



# Hybrid stacked ensemble combined with genetic algorithms for diabetes prediction

Jafar Abdollahi<sup>1</sup> · Babak Nouri-Moghaddam<sup>1</sup>

Received: 21 April 2021 / Accepted: 27 February 2022 / Published online: 21 March 2022  
© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2022

## Abstract

Diabetes is currently one of the most common, dangerous, and costly diseases globally caused by increased blood sugar or a decrease in insulin in the body. Diabetes can have detrimental effects on people's health if diagnosed late. Today, diabetes has become one of the challenges for health and government officials. Prevention is a priority, and taking care of people's health without compromising their comfort is an essential need. In this study, the ensemble training methodology based on genetic algorithms was used to diagnose and predict the outcomes of diabetes mellitus accurately. This study uses the experimental data, actual data on Indian diabetics on the University of California website. Current developments in ICT, such as the Internet of Things, machine learning, and data mining, allow us to provide health strategies with more intelligent capabilities to accurately predict the outcomes of the disease in daily life and the hospital and prevent the progression of this disease and its many complications. The results show the high performance of the proposed method in diagnosing the disease, which has reached 98.8%, and 99% accuracy in this study.

**Keywords** Diabetes · Stacked generalization · Prediction · Internet of Things · Ensemble learning

## 1 Introduction

Chronic diabetes (CD) is one disease that affects the body's metabolism and causes structural changes. In 2014, the number of patients increased from 100 to 422 million [1–3]. Diabetes is typically divided into type 1, type 2 [4], and gestational diabetes. Type 2 is increasing with a high prevalence worldwide and is one of the leading causes of death. Because regardless of age and gender, it threatens them due to the lack of insulin in the body [5]. Increased blood sugar is associated with the risk of death in the community due to pneumonia, stroke, acute myocardial infarction, etc.

However, its effect is on the vital organs is so harmful that it is considered the mother of all diseases. There is a risk of miscarriage, kidney failure, heart attack, blindness, and other chronic and deadly diseases in diabetic patients. Therefore, it is essential to diagnose diabetes faster [6–8].

The promising emerging potential of the Internet of Things (IoT) for connected medical devices and sensors plays a vital role in the next-generation healthcare industry for quality patient care. Due to the increasing number of elderly and disabled people, there is an urgent need for real-time health care infrastructure to analyze patient health care data to prevent preventable deaths [9]. Also, in intelligent health, modern wearable devices have gradually increased their capabilities in recent decades. They are equipped with several internal and external sensors to detect many vital signs [10].

Designing and implementing a far-off monitors system allows physicians and caregivers to remember peopling health in the least time. Current developments in ICT like the net of Things, Machine learning (ML), and data processing will enable us to produce health strategies with more intelligent capabilities to accurately predict the outcomes of the disease in the way of life, and therefore the hospital.

Besides, medical advances in recent decades have significantly increased life expectancy while significantly reducing mortality [11]. According to La et al.' study [12], the benefits of personal health care with IoT requirements are divided

✉ Babak Nouri-Moghaddam  
babaknouriit85@gmail.com

<sup>1</sup> Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran



Fig. 1 General IoT-based health monitoring system [14, 15]

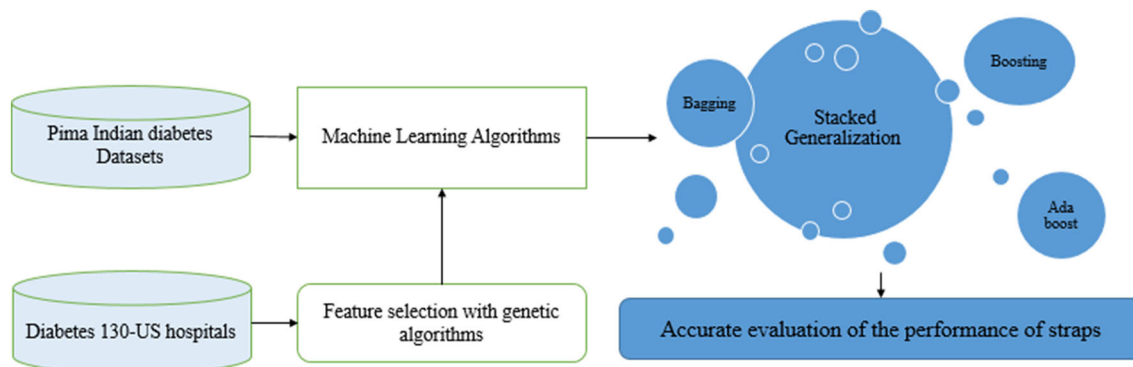


Fig. 2 Schematic of the model proposed framework

into three general categories, such as. First, increase the likelihood of early detection of potential and ongoing diseases without visiting clinics.

Promote frequent assessments of health conditions and awareness of preventive health care needs. For example, the diagnosis of the disease with a set of measurements for a period can have been effectively accessible through routine examinations in clinics. Because routinely, in clinics and hospitals, doctors first take vital signs, then perform medical tests on the accepted vital signs to diagnose the disease.

Rahmani et al. [13] used IoT technology to offer an objective and structured approach to improving human health. This approach will vary the health sector's IoT-based devices regarding social benefits, influence, and cost-effectiveness. Due to the character of the IoT calculations, all health institutions (people, equipment, medicine) are often continuously monitored and managed. Therefore, using these technologies within the health industry can improve the standard and costs of medical aid by automating tasks already carried out by humans. Figure 1 shows the overall IoT-based health monitoring system [14, 15].

As a result, the prediction accuracy of a model may be high even with optimal parameters. Not surprisingly, the models produced by MLR and carrier support regression with a linear core are not statistically distinct and

perform significantly better than other methods in the IWPC Ensemble [16]. One way to beat the restrictions of one algorithm is to mix the benefits of several algorithms to exceed the limit of one machine learning algorithm (e.g., the Ensemble method). Recently, the “bagging” Ensemble method has been went to predict diabetes [17, 18].

Stack generalization is another Ensemble method that uses a higher-level model to combine lower-level models to achieve higher prediction accuracy [19, 20]. Unlike bagging and boosting approaches, which can only combine machine learning algorithms of the same type, stacked generalization can combine different algorithms through a meta-machine learning model to maximize generalization accuracy. The objectives of the study or the research question are as follows.

- Using a hybrid machine learning model to diagnose diabetes.
- Significant improvement in a forecast accuracy
- Use several models in combination
- Achieve a high level of reliability in classification.
- Use multiple models to increase the estimate of the final model.
- Improved accuracy and reduced error compared to single-core models.

Also, the contributions of this paper can be summarized as follows:

- Machine learning ensemble models for Diabetes (T2D) prediction demonstrated high performance.
- Comparing results with the most related researches according to the literature.
- Examining the benefits of ensemble methods proposed recently for prediction.
- Using hybrid stacked generalization (SG) based Meta-heuristics approach in the diagnosis of diabetes.

In the last several years, the number of persons diagnosed with diabetes has increased exponentially. According to a health report, 347 million individuals globally have diabetes. Diabetes is a disease that affects both the elderly and the younger generation [21–23]. Diagnosing diabetes in its early stages is also difficult. This diagnosis will aid in the decision-making process of the medical system and will help us save lives from diabetes. Therefore, early prediction of diabetes is significant to save a person from diabetes.

The type 2 diabetes data set, which contains nine valuable factors and 768 records, and another dataset with 55 useful variables and 100,000 records, was used in this investigation. Tables 2 and 3 lists the variables and their abbreviations. Finding a pattern in a vast data set is all about data analysis. This allows us to draw particular conclusions from the data supplied. Different machine learning methods can be used to perform the analysis. However, investigations have demonstrated that none of the algorithms can adequately solve a problem independently. This paper presents two sets of machine learning algorithms for diabetes prediction. The classification-based algorithm is one, and the ensemble learning algorithm is the other.

Artificial intelligence (AI) research in healthcare is quickly advancing, with possible applications being proven in a variety of medical fields. We employ ensemble learning deliberately to look for superior prediction performance or classification accuracy [24–27]. We use nine classifiers in classification. Random forest classifier, Support Vector Machine, Decision Trees, K-Nearest Neighbors, Gradient Boosting, Multilayer Perception, Extra Tree Classifier, AdaBoost classifier, and basic Gaussian bays are the several types of classification algorithms. We used a learning-based Genetic Algorithm (GA) for the ensemble learning approach. These ten algorithms were applied and compared to evaluate the accuracy of diabetes prediction for two different approaches to machine learning, and they scored 98% on average, which is higher than previous machine learning algorithms.

*The rest of the article is as follows:* we will review the method in section two and examine the results in section three. Then, we will review the discussion in the fourth

section. Finally, in the last section, we will review the conclusions and future work.

## 2 Related work

The present study deals with an ensemble stacking-based learning methodology for detecting diabetes. This section provides a brief review of the literature on numerous meta-heuristic optimization-based and ensemble learning-based prediction methods of diabetes. Measures have been taken to reduce the number of chronic disease diagnostic tests to reduce overall costs. One of the possible solutions is to use machine learning techniques in healthcare data, which are used to find frequent patterns in an extensive database to obtain helpful information.

Machine learning methods are instrumental in diagnosing diabetes and increasing its efficiency. One of the most important challenges for machine learning researches is accurately diagnosing diabetes. For example, Fatima et al. [28] have studied different machine learning methods to diagnose various diseases such as cardiopathy, diabetes, liver, and hepatitis and have succeeded in diagnosing this disease. Besides, Alić et al. [29] have used artificial neural networks and Bayesian networks to diagnose and classify diabetes, aiming to evaluate artificial neural networks and Bayesian networks and their application to classify Type 2 diabetes (T2D) cardiovascular disease. Kumar et al. [30] used three distinct data mining arrangement methods, e.g., Simple Biz (NB), support vector machine (SVM), and decision tree decision-making approach to potential approaches to predict the likelihood of heart disease for people with diabetes. However, the accuracy of their prediction is dependent on the accuracy of their prediction. Medical bioinformatics analyses were used by Saru et al. to predict diabetes [31].

Subramaniyan et al. [32] was predictive analytics using machine learning to evaluate massive data to anticipate future difficulties in diabetic patients. Yang et al. [33] also created a computer method that combined multiple forms of physical examination data to predict the risk of diabetes. Today, ML offered various tools for efficient data analysis. Especially in the last few years, the digital revolution has provided affordable and accessible tools for collecting and saving data. Data collection and examination machines are located in new and modern hospitals to collect and share data in large information systems. ML technology is very effective for analyzing medical data and has an influential role in solving diagnostic problems. Correct diagnostic data is presented as medical history or report in modern hospitals or their specific information department. These techniques are accustomed to classifying the information set [28, 34].

To evaluate the early diagnosis of diabetes, Zeki et al. [35] used three DM methods: Nave Bayes (NB), logistic

regression (LR), and Random Forest (RF). The findings of the RF test, according to their research, showed that it has the best level of accuracy when compared to other procedures. In addition, according to the Kalyankar et al. study [36], the machine learning technique in the Hadoop MapReduce environment was used to detect missing values and discover trends in the Pima Indian Diabetes dataset. According to the patient's risk level, this research can forecast the forms of diabetes mellitus, associated future hazards, and the type of therapy supplied.

Predictive algorithms based on data mining for evaluating diabetes data can aid in the early diagnosis and prediction of the condition and important link events such as hypo/hyperglycemia. Various approaches have been developed to diagnose, forecast, and classify diabetes [14].

The Internet of Things (IoT) is increasingly being utilized to implement various applications, particularly as the amount of data available grows. The Internet of Things can be used in various applications, including patient monitoring systems. For example, we can use the Internet of Things to evaluate data and offer it to physicians and paramedics in the health industry. We can identify solutions for longer and healthier lives by analyzing, processing, and exploiting the knowledge and information contained in big data on health issues and illness trends in a specific population. In many ways, extensive data analysis improves healthcare insight [37].

## 2.1 Analysis of strengths and weaknesses related work

Each year, several expenditures are associated with treating and diagnosing patients with diabetes. Diagnostic techniques that have been used in the past are time-consuming. As a result, the most significant and urgent concern is precise forecasting and dependable procedures [38]. Machine learning techniques used to healthcare data could be one option. Diabetes can be diagnosed and managed more efficiently using machine learning approaches. Researchers have investigated several machine learning methods for identifying diseases such as cardiopathy, diabetes, liver, and hepatitis and have been successful in doing so.

Many research suggests that artificial neural networks, random forest networks, and Bayesian networks are the most accurate ways for diagnosing and classifying diabetes when compared to other techniques. However, we offer a paradigm that incorporates artificial neural networks, Bayesian networks, and seven different methods for categorization. Ensemble Learner is the name given to this combination strategy. We proposed our approach for more accurate disease prediction compared to prior models. The experimental research findings also demonstrated that our strategy outperforms artificial neural networks, Bayesian networks, and random forests. Table 1 compares the benefits and

drawbacks of the proposed approach for diagnosing diabetes to previous methods.

## 2.2 Supported methodologies

This section discusses all the supported methodologies used in this study. Machine learning techniques have been widely used in many scientific fields. However, this use in the medical literature is limited partly because of technical difficulties [49, 50].

- *Decision tree*: a decision tree is a decision support tool that uses trees to model [51, 52].
- *Naive Bayes Classifier*: an ensemble of simple classifiers based on probabilities is based on Bayes' theorem, assuming the independence of random variables [53, 54].
- *Artificial Neural Network*: inspired by how the biological nervous system processes data and information for learning and knowledge creation [55, 56].
- *Support Vector Machine*: it is one of the supervised learning methods used for classification and regression [56, 57].
- *C 4.5*: Algorithm C 4.5 is one of the decision tree algorithms, which is very important due to its very high interoperability [58].
- *Random Forest*: one combination learning method for categorization is regression, which works supported training time and, therefore, the output of classes (classification) or for the predictions of every tree individually, supported a structure consisting of the many decision trees [59, 60].
- *K-Nearest Neighbors*: KNN classifier is to classify unlabeled observations by assigning them to the class of the most similar labeled examples. Characteristics of statements are collected for both training and testing datasets. The suitable choice of k features has a significant impact on the diagnostic performance of the KNN algorithm. An oversized k reduces the effects of variance caused by random error but runs the danger of ignoring small but significant patterns [59, 61, 62].

The following figure shows the proposed flowchart [62]. Within the following, we will examine the popular methods of mixing categories.

- *Bagging*: one of the most straightforward and successful combined approaches to improving the classification problem is the Bagging algorithm, commonly used for decision trees. This algorithm is beneficial for bulk data and will work well for unstable learning algorithms, that is, algorithms that change due to changing data.
- *Boosting*: this algorithm aims to combine several weak classifiers and obtain a strong one to improve performance, in which the predictors are trained continuously.

**Table 1** Advantages and disadvantages of the proposed method in diagnosing diabetes compared to other methods

Authors	Approach	Advantages	Disadvantages
Niloy Sikder et al. [39]	Ensemble learning	A valuable tool for mass retinal screening to detect DR	Imbalanced dataset
Baha Ihnaini et al. [40]	Deep ensemble learning	For multidisciplinary diabetic illness prediction and recommendation, the proposed method is superior	–
Mrinal Banchhor et al. [41]	Ensemble Learning(Random Forest)	In tenfold cross-validation, the Random Forest method was determined to have the best test accuracy, with 99.03% and 96.88% accuracy	Low of data
Md. Mahmudul Hasan Sabbir et al. [42]	ensemble learning	The ensemble learning approach's sensitivity, specificity, and accuracy with texture features are higher than any individual learning model, with 97.2% sensitivity, 78.6% specificity, and 92.0% accuracy	–
Leon Kopitar et al. [43]	Glmnet, RF, XGBoost, LightGBM	LightGBM models had the best amount of variable selection stability over time	Our results showed no clinically relevant benefit when more sophisticated prediction models were deployed
L. J. Muhammad et al. [44]	Random-Forest	With an accuracy of 88.76%, the random forest predictive learning-based model appeared to be one of the best-produced models	Low accuracy
Nour Abdulhadi et al. [45]	Random-Forest	The primary goal of this study is to use several machine learning approaches to predict the occurrence of diabetes, particularly in females, at an early stage	Low accuracy
Huma Naz et al. [46]	ANN, NB, DT DL	DL had the greatest results for diabetes onset on the PIMA dataset, with an accuracy rate of 98.07%	–
Adel Al-Zebari et al. [47]	Decision Trees (DT), Logistic Regressions (LR), Discriminant Analysis (DA), Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and ensemble learners	Performance comparison of machine learning algorithms for detecting diabetes illness is conducted	He had accuracy scores ranging from 65.5 to 77.9% on average. The LR approach produces the best accuracy score of 77.9%, while the Coarse Gaussian SVM methodology produces the worst accuracy score of 65.5%
Vandana Rawat et al. [48]	Bagging and AdaBoost	The results computed are quite accurate, with bagging and AdaBoost approaches achieving classification accuracy of 81.77% and 79.69%, respectively	–



- *Ada Boost*: a meta-algorithm is designed to improve the performance and solve the problem of unbalanced categories, which produces a robust and high-quality learner from a combination of three-week learners. This algorithm combines weak learners to produce an accurate classifier [63–67].

This study provides an intelligent monitoring system for patients and older people with chronic diseases using the collected data for effective diagnosis and prediction in non-critical situations to promote smart health and prevent deaths on IoT infrastructure using the intelligent ensemble learning algorithms. Ensemble methods are learning algorithms that construct an ensemble of classifiers and classify new data points by taking a (weighted) vote of their predictions.

The original ensemble method is Bayesian averaging, but newer algorithms include error-correcting output coding, Bagging, and boosting. This article reviews these methods and explains why ensembles often perform better than any classifier.

### 3 Materials and methods

Since the use of an intelligent machine learning algorithm in diagnosing and predicting diseases has not been successful in many scenarios and considering the content and challenges mentioned in the background section, there are many challenges to smart health in IoT, need to be addressed; one of which is the accurate diagnosis and prediction of disease outcomes.

This paper uses the new machine learning approach called Ensemble Learning to diagnose and predict chronic diseases described below. The purpose of this article is to improve the accuracy and speed of diagnosis of chronic diseases in

the context of the intelligent network by which we want to use ensemble learning approaches and a new meta-learner in stacking learning. Stack generalization is an approach that allows researchers to connect several different prediction algorithms to a combination.

#### 3.1 Data set

In this study, the data set of type 2 diabetes available in (<https://archive.ics.uci.edu/ml/datasets/diabetes>) has been used, which has nine useful variables and 768 records. These variables and abbreviations are listed in Table 2. 70% of the data is for training, and 30% of the information is for testing.

In addition, another dataset has been used to teach algorithms. This data has been prepared to investigate factors associated with readmission yet as other outcomes regarding patients with diabetes. The data set of heart patients is available in (<https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008>) has been used, which has 55 useful variables and 100.000 records. Variables and their abbreviations are listed in Table 3. 80% of the data is for training, and 20% of the information is for testing.

#### 3.2 Data preprocessing

On the analysis of big clinical databases, the Knowledge Discovery in Databases (KDD) methodology appears to be appealing. The preprocessing step (data cleaning and management of missing values) is critical in the KDD process since it determines the quality of the results acquired by data mining processes and takes up roughly 80% of the total project time. Data preprocessing is vital to arrange the diabetes type data and Pima Indians data to accept a machine learning model. Separating the training and testing data sets ensures that the model learns only from training data and tests

**Table 2** Description of the Pima Indian diabetes data set

Pima Indian diabetes datasets					
Dataset	Sample size	Feature size including class label	Classes	Presence of missing attribute	Presence of noisy attributes
Pima Indian diabetes	768	9	2	No	No

**Table 3** Description of the diabetes 130-US hospitals for 1999–2008 data set

Diabetes 130-US hospitals for years 1999–2008 data set					
Dataset	Sample size	Feature size including class label	Classes	Presence of missing attribute	Presence of noisy attributes
Diabetes 130-US hospitals for years 1999–2008 data set	100.000	55	Multivariate	Yes	No

its performance with the testing data. Therefore, the data set used was divided into training and test data. The training data contain 70% of the data set, and the test and validation data include 15% each. At first, all the data was shuffled.

### 3.3 Building and training the stacked-generalization model

This article develops a stacking-based evolutionary ensemble learning system, “Stacked Generalization based Metaheuristics,” to predict the onset of Type-2 diabetes mellitus (T2DM) within five years. Before learning, as a data preprocessing step, the missing values and outliers were identified and imputed with the median values. Several machine learning optimization algorithms are utilized for base learner selection, which simultaneously maximizes the classification accuracy and minimizes the ensemble complexity. As for model combination, Bagging, Boosting, and Ada boost are employed as a meta-classifier that combines the predictions of the base learners [68].

The comparative results demonstrate that the proposed stacking genetic method outperforms several individual ML and conventional ensemble approaches. Figure 2 depicts the learning process with stacked generalization based on the model selection from 9 (Table 5) base learners and three stacking-based combination methods.

Stack generalization is a different technique for combining several different classifiers such as decision tree, artificial neural network, support vector machine, etc., which consists of two stages:

Basic learners at level zero and stacking model learners at level one; at level zero, several different models are used to learn from the dataset, and the output of each model is used to make a new dataset. For example, Fig. 3 shows the Stacking algorithm [69].

### 3.4 Hyperparameter tuning

Algorithms: an overview. Figure 4 shows a flow chart of the algorithms we applied in this study. The stack generalization learning algorithm is shown in Fig. 3. Here, the Pima Indian diabetes and Diabetes 130-US hospitals for years 1999–2008 dataset is considered for testing all the models. The source of this dataset is the UCI repository [70–74]. To determine whether ensemble predictors constructed using stacked generalization improve the prediction accuracy for diabetes, we constructed different stacked generalization frameworks using the same parameters in individual algorithms. Of course, we use Feature selection to select the best feature to improve the accuracy suggested method.

Feature selection reduces the number of attributes while keeping a subset of the original features. Feature selection is frequently used in data preparation to find previously

unknown features useful to classification tasks and remove unnecessary or redundant features. The goal of feature selection is to boost classification accuracy. GA is used to reduce insignificant features in this study. We defined chromosomes as a mask for characteristics to achieve this goal [75].

In this work, GA is used to eliminate insignificant features. In order to reach this purpose, we defined chromosomes as a mask for features. To put it another way, each chromosome is a collection of characteristics. The number of characteristics indicating a diabetes patient’s specification is equal to the size of the chromosome (number of genes). As previously stated, a chromosome is represented by a binary string that

---

**Algorithms 1: Pseudo Code of Ensemble Algorithm**

---

```

1: Input: Data set  $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ ;
2:   First-level learning algorithms  $L_1, \dots, L_T$ ;
3:   Second-level learning algorithm  $L$ .
4: Process:
5:   for  $t = 1, \dots, T$ :
6:      $h_t = L_t(D)$  // Train a first-level individual learner  $h_t$  by applying the first-level learning algorithm  $L_t$  to the original data set  $D$ 
7:   end for
8:    $D' = \emptyset$  // Generate a new data set
9:   for  $i = 1, \dots, m$ :
10:    for  $t = 1, \dots, T$ :
11:       $z_{it} = h_t(x_i)$  // Use  $h_t$  to classify the training example  $x_i$ 
12:    end for
13:     $D' = D' \cup \{(z_{i1}, z_{i2}, \dots, z_{iT}), y_i\}$ 
14:  end for
15:   $h' = L(D')$  // Train a second-level learner  $h'$  by applying the second-level learning algorithm  $L$  to the new data set  $D'$ 
16: Output:  $H(x) = h'(h_1(x), \dots, h_T(x))$ 

```

---

Fig. 3 Stacking algorithm

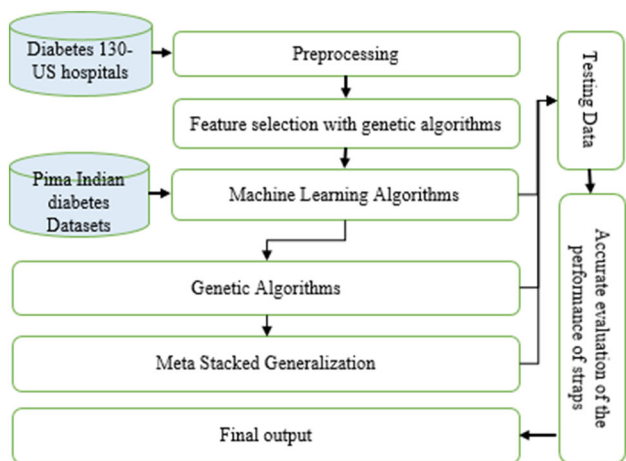


Fig. 4 Proposed of research

Algorithms 2: Algorithm GA	
1	formulate initial population
2	randomly initialize population
3	repeat
3.1	evaluate objective function
3.2	find fitness function
3.3	apply genetic operators
3.3.1	reproduction
3.3.2	crossover
3.3.3	mutation
4	until stopping criteria

**Fig. 5** The working principle of a simple genetic algorithm

**Table 4** Genetic algorithms parameters

Parameter name	Value
Population size	50
Number of generation	100
Probability of crossover	0.6
Probability of mutation	0.033
Type of mutation	Uniform
Type of selection	Rank-Based
Stall generation limit	10
Stall time limit	Infinite
Elite count	2
Fitness function	Linear ranking
Encoding	Binary

is either 0 or 1. A value of 1 indicates that the related feature is selected, whereas 0 indicates that it is not.

### 3.5 Genetic algorithm

GA is one of the initial population-based random algorithms proposed in history. Select, cross, and mutant are the most common GA operators [76]. These algorithms use recombinant operators to store crucial information in simple chromosome-like structures that encode a possible solution to a specific problem. Although the wide variety of genetic algorithms is very wide [77], GAs are frequently regarded as performance optimizers. A simple genetic algorithm's operating concept is depicted in Fig. 5.

#### 3.5.1 Setting the GA parameters

Other parameters exist in GA. Executive parameters such as mutation and elite rates and structural characteristics such as population size must be modified. However, we employed the most frequent values for these factors, which yielded satisfactory results. The significance of the GA parameters utilized in this experiment is shown in Table 4.

The conventional GA method was employed in this experiment. There are four significant steps in the GA method: (1) The features were coded as genomes in binary, with '1' denoting selection and '0' denoting non-selection (phenotypes labeled as '0' denote features that were eliminated, whereas phenotypes labeled as '1' denote features that were chosen). Thus, phenotypes labeled as '0' are referred to as decreased features, whereas phenotypes labeled as '1' are very significant features). Then, based on the concept of phenotypes, each genotype creates a collection of subsets. The proposed approach uses these subsets as training sets.

Then, (1) each chromosome was evaluated using the fitness function, and the best features were picked; (3) the chromosomes were modified using crossover and mutation to form a new generation of the population; and (4) the new generation continued to (2) until halting criteria were satisfied.

A population of 50 people was estimated. This ranking was utilized as a part of a selection strategy in which people who scored in the top half of their fitness levels were chosen to have children. The following population was created using a single point crossover with a ratio of 0.6 and a single point jump with a ratio of 0.033. First, the elitist strategy was set at 2, which meant that the two most minor members of the current generation were included in the next population. Then, the number of generations with the best value was set to twenty for the same fitness. The number of generations was finally set to 100. There were 100 GA executions, each with different initial conditions (some data splitting).

#### 3.5.2 Fitness function

The classifier fit was evaluated using a linear ranking value in this work. A statistical measure of the agreement between expected and actual values is linear ranking. The classifier's performance is determined by how the training set is generated. Cross-validation was done using a repeated sub-random sampling technique due to the small data. Data were randomly partitioned into 70% training sets and 30% validation sets before each GA analysis. Because of the linear ranking value, each stage's performance was recorded. Due to the value of genome fit, the center of the ten linear ranking values was recorded [76, 77].

#### 3.5.3 Crossover (recombination)

Two-parent solutions are employed in crossovers to care for a toddler. The selection (production) procedure then comes to a close, followed by the event of higher persons. To pair two chromosomes, this study uses a single-point crossover. Two chromosomes are cut once, and the slices are swapped between them.



### 3.5.4 Mutation

When the crossing procedure is finished, the strings are inserted to perform the mutation process. Small amounts are rotated from 0 to 1 and 1 to 0 in touch mutations [76, 77].

### 3.6 Machine learning parameters

Hyperparameters are variables whose values influence the learning process and affect the learning algorithm's model parameters. As the prefix 'hyper\_' suggests, they are 'top-level' parameters that regulate the learning process and the model parameters that come from it, as the prefix 'hyper\_' suggests. Before you start training your model, as a machine learning engineer, you choose and establish hyperparameter values that your learning algorithm will employ.

This paper investigates many approaches to establish which factors do not affect model performance, a set of parameters that may have an additional impact on model performance and offers appropriate parameters for diabetes research to show the issues involved with identifiability. The purpose of parameter suggestions is to locate the collection of parameter values that reduce your cost function to the smallest possible value.

For Machine Learning Algorithms, there are six Must-Know Parameters, such as:

1. Support Vector Machines have a C parameter.
2. Support Vector Machines' gamma parameter.
3. Maximum depth for Decision Trees.
4. For Decision Trees, min impurity decrease.
5. Tree count for Random Forest and GBDT.
6. *K*-Nearest Neighbors has *N* neighbors.

We have the trained model parameters at the end of the learning process, which is effectively what we refer to as the model.

*Support Vector Machine (SVM)*: like gradient boosting, the SVM algorithm is viral, very effective, and has many hyperparameters to tweak. The choice of the kernel that will regulate how the input variables are projected is maybe the most significant parameter. This work evaluated the subset of selected genes using an SVM classification model with an RBF kernel. There are numerous to pick from, but linear, polynomial, and RBF are most frequent. The penalty (C), which can take on various values and has a significant impact on the geometry of the resulting areas for each class, is another important parameter.

*Random Forest (RF)*: the number of random features to sample at each split point (max features) is the most critical parameter in Random Forest. In this study, a range of integer values, such as 1 to 20, or 1 to half the number of input features, are tested. However, after careful considera-

tion and testing, we decided to change this number to zero. The number of trees (n estimators) is another key element for the random forest. This should ideally be increased until the model shows no more improvement. However, a log scale from 10 to 1000 could be helpful too. In this paper, Ring was utilized to determine the optimal settings. For this parameter, the best-selected scale is 100.

*K-Nearest Neighbors (KNN)*: the most important hyperparameter for KNN is the number of neighbors (n neighbors). We tested values ranging from 1 to 21 in this research. In this paper, the best selected for this parameter N neighbors = 5 was used. This study also looked at distance measures (metrics) for determining the composition of a neighborhood. For the metric, I used Minkowski after checking.

*Multilayer Perceptron (MLP)*: a variety of hyperparameters, such as the number of hidden neurons, layers, and iterations, must be tuned in order for MLP to work. Grid search is a method for optimizing model hyperparameters. It would be best to give a dictionary of hyperparameters to evaluate in the param grid argument when creating this class. This is a map containing the model parameter's name and an array of values to attempt. Grid-search was utilized in this paper to discover the optimal parameter for this technique. Grid-search is configured as follows.

```
parameter_space = {
    'hidden_layer_sizes': [(10, 30, 10), (20,)],
    'activation': ['tanh', 'relu'],
    'solver': ['sgd', 'adam'],
    'alpha': [0.0001, 0.05],
    'learning_rate': ['constant', 'adaptive'],
}
```

*Decision tree (DTree)*: DTree are a great technique to categorize classes because, unlike Random Forests, they are transparent or white box classifiers, which means we can see the logic behind their classification. The function for determining a split's quality. The criterion "Gini" for Gini impurity and "entropy" for information gain is supported. In this paper, entropy was used as a criterion parameter. The entropy measure is used as the impurity measure, and information gain splits a node to deliver the highest information gain. On the other hand, Gini Impurity examines the divergences between the probability distributions of the target attribute's values and splits a node so that the least amount of impurity is produced. The strategy used to choose splitter "best" the split at each node used.

*AdaBoost*: the number of decision trees employed in the ensemble is an essential hyperparameter for the AdaBoost method. For the model to perform successfully, there must be many trees put to it, often hundreds, if not thousands. The "n estimators" option can specify the number of trees. This parameter was set at 100 in this study.

*Naive Bayes (NB)*: classifiers are scalable, with several parameters proportional to the number of variables (features/predictors) in a learning problem. The parameter set for the Naive Bayes classifier is somewhat narrow. Depending on the implementation, the number of classes may be the only parameter we do not influence over in actuality.

*Extra Trees*: it is simple to use because it only includes a few key hyperparameters and logical rules for tuning them. The number of decision trees in the ensemble, the number of input features to select and consider for each split point randomly, and the minimum number of samples necessary in a node to establish a new split point are the three significant hyperparameters to tune in the algorithm. The number of decision trees utilized in the ensemble is an essential hyperparameter for the Extra Trees technique. This setting is set to auto default in this paper. The Extra Trees algorithm, like Random Forest, is unaffected by the value utilized, despite being a critical hyperparameter to control. It is set via the `max_features` argument and defaults to the square root of the number of input features. In this case, for our test dataset, this would be three features.

*Gradient boosting (GB)*: among data scientists, GB is a very popular prediction model. The following are the parameters used in this algorithm.

- Learning rate: this influences how much each tree affects the final result. This parameter was set to 0.1 in this study.
- $N$  estimators: this is the number of sequential trees modeled. Even though GBM is fairly resilient when dealing with many trees, it can nonetheless overfit at times. As a result, a CV should modify this for a specific learning rate.
- Loss: the loss function that must be minimized in each split. For classification and regression cases, it can have a variety of values. In most cases, the default settings are sufficient. The term 'deviance' for loss refers to deviance (= logistic regression) for classification with probabilistic outputs, employed in this paper.

In Table 5—RF is the Random Forest; KNN— $k$ -Nearest Neighbors algorithm; MLP—multilayer Perceptron; Ada Boost—Ada Boost; D Tree—decision tree algorithm; NB—Naive Bayes; GBC—gradient boosting classifier algorithm; SVM—Support vector machine, Extra Trees—Extremely Randomized Trees Classifier.

### 3.7 Performance measuring attributes

For the study, Jupyter notebook was used for implementation, and Python, the programming language, was used for coding. For the study, Jupyter notebook was used for implementation, and Python, the programming language, was used for coding. Among all models, we selected the model with the most predictive accuracy.

**Table 5** Hyperparameter tuning Machine Learning algorithms

Technique	Hyperparameter
RF	$N$ estimators = 100, criterion = entropy, max depth = 10, max features = $n$ features, $n$ classes = 2
KNN	$N$ neighbors = 5, weights = uniform, algorithm = auto, metric = minkowski
MLP	Hidden layer sizes = (100,), activation = relu, solver = adam, batch size = 100, learning rate = adaptive, max iter = 100
Ada Boost	Base estimator = Decision Tree Classifier(max depth = 1), $n$ estimators = 100, learning rate = 1, algorithm = SAMME.R, $n$ classes = 2,
D Tree	Criterion = entropy, splitter = best, max depth = 3, max features = auto
NB	Priors = $n$ classes, epsilon = float, sigma = $n$ classes
GBC	Loss = deviance, learning rate = 0.1, $n$ estimators = 50, criterion = friedman mse,
SVM	Kernel = RBF, degree = 3, gamma = scale
Extra Trees	$N$ estimators = 50, criterion = gini, max depth = 3, max features = auto

**Table 6** Confusion matrix

Confusion matrix	Classified as:	
	Negative	Positive
Actual class		
Negative	TN	FP
Positive	FN	TP

This article efficiently used cost–benefit analysis (the disruption matrix), ROC curve, and other model selection issues such as accuracy. Performance measurement is used to determine the effectiveness of the classification algorithm so that, in the case of two-dimensional classification problems, one can show the cost of classification with a cost matrix for two types of false positive (FP) and false-negative (FN) errors and two types of classification into the positive true (TN) and negative true (TN) that give different costs and benefits. As shown in Tables 6 and 7 [6, 49, 50, 78].

## 4 Results and discussion

Diabetes is a condition in which blood flow is obstructed throughout the body. The retinal blood vessels may leak in this disorder, resulting in retinal edema [79]. Four learning strategies for extracting patterns from data have been described based on data types: supervised, semi-supervised, unsupervised, and reinforced. Labeled data is challenging to access in machine learning, but unlabeled data is frequently collected and accessed quickly. In most initiatives, however, most of the data is unlabeled, but some are [80]. So, in

**Table 7** Detail descriptions about the performance measures [6, 49, 50, 78]

Performance measure	Description
TP	When the positive samples are classified accurately
TN	When the negative samples are classified accurately
FP	When the negative examples are misclassified
FN	When the positive samples are misclassified
Accuracy	It is the overall classification accuracy percentage resulting from a standard classifier
Sensitivity(Sen)	It determines the proportion of true positive samples in total samples and is called as True Positive Rate (TPR) $S_n = TP/(TP + FN)$
Specificity(Spe)	It identifies the proportion of true negative samples in total samples and is called as False Positive Rate (FPR) $S_p = TN/(TN + FP)$

machine learning and data mining, the primary assumption is that the training and future data have the same distribution and properties [81].

During the past years, medical service providers have always manually examined patients' vital signs and diagnosed and predicted the disease based on patient records and research findings. In this study, intelligent machine-learning algorithms are used to diagnose effectively and accurately predict the outcomes of the disease, in which cases such as age, gender, blood pressure, cholesterol, smoking, etc., are considered in the diagnosis of this disease. Finally, the risk of the disease against the mentioned diseases is determined. Table 8 compares the similar works of others with ours.

Table 8 shows the results. Our model's ability to forecast people with diabetes When compared to the findings of other researchers, it is high, with acceptable accuracy. These models can be integrated into an online computer software to assist doctors in predicting the onset of diabetes in patients and offering required preventive measures.

The success of an ensemble learning system is based on a variety of classifiers that make it up. If all classifiers present the same output, it is impossible to correct a possible error. So, they are more likely to have different errors on different samples. If each classifier presents another mistake, you can reduce the total error after their strategic combination. So, such a set of classifiers must be diverse. This diversity can be achieved in different ways, as shown in Fig. 6 [86].

**Table 8** Comparison of work results with previous works

Years	Work	Approach	Accuracy (%)
2018	Singh, P. P. et al. (2018) [82]	Neural Network Method	88.5%
2018	Bhuvaneswari, G., and Manikandan, G. (2018) [83]	Feature Selection based on Genetic Algorithm	82.3
2017	Enrique V. Carrera et al. [84]	Decision tree and Support Vector Machine	SVM: 80.4%, Decision Tree is 91.0%
2018	Emrana Kabir Hashi et al. [85]	Use Machine Learning Algorithms	C4.5 and KNN were 90.43% and 76.96%
2019	Adel Al-Zebari et al. [47]	DT, LR, DA, SVM, k-NN, and ensemble learners	77.9%
2019	Vandana Rawat et al. [48]	bagging and AdaBoost	81.77% and 79.69%
2020	Leon Kopitar et al. [43]	Glmnet, RF, XGBoost, LightGBM	(0.842), (0.846), (0.859), (0.881)
2020	L. J. Muhammad et al. [44]	random-forest	86.28%
2021	Niloy Sikder et al. [39]	Ensemble Learning	94.20%
2021	Nour Abdulhadi et al. [45]	random-forest	82%
2022	This paper (Suggest Method (ST-GA))	Stacked generalization-genetic algorithm	99

- Use different training datasets is to train classifiers.
- Use different training parameters is for different classifiers.
- Use different classifiers.

Today, due to the lack of knowledge about using different data around us, they are neglected by managers. In contrast, if these seemingly insignificant data are purposefully stored and then mined, it will generate much knowledge and help us make managerial decisions. In this study, the genetic algorithm with logistic regression and random forest is used to select the appropriate feature is based on the correct diagnosis of the desired class, after applying statistical and probabilistic approaches in the data set and also preprocessing to remove redundant and lost data to extract features that have more variance in the complications of diabetes.

The genetic algorithm selects a subset of the most essential qualities for classification [87]. Logistic regression algorithm and a random forest to calculate the accuracy and examine the features with more variance in the rapid diagnosis of diabetes. We applied the data set properties as input to both algorithms. Then this algorithm calculated the accuracy of

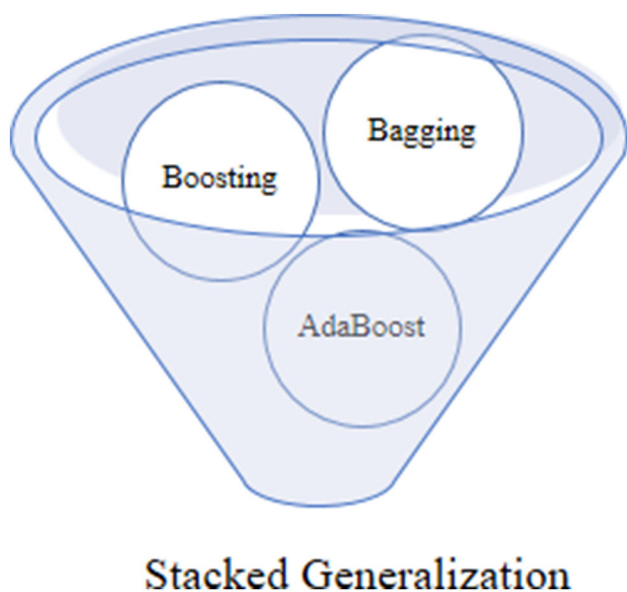


Fig. 6 Proper methods of ensemble learning

each feature. According to the results, logistic regression has a higher performance than the random forest classifier 0.70 times with 93% accuracy and five optimal features (Tables 9 and 10).

In the following—Features Selection Diabetes 130-US hospitals for years 1999–2008 Data Set for classification model attempts to select a minimally sized subset according to the following criteria: (1) The classification accuracy should increase; (2) the values for the selected features should have to close as a possible to the original class distribution. The feature selection results on the diabetes dataset are shown below.

Individual: [0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0].

Table 9 Comparison of holdout approach accuracy, sensitivity, and specificity rates

	Pima Indian diabetes datasets			Diabetes 130-US hospitals for years 1999–2008 data set		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
RF	93	95	93	95	94	92
KNN	83	92	82	88	89	92
MLP	90	85	88	93	90	89
Ada boost	92	94	88	86	88	87
D tree Classifier	93	90	96	94	89	92
NB	83	85	77	88	79	83
GBC	80	95	66	89	92	88
SVM	70	65	72	80	79	78
Extra Tree	80	86	79	91	93	92
Suggest Method (ST-GA)	98	97	100	99.06	99	100

Diagnosis of detection diabetes was compared in rates among various algorithms with the holdout approach

Feature Subset: ['age', 'discharge\_disposition\_id', 'time\_in\_hospital', 'num\_lab\_procedures', 'num\_procedures', 'num\_medications', 'diag\_2', 'number\_diagnoses'].

### 4.1 What is lacking in the current knowledge?

We chose ensemble learning in this work since it usually outperforms any trained models [24]. It has been successfully applied to both supervised (Regression, [25] Classification, and Distance learning [26]) and Unsupervised (Density estimation) learning tasks [27]. It has also been used to figure out how much packing fault [38, 69]. Using a series of models instead of a single model is advantageous for various reasons:

- *Performance*: compared to a single model, there is a significant performance improvement.
- *Error reduction*: predictive errors in machine learning models can be described using bias and variance.

As a result, this paper aims to identify the limitations of machine learning algorithms employed by other researchers in the accurate diagnosis of diabetes and compare them to the ensemble learning approach for better outcomes.

### 4.2 Strengths of the study

Using these findings, this chapter describes the limitations of machine learning models used for diabetes diagnosis using the dataset, with the goal of highlighting critical issues such as data quality, data quantity, explainability, and data privacy while getting quick results. The following are the article's strengths:

- Combining the greatest machine learning architectures for voting.

**Table 10** Comparison of accuracy, sensitivity, and specificity rates based on the  $K$ -fold cross-validation approach

	Diabetes 130-US hospitals for years 1999–2008 data set			Pima Indian diabetes datasets			Model
	$K$ fold = 15	$K$ fold = 10	$K$ fold = 5	$K$ fold = 15	$K$ fold = 10	$K$ fold = 5	
	95	94	93	90	93	87	RF
	88	87	88.06	91	94	83.26	KNN
	93	93	92	89	90	89	MLP
	86	86	87	78	52	50	Ada boost
	94	93.56	93	95	95	78	D tree Classifier
	88	89	86	55	55	55	NB
	89	90	88	95	95	73	GBC
	80	79	82	93	94	76	SVM
	91	91	89	78	78	78	Extra Tree
	99	99.01	98	98	98.8	97	Suggest Method (ST-GA)

Diagnosis of detection diabetes was compared in rates among various algorithms with the  $K$ -fold cross-validation approach

- Achieve a high level of classification reliability.
- When compared to single-core models, accuracy and error have improved.
- A comparison of the proposed method's outcomes.

Future research could resolve several significant shortcomings in this study. The study's first goal was to predict diabetes. and similar things:

- The sample size.
- Data that is not readily available or is not trustworthy.
- Inadequate access to hospital information.

## 5 Conclusion and future work

Diabetes has become one of the most important concerns of people and officials due to irreversible complications and its high prevalence. The PID and Diabetes 130-US hospital's 1999–2008 database was used to diagnose diabetes in this study. Data mining methods have been widely used in medicine and health care to diagnose and prevent diseases, choose treatment methods, and predict deaths and treatment costs during the last years. For this purpose, we used an ensemble learning algorithm called stacked generalization based on genetic algorithms to classify diabetic patients based on the observed complications. This study aimed to combine data mining algorithms to show that combining models can improve models. The highest accuracy was obtained using the proposed Stack Generalization algorithm according to the methods used Intelligence.

## 5.1 Contribution

A recent study has developed many machine learning algorithms for predicting diabetes. The ensemble learning method for the best diabetes prediction is key to this study. Through the site, we have gathered patient information. Following data collection, only the relevant features were eliminated from each data set to improve the proposed model's accuracy and remove unrelated features that slowed down calculation. The ensemble learning approach works more carefully in larger healthcare datasets and gives better outcomes. Finally, we developed a diabetes data gathering and ensemble learning approach for accurate and timely prediction. Our recommended system's total performance is between 98.8 and 99.9%. As a result, new medical researchers will benefit from future research and academic practice, particularly for Internet of Things-based prediction systems.

## 5.2 Future work

*In future research*, considering the importance of diagnosing the disease, we intend to expand the research in the field of diagnosing diseases such as breast cancer metastasis, lung cancer, Covid-19 by data mining tools and proposed algorithm and also develop and implement the subjects 1—Consumption of drugs and supplements (drug interaction) and 2—Provide solutions to caregivers and 3—Introduce a specialist related to the disease and 4—Online medical services.

**Author contributions** JA: Designed and performed experiments, analyzed data. BNM: supervised the findings of this work and co-wrote the paper. All authors discussed the results and contributed to the final manuscript.

**Funding** None.



## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Garcia-Molina, L., Lewis-Mikhael, A.M., Riquelme-Gallego, B., Cano-Ibanez, N., Oliveras-Lopez, M.J., Bueno-Cavanillas, A.: Improving type 2 diabetes mellitus glycaemic control through lifestyle modification implementing diet intervention: a systematic review and meta-analysis. *Eur. J. Nutr.* **59**(4), 1313–1328 (2020)
- Liang, Y.Z., Li, J.J.H., Xiao, H.B., He, Y., Zhang, L., Yan, Y.X.: Identification of stress-related microRNA biomarkers in type 2 diabetes mellitus: a systematic review and meta-analysis. *J. Diabetes* **12**(9), 633–644 (2020)
- Zhang, Y., Pan, X.F., Chen, J., Xia, L., Cao, A., Zhang, Y., et al.: Combined lifestyle factors and risk of incident type 2 diabetes and prognosis among individuals with type 2 diabetes: a systematic review and meta-analysis of prospective cohort studies. *Diabetologia* **63**(1), 21–33 (2020)
- Wong, J.J., Addala, A., Abujaradeh, H., Adams, R.N., Barley, R.C., Hanes, S.J., et al.: Depression in context: Important considerations for youth with type 1 vs type 2 diabetes. *Pediatr. Diabetes* **21**(1), 135–142 (2020)
- Australia, Healthdirect. Type 2 diabetes (2022). <https://www.healthdirect.gov.au/type-2-diabetes>
- Abdollahi, J., Moghaddam, B.N., Parvar, M.E.: Improving diabetes diagnosis in smart health using genetic-based Ensemble learning algorithm. Approach to IoT Infrastructure. *Future Gen. Distrib. Syst. J.* **1**, 23–30 (2019)
- Wang, S., Ma, P., Zhang, S., Song, S., Wang, Z., Ma, Y., et al.: Fasting blood glucose at admission is an independent predictor for 28-day mortality in patients with COVID-19 without previous diagnosis of diabetes: a multi-centre retrospective study. *Diabetologia* (2020). <https://doi.org/10.1007/s00125-020-05209-1>
- Debata, P.P., Mohapatra, P.: Diagnosis of diabetes in pregnant woman using a Chaotic-Jaya hybridized extreme learning machine model. *J. Integr. Bioinform.* **18**, 81–99 (2020)
- Hossain, M.S., Muhammad, G.: Cloud-assisted industrial internet of things (iiot)-enabled framework for health monitoring. *Comput. Netw.* **101**, 192–202 (2016)
- Temko, A.: Accurate wearable heart rate monitoring during physical exercises using PPG. *IEEE Trans. Biomed. Eng.* (2017). <https://doi.org/10.1109/TBME.2017.2676243>
- Abawajy, J.H., Hassan, M.M.: Federated internet of things and cloud computing pervasive patient health monitoring system. *IEEE Commun. Mag.* **55**(1), 48–53 (2017)
- La, H.J.: A conceptual framework for trajectory-based medical analytics with IoT contexts. *J. Comput. Syst. Sci.* **82**(4), 610–626 (2016)
- Rahmani, A.M., Gia, T.N., Negash, B., Anzanpour, A., Azimi, I., Jiang, M., Liljeberg, P.: Exploiting smart e-health gateways at the edge of healthcare internet-of-things: a fog computing approach. *Futur. Gener. Comput. Syst.* **78**, 641–658 (2017)
- Verma, P., Sood, S.K.: Cloud-centric IoT based disease diagnosis healthcare framework. *J. Parallel Distrib. Comput.* **116**, 27–38 (2018)
- Din, I.U., Guizani, M., Rodrigues, J.J., Hassan, S., Korotaev, V.V.: Machine learning in the Internet of Things: designed techniques for smart cities. *Futur. Gener. Comput. Syst.* **100**, 826–843 (2019)
- International Warfarin Pharmacogenetics C, Klein, T.E., Altman, R.B., Eriksson, N., Gage, B.F., Kimmel, S.E., et al.: Estimation of the warfarin dose with clinical and pharmacogenetic data. *N. Engl. J. Med.* **360**(8), 753–764 (2009). (pmid:19228618; PubMed Central PMCID: PMCPMC2722908)
- Hu, Y.H., Wu, F., Lo, C.L., Tai, C.T.: Predicting warfarin dosage from clinical data: a supervised learning approach. *Artif. Intell. Med.* **56**(1), 27–34 (2012). (Epub 2012/04/28. pmid:22537823)
- Tao, Y., Chen, Y.J., Fu, X., Jiang, B., Zhang, Y.: Evolutionary ensemble learning algorithm to modeling warfarin dose prediction for Chinese. *IEEE J. Biomed. Health Inform.* **23**(1), 395–406 (2018). (Epub 2018/07/12. pmid:29993619)
- Breiman, L.: Stacked regressions. *Mach. Learn.* **24**(1), 49–64 (1996)
- Wang, S.Q., Yang, J., Chou, K.C.: Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. *J. Theor. Biol.* **242**(4), 941–946 (2006). (Epub 2006/06/30. pmid:16806277)
- Palimkar, P., Shaw, R. N., Ghosh, A.: Machine learning technique to prognosis diabetes disease: random forest classifier approach. In: *Advanced Computing and Intelligent Technologies*, pp. 219–244. Springer, Singapore (2022)
- Ahmad, H.F., Mukhtar, H., Alaqail, H., Seliaman, M., Alhumam, A.: Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Appl. Sci.* **11**(3), 1173 (2021)
- Li, J., Yuan, P., Hu, X., Huang, J., Cui, L., Cui, J., et al.: A tongue features fusion approach to predicting prediabetes and diabetes with machine learning. *J. Biomed. Inform.* **115**, 103693 (2021)
- Dietterich, T.G.: Ensemble learning. *Handb. Brain Theory Neural Netw.* **2**(1), 110–125 (2002)
- Polikar, R.: Ensemble learning. In: *Ensemble Machine Learning*, pp. 1–34. Springer, Boston (2012)
- Zhou, Z. H.: Ensemble learning. In: *Machine Learning*, pp. 181–210. Springer, Singapore (2021)
- Sagi, O., Rokach, L.: Ensemble learning: A survey. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* **8**(4), e1249 (2018)
- Fatima, M., Pasha, M.: Survey of machine learning algorithms for disease diagnostic. *J. Intell. Learn. Syst. Appl.* **9**(01), 1 (2017)
- Alić, B., Gurbeta, L., Badnjević, A.: Machine learning techniques for classification of diabetes and cardiovascular diseases. In: *Embedded Computing (MECO), 2017 6th Mediterranean Conference on*, pp. 1–4. IEEE (2017)
- Kumar, A., Kumar, P., Srivastava, A., Kumar, V. A., Vengatesan, K., Singhal, A.: Comparative analysis of data mining techniques to predict heart disease for diabetic patients. In: *International Conference on Advances in Computing and Data Sciences*, pp. 507–518. Springer, Singapore (2020)
- Saru, S., Subashree, S.: Analysis and prediction of diabetes using machine learning. *Int. J. Emerg. Tech. Innov. Eng.* **5**(4) (2019)
- Subramanian, S., Regan, R., Perumal, T., Venkatachalam, K.: Semi-supervised machine learning algorithm for predicting diabetes using big data analytics. In: *Business Intelligence for Enterprise Internet of Things*, pp. 139–149. Springer, Cham (2020)
- Yang, H., Luo, Y., Ren, X., Wu, M., He, X., Peng, B., et al.: Risk prediction of diabetes: big data mining with fusion of multifarious physical examination indicators. *Inform. Fusion* **75**, 140–149 (2021)
- Beunza, J.J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., et al.: Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *J. Biomed. Inform.* **97**, 103257 (2019)
- Zeki, A. M., Taha, R., Alshakrani, S.: Developing a predictive model for diabetes using data mining techniques. In: *2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, pp. 24–28. IEEE (2021)
- Kalyankar, G. D., Poojara, S. R., Dharwadkar, N. V.: Predictive analysis of diabetic patient data using machine learning and

- Hadoop. In: 2017 international conference on I-SMAC (IoT in social, mobile, analytics and cloud)(I-SMAC), pp. 619–624. IEEE (2017)
37. Rghioui, A., Naja, A., Oumnad, A.: Diabetic patients monitoring and data classification using IoT application. In: 2020 International Conference on Electrical and Information Technologies (ICEIT), pp. 1–6. IEEE (2020).
  38. Joshi, T.N., Chawan, P.P.M.: Diabetes prediction using machine learning techniques. *Ijera* **8**(1), 9–13 (2018)
  39. Sikder, N., Masud, M., Bairagi, A.K., Arif, A.S.M., Nahid, A.A., Alhomyani, H.A.: Severity classification of diabetic retinopathy using an ensemble learning algorithm through analyzing retinal images. *Symmetry* **13**(4), 670 (2021)
  40. Ihnaini, B., Khan, M.A., Khan, T.A., Abbas, S., Daoud, M.S., Ahmad, M., Khan, M.A.: A smart healthcare recommendation system for multidisciplinary diabetes patients with data fusion based on deep ensemble learning. *Comput. Intell. Neurosci.* (2021). <https://doi.org/10.1155/2021/4243700>
  41. Banchhor, M., Singh, P.: Comparative study of ensemble learning algorithms on early stage diabetes risk prediction. In: 2021 2nd International Conference for Emerging Technology (INCET), pp. 1–6. IEEE (2021)
  42. Sabbir, M. M. H., Sayeed, A., Jamee, M. A. U. Z.: Diabetic retinopathy detection using texture features and ensemble learning. In: 2020 IEEE Region 10 Symposium (TENSYP), pp. 178–181. IEEE (2020)
  43. Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., Stiglic, G.: Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Sci. Rep.* **10**(1), 1–12 (2020)
  44. Muhammad, L.J., Algehyne, E.A., Usman, S.S.: Predictive supervised machine learning models for diabetes mellitus. *SN Comput. Sci.* **1**(5), 1–10 (2020)
  45. Abdulhadi, N., Al-Mousa, A.: Diabetes detection using machine learning classification methods. In: 2021 International Conference on Information Technology (ICIT), pp. 350–354. IEEE (2021)
  46. Naz, H., Ahuja, S.: Deep learning approach for diabetes prediction using PIMA Indian dataset. *J. Diabetes Metab. Disord.* **19**(1), 391–403 (2020)
  47. Al-Zebari, A., Sengur, A.: Performance comparison of machine learning techniques on diabetes disease detection. In: 2019 1st international informatics and software engineering conference (UBMYK), pp. 1–4. IEEE (2019)
  48. Rawat, V., Suryakant, S.: A classification system for diabetic patients with machine learning techniques. *Int. J. Math. Eng. Manage. Sci.* **4**(3), 729–744 (2019)
  49. Abdollahi, J., Nouri-Moghaddam, B., Ghazanfari, M.: Deep neural network based ensemble learning algorithms for the healthcare system (diagnosis of chronic diseases). *arXiv preprint arXiv:2103.08182* (2021)
  50. Abdollahi, J., Nouri-Moghaddam, B., Ghazanfari, M.: Deep neural network based ensemble learning algorithms for the healthcare system (diagnosis of chronic diseases) (2021). *arXiv preprint arXiv:2103.08182*
  51. Priyanka, N.A., Kumar, D.: Decision tree classifier: a detailed survey. *Int. J. Inf. Decis. Sci.* **12**(3), 246–269 (2020)
  52. Quinlan, J.R.: Induction of decision trees. *Mach. Learn.* **1**(1), 81–106 (1986)
  53. Jadhav, S.D., Channe, H.P.: Comparative study of K-NN, naive Bayes and decision tree classification techniques. *Int. J. Sci. Res.* **5**(1), 1842–1845 (2016)
  54. Karthikeyan, T., Thangaraju, P.: PCA-NB algorithm to enhance the predictive accuracy. *Int. J. Eng. Tech* **6**(1), 381–387 (2014)
  55. Amato, F., López, A., Peña-Méndez, E.M., Vañhara, P., Hampl, A., Havel, J.: Artificial neural networks in medical diagnosis. *J. Appl. Biomed.* (2013). <https://doi.org/10.2478/v10136-012-0031-x>
  56. Vijayarani, S., Dhayanand, S., Phil, M.: Kidney disease prediction using SVM and ANN algorithms. *Int. J. Comput. Bus. Res. (IJCBR)* **6**(2), 1–12 (2015)
  57. Platt, J. C.: 12 fast training of support vector machines using sequential minimal optimization. *Adv. Kernel Methods* 185–208 (1999)
  58. Ruggieri, S.: Efficient C4.5 [classification algorithm]. *IEEE Trans. Knowl. Data Eng.* **14**(2), 438–444 (2002)
  59. Liaw, A., Wiener, M.: Classification and regression by random Forest. *R News* **2**(3), 18–22 (2002)
  60. Izquierdo-Verdiguier, E., Zurita-Milla, R.: An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **88**, 102051 (2020)
  61. Zhang, Z.: Introduction to machine learning: k-nearest neighbors. *Ann. Transl. Med.* **4**(11), 218 (2016)
  62. Naimi, A., Balzer, L. B.: Stacked generalization: an introduction to super learning. *bioRxiv* 172395 (2017)
  63. Bui, D.T., Ho, T.C., Pradhan, B., Pham, B.T., Nhu, V.H., Revhaug, I.: GIS-based modeling of rainfall-induced landslides using data mining-based functional trees classifier with AdaBoost, Bagging, and MultiBoost ensemble frameworks. *Environ. Earth Sci.* **75**(14), 1101 (2016)
  64. Del Jesus, M.J., Hoffmann, F., Navascués, L.J., Sánchez, L.: Induction of fuzzy-rule-based classifiers with evolutionary boosting algorithms. *IEEE Trans. Fuzzy Syst.* **12**(3), 296–308 (2004)
  65. Kearns, M.: Boosting theory towards practice: Recent developments in decision tree induction and the weak learning framework. In: *Proceedings of the National Conference on Artificial Intelligence*, pp. 1337–1339 (1996)
  66. Choi, S., Kim, Y. J., Briceno, S., Mavris, D.: Prediction of weather-induced airline delays based on machine learning algorithms. In: 2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC), pp. 1–6. IEEE (2016)
  67. Schapire, R. E.: The boosting approach to machine learning: An overview. In: *Nonlinear estimation and classification*, pp. 149–171. Springer, New York (2003)
  68. Barik, S., Mohanty, S., Mohanty, S., Singh, D.: Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques. In: *Intelligent and Cloud Computing*, pp. 399–409. Springer, Singapore (2021)
  69. Sewell, M.: Ensemble learning. *RN* **11**(02), 1–34 (2008)
  70. Singh, N., Singh, P.: A stacked generalization approach for diagnosis and prediction of type 2 diabetes mellitus. In: *Computational Intelligence in Data Mining*, pp. 559–570. Springer, Singapore (2020)
  71. Singh, N., Singh, P.: Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybern. Biomed. Eng.* **40**(1), 1–22 (2020)
  72. Kuo, K.M., Talley, P., Kao, Y., Huang, C.H.: A multi-class classification model for supporting the diagnosis of type II diabetes mellitus. *PeerJ* **8**, e9920 (2020)
  73. Bernardini, M., Moretini, M., Romeo, L., Frontoni, E., Burattini, L.: Early temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: a multiple instance boosting approach. *Artif. Intell. Med.* **105**, 101847 (2020)
  74. Mahendran, N., Vincent, P.D.R., Srinivasan, K., Sharma, V., Jayakody, D.K.: Realizing a stacking generalization model to improve the prediction accuracy of major depressive disorder in adults. *IEEE Access* **8**, 49509–49522 (2020)
  75. Khan, F.A., Zeb, K., Alrakhami, M., Derhab, A., Bukhari, S.A.C.: Detection and prediction of diabetes using data mining: a comprehensive review. *IEEE Access* (2021). <https://doi.org/10.1109/ACCESS.2021.3059343>
  76. Mirjalili, S.: *Evolutionary Algorithms and Neural Networks*. Studies in Computational Intelligence, vol. 780. Springer, Cham (2019)

77. Mathew, T. V.: Genetic algorithm. Report submitted at IIT Bombay (2012)
78. Abdollahi, J., Keshandehghan, A., Gardaneh, M., Panahi, Y., Gardaneh, M.: accurate detection of breast cancer metastasis using a hybrid model of artificial intelligence algorithm. *Arch. Breast Cancer* (2020). <https://doi.org/10.32768/abc.20207118-24>
79. Zadeh, H.G., Jamshidi, H., Fayazi, A., Gholizadeh, M.H., Toussi, C.A., Danaeian, M.: A new model for retinal lesion detection of diabetic retinopathy using hierarchical self-organizing maps. *Iran J. Comput. Sci.* **3**(2), 93–101 (2020)
80. Bagherzadeh, J., Asil, H.: A review of various semi-supervised learning models with a deep learning and memory approach. *Iran J. Comput. Sci.* **2**(2), 65–80 (2019)
81. Iqbal, M.S., Luo, B., Khan, T., Mehmood, R., Sadiq, M.: Heterogeneous transfer learning techniques for machine learning. *Iran J. Comput. Sci.* **1**(1), 31–46 (2018)
82. Singh, P. P., Prasad, S., Das, B., Poddar, U., Choudhury, D. R.: Classification of diabetic patient data using machine learning techniques. In: *Ambient Communications and Computer Systems*, pp. 427–436. Springer, Singapore (2018)
83. Bhuvaneshwari, G., Manikandan, G.: A novel machine learning framework for diagnosing the type 2 diabetics using temporal fuzzy ant miner decision tree classifier with temporal weighted genetic algorithm. *Computing* **100**(8), 759–772 (2018)
84. Carrera, E.V., González, A., Carrera, R.: Automated detection of diabetic retinopathy using SVM. In: *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*, pp. 1–4, August 2017. IEEE (2017). <https://doi.org/10.1109/INTERCON.2017.8079692>
85. Sharma, P., Saxena, S., Sharma, Y. M.: An efficient decision support model based on ensemble framework of data mining features assortment & classification process. In: *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, pp. 487–491. IEEE (2018)
86. Wolpert, D. H.: Stacked generalization (1992)
87. Leardi, R.: Application of a genetic algorithm to feature selection under full validation conditions and to outlier detection. *J. Chemom.* **8**(1), 65–79 (1994)

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