STM JOURNALS

Journal of
Artificial Intelligence
Research & Advances

ISSN: 2395-6720
Volume 9, Issue 1, 2022
DOI Journal: 10.37591/JoAIRA

http://computers.stmjournals.com/index.php?journal=JoAIRA&page=index

Review JoAIRA

Machine Learning Techniques for Diabetes Prognosis

Uzma Khan, Mayur Patil*

Abstract

Diabetes is a long-term illness that has the potential to devastate the global health-care system. Diabetes will affect 522 million people worldwide by 2033. Diabetes develops in eagerness, hunger, and the process of excreting due to high blood sugar levels.. Among other problems, diabetes is a primary cause of blindness, renal failure, amputations, heart failure, and stroke. Our bodies convert food into sugars, or glucose, when we consume it. At that point, our pancreas is supposed to release insulin. Insulin is the key molecule that allows glucose to enter and be used for energy in our cells. This system, however, does not operate in the case Diabetes Type 1 and Type 2 diabetes are the most common, but there are many others, including gestational diabetes, which develops during pregnancy, and the other forms of Diabetes. Machine learning is a new discipline of data science that studies how machines learn from their past experiences. The goal of the research is to develop a module that can accurately identify diabetes in a patient early, by combining the outcomes of various machine learning approaches. Among the approaches used are K-Nearest Neighbor, Logistic Regression, Random Forest, and Support Vector Machine. The model's accuracy when utilizing each of the algorithms is pre-determined. The model accompanying the capital accuracy for forecasting diabetes is therefore preferred.

Keywords: Diabetes prediction, Machine Learning, K closest neighbor (KNN), Logistic Regression (LR), Support vector Machine (SVM), Random Forest (RF).

INTRODUCTION

Diabetes is a fearful condition all over the world. Diabetes can be caused by a variety of factors, including obesity, excessive blood glucose levels, and so on. It infects the hormone insulin, causing crab metabolism to be disrupted and blood sugar levels to drop. In diabetes, there is not enough insulin produced by the body. Diabetes expands as it influences 432 million community general, mainly in reduced- and middle-undertake countries with its own government, in accordance with the World Health Organization this figure keep increasing in activity from 460 billion by the year 2035. However, diabetes is prevalent in a number of countries, including Canada, China, and India. The population of the number of diabetics in India has lately increased to 110 million, bringing the total number of diabetics in the country to 45 million. Diabetes is the leading cause of death worldwide. Diabetes can be treated, and human lives saved if disorders like diabetes are detected early. To that end, this study looks into diabetes prediction using a variety of diabetes-related characteristics. To anticipate this disease, we used the Pima Indian Diabetes Dataset as a foundation, and we used multiple methods like Machine Learning classification and ensemble techniques. Machine learning is a method of explicitly instructing computers or machines. By constructing multiple classification and ensemble models from acquired datasets, various Machine Learning techniques deliver efficient results for collecting knowledge.

***Author for Correspondence**
Mayur Patil
E-mail: mayurpatil1429@gmail.com

Student, Master of Computer Applications, Thakur Institute of Management Studies Career Development and Research, Maharashtra, India

Received Date: April 01, 2022
Accepted Date: April 15, 2022
Published Date: April 20, 2022

Citation: Uzma Khan, Mayur Patil. Machine Learning Techniques for Diabetes Prognosis. Journal of Artificial Intelligence Research & Advances. 2022; 9(1): 25–32p.

Diabetes can be predicted using data. Different Machine Learning algorithms are capable of making predictions, but selecting the optimum methodology is difficult. As a result, we use common classification and ensemble algorithms on the dataset to make predictions [1–3].

LITERATURE REVIEW

Kumar *et al*. [4]

The suggested random Forest algorithm for diabetes prediction develops a system that can detect early diabetes using the Random Forest algorithm in machine learning, making a more accurate prognosis for a patient. The proposed model yields the best diabetic prediction results, suggesting that the prediction system can effectively forecast diabetes disease, [5–7] efficiently and, more importantly, instantly.

Nnamoko *et al*. [8]

They presented predicting diabetes onset. They used five widely used classifiers for the ensembles and a meta-classifier to aggregate their results in an ensemble supervised learning approach. The conclusions are stated and distinguished from other studies that have used the same dataset in the literature. It is demonstrated that diabetes onset prediction can be done more accurately utilizing the proposed strategy.

Joshi *et al*. [6]

Diabetes prediction made available using Machine Learning Techniques employs three supervised machine learning techniques to attempt to predict diabetes. Algorithms used are SVM, Logistic Regression, and Artificial Neural Networks (ANN). This project suggests an excellent method for detecting diabetes illness earlier.

Shetty *et al*. [9]

This project suggests an excellent method for detecting diabetes illness earlier.

Faruque *et al.* [5]

They proposed data mining package for diabetic illness prediction in healthcare, an intelligent study on diabetes prediction utilizing machine learning algorithms has been proposed. Six distinct machine learning algorithms were used. Evaluations are containing the algorithms' conduct and accuracy. A variety of machine learning techniques are compared in order to determine which algorithm is most suitable for diabetes prediction. Researchers are interested in diabetes prediction in order to train a program to detect if a patient is diabetic or not by using an appropriate classifier on a dataset. The classification procedure, according to earlier research, has not significantly improved. As diabetes prediction is a critical topic in computers, a system is required to address the concerns identified [9].

PROPOSED METHODOLOGY

The major goal of this study is to look for a data model that can better predict diabetes with a finer precision. To predict diabetes, we used dissimilar classification and ensemble algorithms.

Dataset Description

In the collection, many characteristics of 769 patients have been provided in Table 1. The dossier was composed from the UC Irvine Pima Indian Diabetes Dataset warehouse.

Table 1. Dataset Description.

S.N.	Attributes
1	Pregnancy
2	Glucose
3	Blood Pressure
4	Skin thickness
5	Insulin
6	BMI (Body Mass Index)
7	Diabetes Pedigree Function
8	Age

The class variable of each data point is the ninth attribute. This class variable displays the outcome 0 and 1 for diabetics, indicating whether they are positive or negative.

Distribution of Diabetic Patient

We created a model to predict diabetes, but the dataset was slightly skewed, with about 600 classes tagged as 0 indicates no diabetes and 267 as 1 means diabetic as shown in Figure 1.

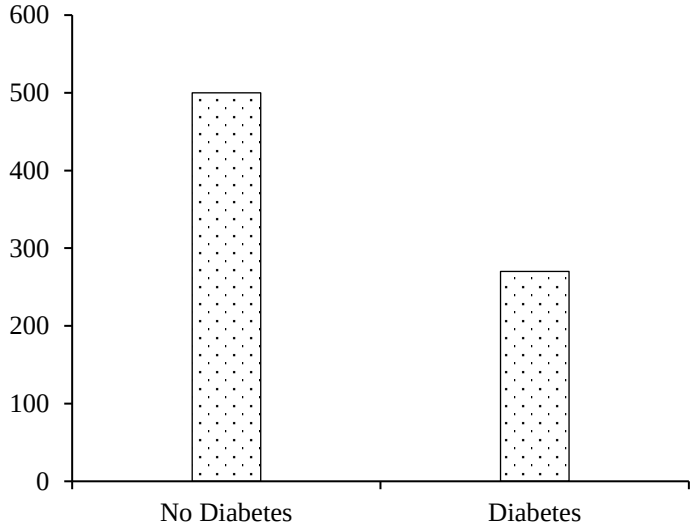

Figure 1. Ratio of Diabetic and Non Diabetic Patient.

Data Pre-processing

As a first step, data pre-processing is crucial. Missing values and other contaminants are common in healthcare data, which can reduce the data's efficacy. Using pre-processing as constituent, the excavating process develops the influence and value of the results process. This method should be performed for correct results and good indicator when utilizing Machine Learning Techniques on a dataset. Two steps are wanted to pre-process this dataset:

1. *Missing values removal:* Get rid of all instances with a value of zero (0). It is impossible to have a value of zero. As a result, this instance is no longer valid. We create feature subsets by removing irrelevant features/instances, a technique known as features subset selection, which decreases data dimensionality and speeds up work.
2. *Splitting of data*: Data is normalized in training and testing the model after it has been cleaned. When the data is spewed out, we train the algorithm on the training data set while ignoring the test data set. Logic and algorithms, as well as the values of the feature in the training data, will be used to generate the training model. The goal of normalization is to level the playing field for all traits.

Apply Machine Learning

We used Machine Learning methods after the data was ready. To predict polygenic disorder, we tend to use a spread of classification and ensemble techniques. The strategies were placed to take a look at employing a diabetes dataset from Pima Indians.[10] The main purpose is to assess the performance of various methods and determine their accuracy using Machine Learning techniques as well as to identify responsible and important features that play a significant part in prediction. The techniques are as follows.

Support Vector Machine

SVM, i.e., support vector machine is Associate in nursing contraction for Support Vector Machine, a directed machine intelligence system. SVM is the most commonly used second hand categorization procedure. SVM creates a hyperplane to separate two classes. In a large enough space,

you can form a hyperplane or a series of hyperplanes. This hyper plane can also be used for classification or regression. SVM can categorize entities that are not supported by data and separates them into specified classes. Separation is accomplished through the utilization of a hyperplane that separates to the closest coaching purpose in any class.

Algorithm

- Choose the plane that best separates the class.
- To recognize best choice energetic plane, you must reckon the margin, that is, the distance middle from two points the planes and the dossier.
- When the distance between classes is small, the chances of miscarriage are great, and vice versa. As a result, we must, choose the hyper plane that best divides the class.

K-Nearest Neighbor

The KNN invention is a directed machine learning approach. KNN can be used to address two together, classification and reversion questions. The sluggish prediction method is popular as K-Nearest Neighbor. KNN implies that objects that are similar are close to one other. Data points that are corresponding are commonly found nearby. KNN virus in the arrangement of new work is based on a correspondence rhythmical. The KNN algorithm takes all of the records and categorizes them based on how similar they are. The distance middle from two points, the spots is planned to utilize a timber-like form. To construct a forecast for a new data point, the invention searches the preparation data set for the tightest dossier points, to allure nearest neighbors. Here, K stands for the number of nearest neighbors, which is always a positive integer. The value of a neighbor is picked from a list of classes. The Euclidean distance between two points The two points P and Q are defined by the equation below:, P(p1, p2,.... Pn) and Q(q1,q2, Qn):

$$d(P,Q) = \sum_{i=1}^{n}(P_i - Q_i)^2$$

Algorithm

- Take a look at the Pima Indian Diabetes data collection, which is an example dataset of columns and rows.
- Take a test dataset of attributes and rows.
- Using the formula, find the Euclidean distance:

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{y}\sum_{j=1}^{m}\sum_{l=1}^{n-1}\left(R_{(j,l)} - P_{(i,l)}\right)^2}$$

- Then, decide a random value of K which is the number of nearest neighbors.
- Then, using these minimal distances and Euclidean distances, determine the nth column of each.
- Find out the same output values.

If two together numbers are the unchanging, the patient is diabetic; alternatively, he or she is not.

Logistic Regression

Logistic regression is a categorization algorithm that uses directed knowledge. It is used to determine the likelihood of a binary response given one or more predictors. They could have a continuous or discontinuous nature. When we wish to categorize or separate some data objects into categories, we apply logistic regression. It classifies data in binary form, that is, just in 0 and 1, which is used to classify patients as diabetic positive or negative. The basic purpose of logistic regression is to identify the optimal fit, which describes how the target and predictor variables are connected. Logistic regression is an undeviating regression model.

P=1/1+e–(a+bx) Sigmoid function P=1/1+e–(a+bx)

Ensembling

It is a type of machine learning that involves piecing together bits of data. Ensemble refers to joining several knowledge methods for certain tasks. It is employed because, in terms of accuracy, it exceeds all other individual models. The most common sources of inaccuracy are noise bias and variance, which ensemble techniques can help to reduce or eliminate. Bagging, pushing, ada-pushing,

gradient boosting, balloting, and equating are two prominent ensemble methods. To predict diabetes, we used the Bagging (Random Forest) and Gradient boosting ensemble approaches in this study.

Random Forest

It is an ensemble knowledge-located categorization and regression invention. When distinguished to additional models, it has a bigger level of accuracy. Large datasets are no problem with this strategy. Leo Bremen designed the Random Forest. It is a well-known ensemble learning strategy. By lowering difference, Random Forest increases Decision Tree depiction. It does everything by displaying the way of the classes, categorization, or mean prophecy (reversion) of each timber as a class, and then preparing a large number of resolution trees and metrics.

Algorithm

a. The first stage is to choose the "R" features from the total "m" features, where RM is the number of features.
b. Using the best split, separate the bud into subnodes.
c. Repeat steps a to b until you have achieved the "l" number of nodes.
d. Created a wood by recurrent steps, "a" is the number of opportunities to produce "n" wood.
e. The Gin-Index Cost Function is used by the random forest to find the best split:

$$\text{Gini} = \sum_{k=1}^{n} p_k * (1 - p_k)$$

Where; k= each class and p= probability of training instances.

The first stage is to look at the options and utilize the foundations of each arbitrarily generated decision tree to predict the outcome and store the predicted outcome at intervals around the target location.

Second, count the votes for each anticipated target and, as a result of the random forest formula's ultimate prediction, admit the predicted target with the most votes.

Random Forest has a number of options that produce accurate predictions for a variety of applications. Overview of the process has been shown in Figure 2.

MODEL BUILDING

This is the most crucial phase, which includes the development of a diabetes prediction model. For Predicting Diabetes, we applied an assortment of machine intelligence algorithms as previously established.

Procedure of Proposed Methodology

Step 1. Import the relevant libraries and the diabetes dataset.
Step 2. You can remove any lost data by pre-processing the dossier.
Step 3. Divide the dataset apart, accompanying the Training 80% of the data and the Test set taking 20%.
Step 4. Select a machine intelligence treasure from K-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, or Gradient Boosting.
Step 5. Create a classifier model utilizing the machine learning method utilizing the preparation set.
Step 6. Using the likely machine intelligence invention and the test set, test the Classifier model.
Step 7. Compare and contrast each classifier's experimental performance results.
Step 8. After inspecting multiple measures, select the best operating algorithm.

EXPERIMENTAL RESULTS

Several measures were made in this project. The suggested method employs a variety of classification and ensemble algorithms and is written in Python. These are some of the most common Machine Learning strategies for getting the maximum precision out of data. In this review, we can visualize that the random jungle classifier outpaces outperforms the rest. Overall, we employed the greatest Machine Learning approaches to forecast performance and obtain excellent accuracy (Figure 3).

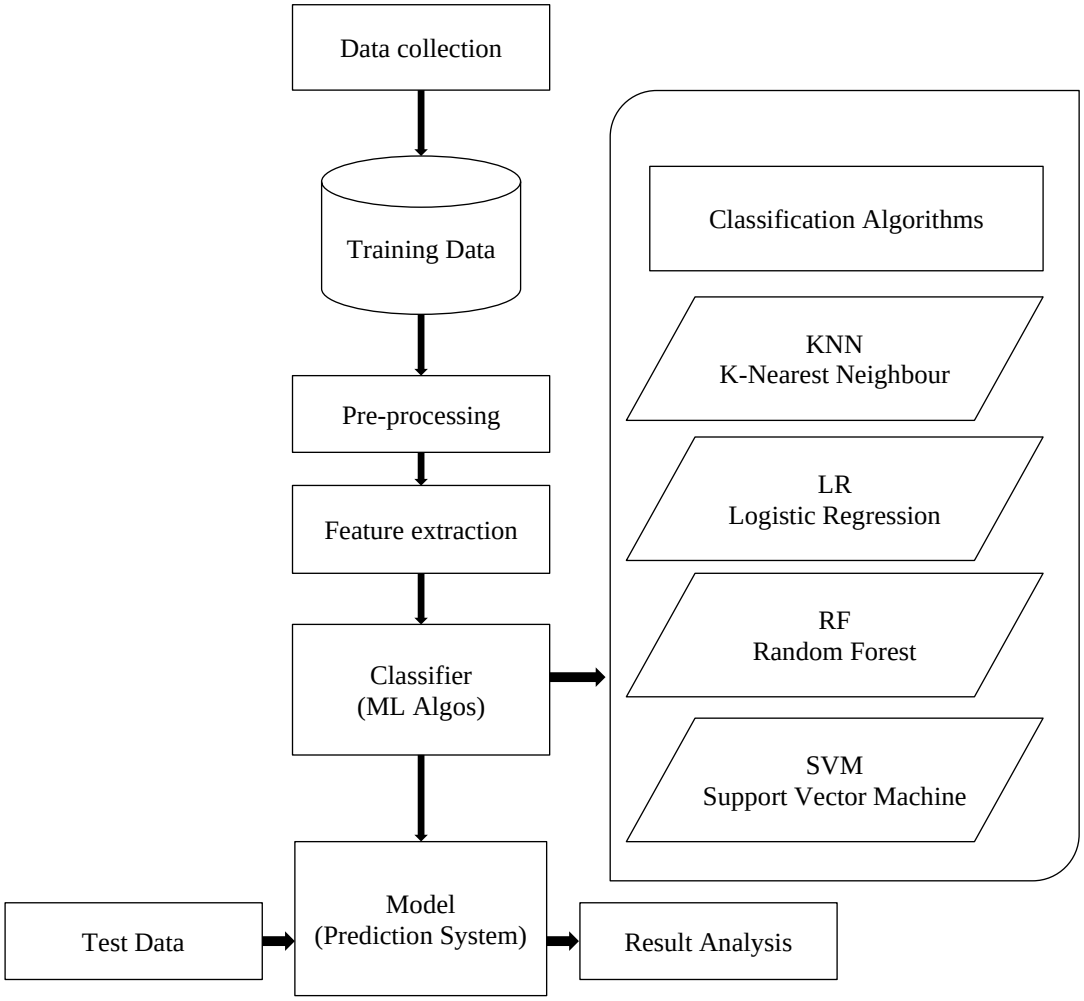

Figure 2. Overview of the Process.

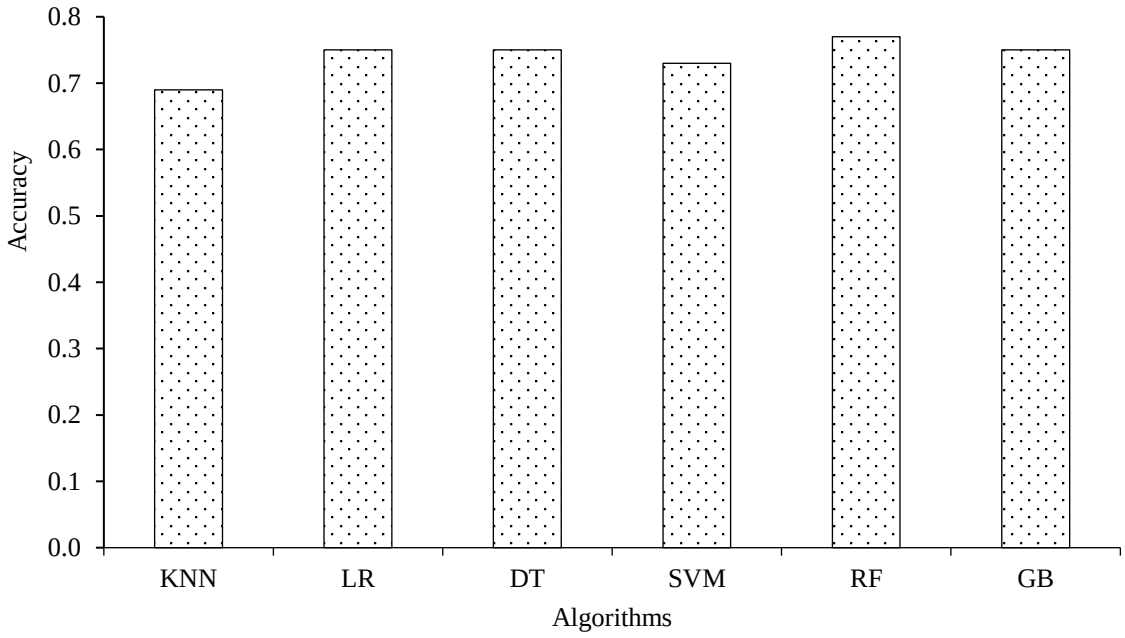

Figure 3. Accuracy Result of Machine learning methods.

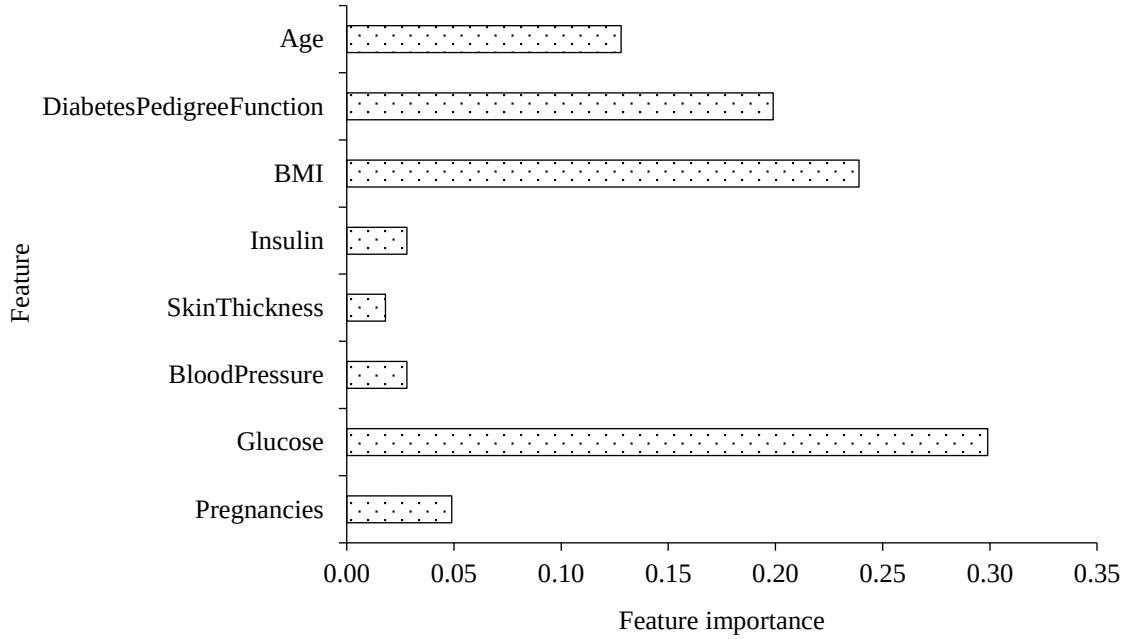

Figure 4. Feature Importance Plot for Random Forest.

The random forest algorithm has a quality that is useful for prediction. The importance of each feature that plays a significant role in diabetes has been plotted, with the X-hinge designating importance and the Y-spindle show feature titles in Figure 4.

CONCLUSION

The fundamental goal of this work was to develop and implement diabetes prediction using machine learning approaches, as well as to assess their performance, which was completed effectively. The recommended classification and ensemble learning technique employs SVM, KNN, random forest, decision tree, logistic regression, and gradient boosting classifiers and a classification accuracy of 79% was achieved. The findings of the study could aid doctors in making early predictions and judgments in order to cure diabetes and save people's lives.

REFERENCES

1. Kayal Vizhi, Aman Dash. Diabetes Prediction Using Machine Learning. International Journal of Animal Science and Technology (IJAST). 2020 May; 29(6): 2842–2852.
2. Giri B, Ghosh NS, Majumdar R, Ghosh A. Predicting Diabetes Implementing Hybrid Approach. In 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE. 2020 Jun 4; 388–39.
3. Misra A, Gopalan H, Jayawardena R, Hills AP, Soares M, Reza-Albarrán AA, Ramaiya KL. Diabetes in developing countries. J Diabetes. 2019 Jul; 11(7): 522–539.
4. Vijiya Kumar K, Lavanya B, Nirmala I, Caroline SS. Random Forest algorithm for the prediction of diabetes. In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE. 2019 Mar 29; 1–5.
5. Faruque MF, Sarker IH. Performance analysis of machine learning techniques to predict diabetes mellitus. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), IEEE. 2019 Feb 7; 1–4.
6. Joshi TN, Chawan PP. Diabetes prediction using machine learning techniques. Int J Eng Res Appl (IJERA). 2018 Jan; 8(1): 9–13.
7. Dutta D, Paul D, Ghosh P. Analysing feature importance for diabetes prediction using machine learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), IEEE. 2018 Nov; 924–928.

8. Nnamoko N, Hussain A, England D. Predicting diabetes onset: an ensemble supervised learning approach. In 2018 IEEE Congress on Evolutionary Computation (CEC), IEEE. 2018 Jul 8; 1–7.
9. Shetty D, Rit K, Shaikh S, Patil N. Diabetes disease prediction using data mining. In 2017 international conference on innovations in information, embedded and communication systems (ICIIECS), IEEE. 2017 Mar 17; 1–5.
10. Barakat N, Bradley AP, Barakat MN. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed. 2010 Jan 12; 14(4): 1114–20.

