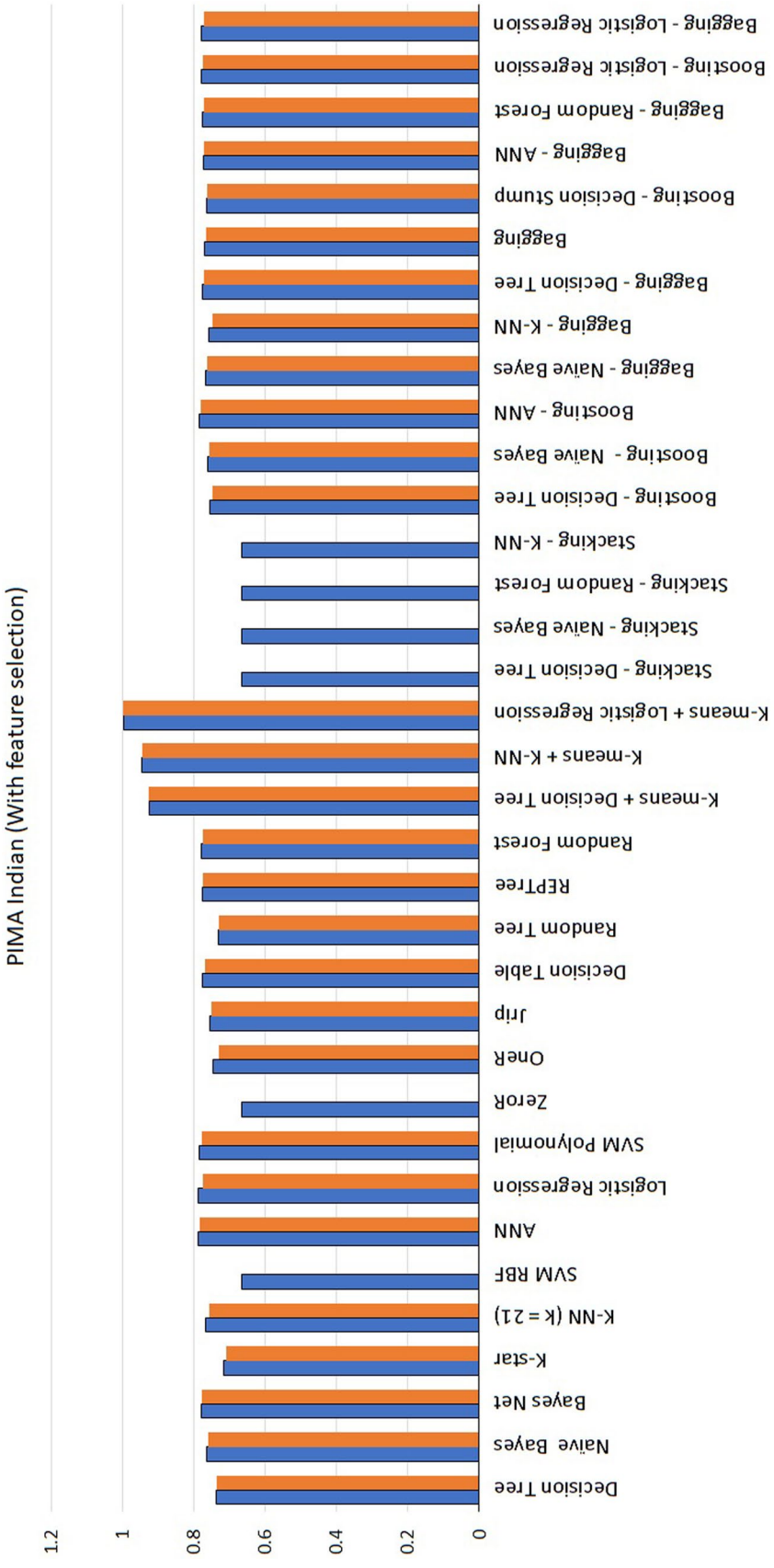

Fig. 7 Accuracy and F-measure of the classification algorithms for the PIMA Indian Dataset with feature selection

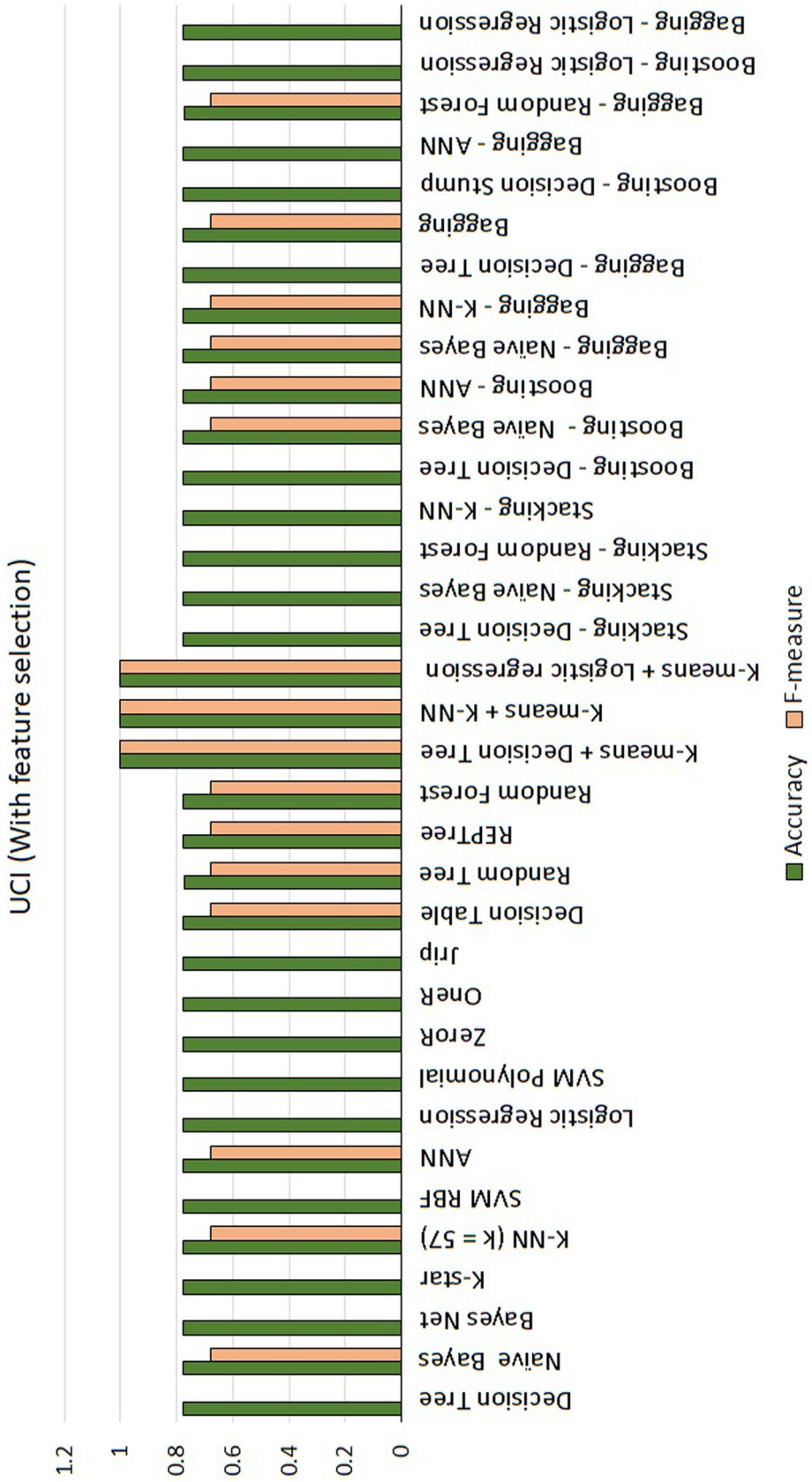

Fig. 8 Accuracy and F-measure of the classification algorithms for the UCI Dataset with feature selection

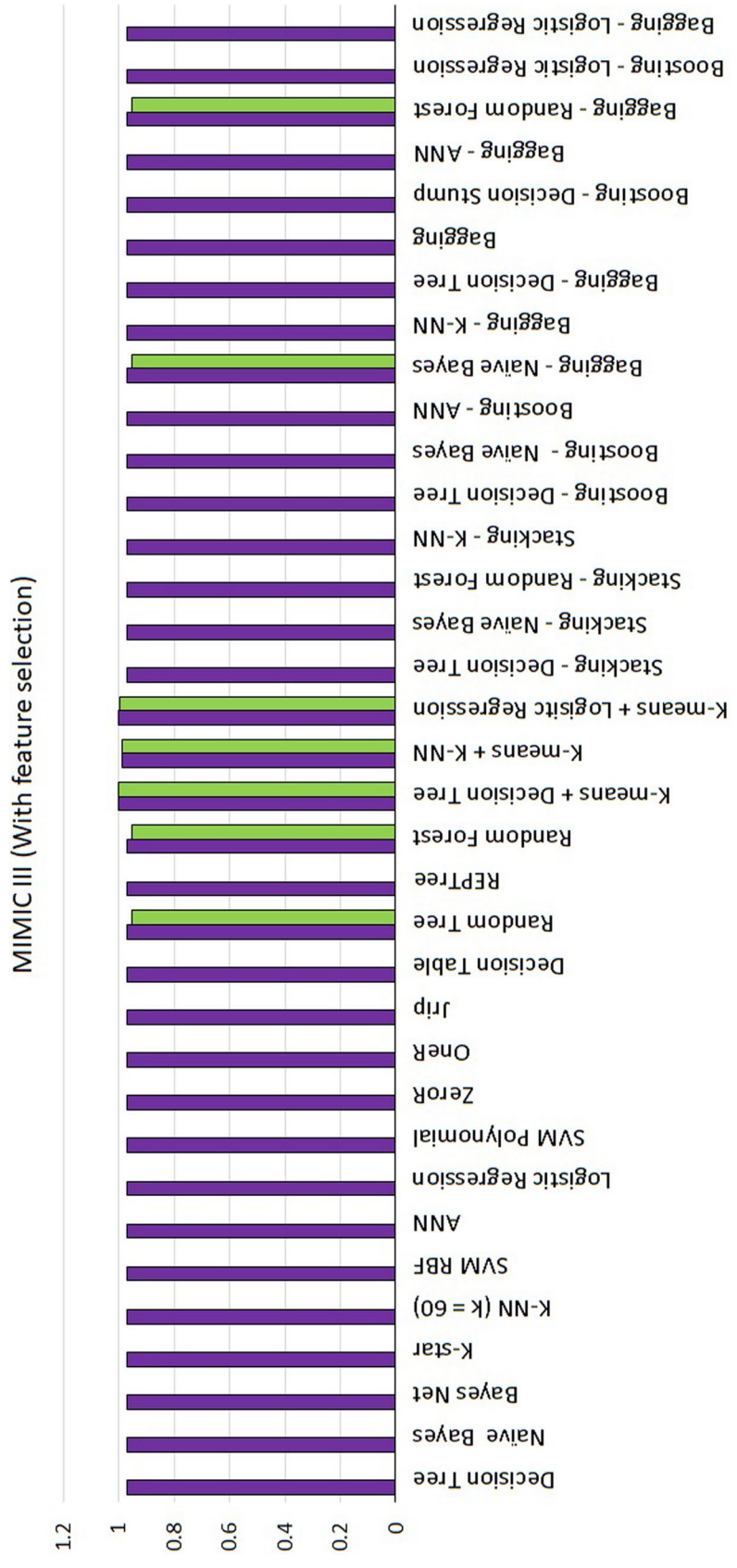

Fig. 9 Accuracy and F-measure of the classification algorithms for the MIMIC III Dataset with feature selection

Table 5 Execution time for model building and validation with feature selection

Algorithms	The execution time of model building and validation (min)		
	PIMA Indian	UCI	MIMIC III
DT	0	0.216	0.016
NB	0	0.016	0
BN	0	0.033	0
K star	0.016	74.667	1.016
K-NN	0	4.65	1.433
SVM-RBF	0.016	574.616	12.51
ANN	0.033	1449.85	46.6
LR	0	0.116	0.234
SVM-polynomial	0	37.216	0.584
ZeroR	0	0	0
OneR	0	0.016	0
JRip	0	0.067	0
DTable	0	0.25	0.05
RT	0	0.034	0.016
REPTree	0	0.05	0.016
RF	0.016	2.116	0.6
K-means + DT	0	0.016	0
K-means + K-NN	0	1.3	0.15
K-means + LR	0	0.067	0.05
Stacking-DT	0	0.033	0
Stacking-NB	0	0.033	0.016
Stacking-RF	0	0.45	0.234
Stacking-K-NN	0	5.567	2.134
Boosting-DT	0	1.167	0.216
Boosting-NB	0	0.483	0.066
Boosting-ANN	0.2	48.05	107.633
Boosting-DS	0	0.1833	0.066
Boosting-LR	0	0.6	0.1216
Bagging-NB	0	0.133	0.033
Bagging-K-NN	0	53.416	14.016
Bagging-DT	0	1.016	0.116
Bagging-REPTree	0	0.55	0.134
Bagging-ANN	0.316	1541.616	470.566
Bagging-RF	0.167	20.766	7.133
Bagging-LR	0	0.983	3.367

When using a classification algorithm for the prediction of type 2 diabetes, the following requirements should be considered.

1. *Accuracy vs F-measure*: Most of the algorithms give a high classification accuracy. However, evaluating the classifier performance using only the accuracy can be misleading. This is because, in the case of an imbalanced dataset, which is very frequent in the health domain, the algorithm might have high accuracy but will not be able to classify the minority class labels as revealed by the F-measure. In such a situation, the prediction results can lead to a life-threatening situation, as a diabetic patient can be classified as non-diabetic. Consequently, we recommend the data scientist include F-measure as one of the evaluation metrics.
2. *Data-driven*: Our experimental results reveal that the performance of the classification algorithms is data-driven. The random forest algorithm is best suitable for prediction in the case of an imbalanced dataset. This is because the algorithm takes a sample of minority class and a similar size sample from the positive class with replacement to form the training dataset for each decision tree constructed.
3. *Feature selection*: We recommend the data scientist and clinicians execute a feature selection algorithm on the dataset before training the classification model. This reduces the execution time, avoids data overfitting as well as will give insights on the most significant features for future consideration without accuracy degradation.
4. *Significant features*: We recommend the clinicians use age (demographic category), ethnicity and family history of diabetes (hereditary category), hypertension, obesity, and cardiovascular disease (medical conditions category), and cholesterol (lifestyle category) for the prediction of type 2 diabetes based on our experimental results. This is also confirmed by the statistical and clinical studies in the literature [79, 80].

Table 6 Evaluated works on classification algorithms

References	Algorithms compared	Dataset	Evaluation metrics
[14]	SVM	National Health and Nutrition Examination Survey [76]	Specificity, Sensitivity, Positive predicted and negative predicted values, Receiver operating curve (ROC)
[15]	SVM, Bagging-REPTree, Boosting-DS, RF	National Inpatient Sample data [77]	ROC, Area Under Curve (AUC)
[26]	DT, SVM, K-NN, ANN, LR	PIMA Indian	Execution time, Accuracy, Recall, Error rate
[37]	DT, Kmeans-DT		True positive and False positive rates, Precision, Recall, F-measure
[40]	K-NN, Kmeans-(K-NN)		Accuracy, Specificity, Sensitivity
[41]	Boosting-DT, Bagging-DT, Stacking-DT		Accuracy, Specificity, Sensitivity, F-measure
[42]	NB, ANN, RT, REPTree, RF, DT		Accuracy, Error
[43]	NB, LR, ANN, SVM, K star, Boosting-DS, Bagging-REPTree, OneR, ZeroR, RT, DT, K-NN, JRip		Accuracy, Error Rate, Execution time
[44]	DT, RT, NB, BN		Accuracy, Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE)
[45]	DT, BN, K-NN, ANN, NB		Accuracy
[16]	NB, K-NN, SVM, DT		Accuracy, Error rate
[17]	Kmeans-LR		Accuracy, F-measure, Kappa statistics, ROC
[18]	NB, ANN, DT, ZeroR, RF, LR		Accuracy, Error rate
[19]	REPTree, K star, OneR, ZeroR, SVM, BN		Accuracy, Precision, Recall, F-measure, ROC
[20]	KNN, DT, NB, SVM, LR, RF		Accuracy, F-measure, Recall, ROC-AUC, Misclassification rate
[21]	DT, K-NN, RF, SVM	UCI Diabetes Data	Accuracy, Sensitivity, Specificity
[22]	NB, RF		Accuracy, F-measure
[23]	ANN, LR	Real data	Specificity, Sensitivity, Positive and negative predicted values
[24]	NB, OneR, ZeroR		Accuracy, F-measure, ROC, Kappa statistics, MAE, RAE, RMSE, RRSE
[25]	DT, ANN, LR, NB, RF, Bagging-DT, Bagging-ANN, Bagging-LR, Bagging-NB, Boosting-DT, Boosting-ANN, Boosting-LR, Boosting-NB		Accuracy, ROC
[27]	NB, DT		Specificity, Sensitivity, Coverage, High-risk
[28]	SVM, ANN, DT, K-NN, BN		Accuracy, Specificity, Sensitivity
[29]	RT, NB, DT, LR		Accuracy, Error rate
[30]	NB, SVM, DT		Accuracy, F-measure
[31]	NB, ANN, K-NN		Accuracy, Kappa statistics, MAE, RAE, RMSE, RRSE, Coverage, Execution time
[32]	LR, K-NN, ANN		Accuracy, Specificity, Sensitivity
[33]	RF, LR, Boosting-DS, DT		Specificity, ROC
[34]	DT, Boosting-DT, Bagging–DT	Canadian Primary Care Sentinel Surveillance Network [78]	AUROC

Table 6 (continued)

References	Algorithms compared	Dataset	Evaluation metrics
[35]	DT, ANN	–	Accuracy
[36]	SVM, LR, ANN		
[38]	RF, DT, ANN	PIMA Indian and Real data	Accuracy, Specificity, Sensitivity
[39]	DT, KNN, NB, RF, Stacking, Bagging-DT, Bagging-KNN, Bagging-RF	PIMA Indian and UCI	Accuracy

Funding This work was supported by the National Water and Energy Center of the United Arab Emirates University under Grant 31R215.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Types of diabetes. https://www.idf.org/aboutdiabetes/what-is-diabetes.html. Accessed 23 Mar 2021
2. International Diabetes Federation—facts and figures. https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html. Accessed 23 Mar 2021
3. Ismail L, Materwala H, Al Kaabi J (2020) Association of risk factors with type 2 diabetes: a systematic review. Comput Struct Biotechnol J. https://doi.org/10.1016/j.csbj.2021.03.003
4. National Institue of Diabetes and Digestive And Kidney Diseases Risk Factors for Type 2 Diabetes | NIDDK. https://www.niddk.nih.gov/health-information/diabetes/overview/risk-factors-type-2-diabetes. Accessed 23 Mar 2021
5. Diabetes UK The risk factors of Type 2 diabetes | Diabetes UK. https://www.diabetes.org.uk/preventing-type-2-diabetes/diabetes-risk-factors. Accessed 23 Mar 2021
6. American Diabetes Association. Complications of type 2 diabetes. https://www.diabetes.org/diabetes/complications. Accessed 23 Mar 2021
7. Licitra L, Trama A, Hosni H (2017) Benefits and risks of machine learning decision support systems. JAMA J Am Med Assoc 318:2354–2354. https://doi.org/10.1001/jama.2017.16627
8. Gulshan V, Peng L, Coram M et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA J Am Med Assoc 316:2402–2410. https://doi.org/10.1001/jama.2016.17216
9. Bejnordi BE, Veta M, Van Diest PJ et al (2017) Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA J Am Med Assoc 318:2199–2210. https://doi.org/10.1001/jama.2017.14585
10. Hyland SL, Faltys M, Huser M et al (2020) Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 26:364–373. https://doi.org/10.1038/s41591-020-0789-4
11. De Silva K, Jönsson D, Demmer RT (2020) A combined strategy of feature selection and machine learning to identify predictors of prediabetes. J Am Med Inf Assoc 27:396–406. https://doi.org/10.1093/jamia/ocz204
12. Coombes CE, Abrams ZB, Li S et al (2020) Unsupervised machine learning and prognostic factors of survival in chronic lymphocytic leukemia. J Am Med Informatics Assoc 27:1019–1027. https://doi.org/10.1093/jamia/ocaa060

13. Leila I, Materwala HP, Karduck A, Adem A (2020) Requirements of health data management systems for biomedical care and research: scoping review. J Med Internet Res. https://doi.org/10.2196/17508
14. Yu W, Liu T, Valdez R et al (2010) Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inf Decis Mak. https://doi.org/10.1186/1472-6947-10-16
15. Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inf Decis Mak. https://doi.org/10.1186/1472-6947-11-51
16. Patel PB, Shah PP, Patel HD (2017) Analyze data mining algorithms for prediction of diabetes. Comput Eng 5:466–473
17. Wu H, Yang S, Huang Z et al (2018) Type 2 diabetes mellitus prediction model based on data mining. Inf Med Unlocked 10:100–107. https://doi.org/10.1016/j.imu.2017.12.006
18. Hina S, Shaikh A, Sattar SA (2017) Analyzing diabetes datasets using data mining. J Basic Appl Sci 13:466–471
19. Larabi-Marie-Sainte S, Aburahmah L, Almohaini R, Saba T (2019) Current techniques for diabetes prediction: review and case study. Appl Sci. https://doi.org/10.3390/app9214604
20. Jakka A, Rani JV (2019) Performance evaluation of machine learning models for diabetes prediction. Int J Innov Technol Explor Eng 8:1976–1980. https://doi.org/10.35940/ijitee.K2155.0981119
21. Kandhasamy JP, Balamurali S (2015) Performance analysis of classifier models to predict diabetes mellitus. Proc Comput Sci 47:45–51. https://doi.org/10.1016/j.procs.2015.03.182
22. Tamilvanan B, Bhaskaran VM (2017) An experimental study of diabetes disease prediction system using classification techniques. IOSR J Comput Eng 19:39–44. https://doi.org/10.9790/0661-1901043944
23. Wang C, Li L, Wang L et al (2013) Evaluating the risk of type 2 diabetes mellitus using artificial neural network: An effective classification approach. Diabetes Res Clin Pract 100:111–118. https://doi.org/10.1016/j.diabres.2013.01.023
24. Mounika M, Suganya SD, Vijayashanthi B, Anand SK (2015) Predictive analysis of diabetic treatment using classification algorithm. Int J Comput Sci Inf Technol 6:2502–2502
25. Nai-arun N, Moungmai R (2015) Comparison of classifiers for the risk of diabetes prediction. Proc Comput Sci 69:132–142. https://doi.org/10.1016/j.procs.2015.10.014
26. Karthikeyani V, Begum I, Tajudin K, Begam I (2012) Comparative of data mining classification algorithm (CDMCA) in diabetes disease prediction. Int J Comput Appl 60:26–31. https://doi.org/10.5120/9745-4307
27. Songthung P, Sripanidkulchai K (2016) Improving type 2 diabetes mellitus risk prediction using classification. In: International joint conference on computer science and software engineering (JCSSE), pp 1–6
28. Heydari M, Teimouri M, Heshmati Z, Alavinia SM (2016) Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran. Int J Diabetes Dev Ctries 36:167–173. https://doi.org/10.1007/s13410-015-0374-4
29. Kumar PS, Umatejaswi V (2017) Diagnosing diabetes using data mining techniques. Int J Sci Res Publ 7:705–709
30. Nithyapriya T, Dhinakaran S (2017) Analysis of various data mining classification techniques to predict diabetes mellitus. Int J Eng Dev Res 5:695–703
31. Sisodia D, Sisodia DS (2018) Prediction of diabetes using classification algorithms. Proc Comput Sci 132:1578–1585. https://doi.org/10.1016/j.procs.2018.05.122
32. Selvakumar S, Kannan KS, GothaiNachiyar S (2017) Prediction of diabetes diagnosis using classification based data mining techniques. Int J Stat Syst 12:183–188
33. Lai H, Huang H, Keshavjee K et al (2019) Predictive models for diabetes mellitus using machine learning techniques. BMC Endocr Disord 1:1–9. https://doi.org/10.1186/s12902-019-0436-6
34. Perveen S, Shahbaz M, Gurgachi A, Keshavjee K (2016) Performance analysis of data mining classification techniques to predict diabetes. Proc Comput Sci 82:115–121. https://doi.org/10.1016/j.procs.2016.04.016
35. Peter S (2014) An analytical study on early diagnosis and classification of diabetes mellitus. Bonfring Int J Data Min 4:07–11. https://doi.org/10.9756/BIJDM.10310
36. Komi M, Li J, Zhai Y, Zhang X (2017) Application of data mining methods in diabetes prediction. In: International conference on image, vision and computing (ICIVC), pp 1006–1010
37. Karegowda AG, Jayaram M, Manjunath A (2012) Rule based classification for diabetic patients using cascaded K-means and decision tree C4.5. Int J Comput Appl. https://doi.org/10.5120/6836-9460
38. Zou Q, Qu K, Luo Y et al (2018) Predicting diabetes mellitus with machine learning techniques. Front Genet. https://doi.org/10.3389/fgene.2018.00515
39. Alehegn M, Joshi RR, Mulay P (2019) Diabetes analysis and prediction using random forest KNN Naïve Bayes and J48: an ensemble approach. Int J Sci Technol Res 8:1346–1354
40. NirmalaDevi M, alias Balamurugan SA, Swathi UV (2013) An amalgam KNN to predict diabetes mellitus. In: IEEE international conference on emerging trends in computing, communication and nanotechnology (ICECCN)
41. Bashir S, Qamar U, Khan FH, Javed MY (2014) An efficient rule-based classification of diabetes using ID3, C4.5 & CART ensembles. In: 12th international conference on frontiers of information technology, pp 226–231
42. Kaur G, Chhabra A (2014) Improved J48 classification algorithm for the prediction of diabetes. Int J Comput Appl 98:13–17. https://doi.org/10.5120/17314-7433
43. Ahmed K, Jesmin T (2014) Comparative analysis of data mining classification algorithms in type-2 diabetes prediction data using WEKA approach. Int J Sci Eng 7:155–160. https://doi.org/10.12777/ijse.7.2.155-160
44. Srikanth P, Deverapalli D (2016) A critical study of classification algorithms using diabetes diagnosis. In: 2016 IEEE 6th international conference on advanced computing (IACC), pp 245–249
45. Devi MR, Shyla JM (2016) Analysis of various data mining techniques to predict diabetes mellitus. Int J Appl Eng Res 11:727–730
46. EMC Education Services (2015) Data science and big data analytics: discovering, analyzing, visualizing and presenting data. Wiley, New York
47. Oliver JJ, Hand D (1994) Averaging over decision stumps. In: European conference on machine learning, pp 231–241
48. Muralidharan V, Sugumaran V (2012) A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of monoblock centrifugal pump using wavelet analysis. Appl Soft Comput 12:2023–2029. https://doi.org/10.1016/j.asoc.2012.03.021
49. Aha DW, Kibler D, Albert MK (1991) Instance-based learning algorithms. Mach Learn 6:37–66. https://doi.org/10.1007/BF00153759
50. Cleary JG, Trigg LE (1995) K*: An instance-based learner using an entropic distance measure. Mach Learn Proc 1995:108–114
51. Homser Jr DW, Lemeshow S, Sturdivant RX (2013) Applied logistic regression
52. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20:273–297. https://doi.org/10.1007/BF00994018
53. Hassoun MH (1995) Fundamentals of artificial neural networks. MIT Press

54. Hall M, Frank E, Holmes G et al (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 11:10–18. https://doi.org/10.1145/1656274.1656278
55. Holte RC (1993) Very simple classification rules perform well on most commonly used datasets. Mach Learn 11:63–90. https://doi.org/10.1023/A:1022631118932
56. Cohen WW (1995) Fast effective rule induction. In: Machine learning proceedings. Elsevier, pp 115–123
57. Kohavi R (1995) The power of decision tables. In: European conference on machine learning, pp 174–189
58. Pfahringer B (2010) Random model trees: an effective and scalable regression method
59. Liaw A, Wiener M (2002) Classification and regression by randomforest. R news 2:18–22
60. Quinlan JR (1987) Simplifying decision trees. Int J Man Mach Stud 27:221–234. https://doi.org/10.1016/S0020-7373(87)80053-6
61. Alsabti K, Ranka S, Singh V (1997) An efficient K-means clustering algorithm
62. Breiman L (1996) Bagging predictors. Mach Learn 24:123–140. https://doi.org/10.1007/BF00058655
63. Freund Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Thirteenth international conference on machine learning, pp 148–156
64. Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259. https://doi.org/10.1016/S0893-6080(05)80023-1
65. Dehghan A, Van Hoek M, Sijbrands EJG et al (2008) High serum uric acid as a novel risk factor for type 2 diabetes. Diabetes Care 31:361–362. https://doi.org/10.2337/dc07-1276
66. Hypertension and Obesity. https://www.obesityaction.org/community/article-library/hypertension-and-obesity-how-weight-loss-affects-hypertension/. Accessed 23 Mar 2021
67. Cardiovascular (Heart) Diseases. https://www.webmd.com/heart-disease/guide/diseases-cardiovascular#1. Accessed 23 Mar 2021
68. Smith JW, Everhart J, Dickson W, et al (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the annual symposium on computer application in medical care, pp 261–265
69. Strack B, Deshazo JP, Gennings C et al (2014) Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed Res Int 2014:11. https://doi.org/10.1155/2014/781670
70. Johnson AEW, Pollard TJ, Shen L et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data. https://doi.org/10.1038/sdata.2016.35
71. Hall MA (1998) Correlation-based feature subset selection for machine learning
72. Hall MA (1999) Feature selection for discrete and numeric class machine learning
73. Feature Selection Algorithms. https://dataminingntua.files.wordpress.com/2008/04/weka-select-attributes.pdf. Accessed 23 Mar 2021
74. Kuhn M, Johnson K (2013) Applied predictive modeling. Springer, New York
75. Fushiki T (2011) Estimation of prediction error by using K-fold cross-validation. Stat Comput 21:137–146. https://doi.org/10.1007/s11222-009-9153-8
76. NHANES - National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/index.htm. Accessed 23 Mar 2021
77. HCUP National (Nationwide) Inpatient Sample (NIS). https://healthdata.gov/dataset/hcup-national-nationwide-inpatient-sample-nis-restricted-access-file. Accessed 23 Mar 2021
78. Canadian Primary Care Sentinel Surveillance Network (CPCSSN). https://cpcssn.ca/. Accessed 23 Mar 2021
79. Zhang N, Yang X, Zhu X et al (2017) Type 2 diabetes mellitus unawareness, prevalence, trends and risk factors: National Health and Nutrition Examination Survey (NHANES) 1999–2010. J Int Med Res 45:594–609. https://doi.org/10.1177/0300060517693178
80. Perry IJ, Wannamethee SG, Walker MK et al (1995) Prospective study of risk factors for development of non-insulin dependent diabetes in middle aged British men. BMJ 310:560–564. https://doi.org/10.1136/bmj.310.6979.560