

# Analyzing the behavior of different classification algorithms in diabetes prediction

Israa Nafea Mahmood, Hasanen S. Abdullah

Department of Computer Science, University of Technology, Al-senah Street, Baghdad, Iraq

## Article Info

### Article history:

Received Jul 26, 2021

Revised Aug 28, 2023

Accepted Sep 20, 2023

### Keywords:

Artificial neural network

Classification algorithms

Diabetes

Random forest

Support vector machine

## ABSTRACT

Diabetes is one of the deadliest diseases in the world that can lead to stroke, blindness, organ failure, and amputation of lower limbs. Researches state that diabetes can be controlled if it is detected at an early stage. Scientists are becoming more interested in classification algorithms in diagnosing diseases. In this study, we have analyzed the performance of five classification algorithms namely naïve Bayes, support vector machine, multi layer perceptron artificial neural network, decision tree, and random forest using diabetes dataset that contains the information of 2000 female patients. Various metrics were applied in evaluating the performance of the classifiers such as precision, area under the curve (AUC), accuracy, receiver operating characteristic (ROC) curve, f-measure, and recall. Experimental results show that random forest is better than any other classifier in predicting diabetes with a 90.75% accuracy rate.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Israa Nafea Mahmood

Department of Computer Science, University of Technology

Al-senah Street, Baghdad, Iraq

Email: cs.19.78@grad.uotechnology.edu.iq

## 1. INTRODUCTION

Diabetes is a well-known chronic disease that poses a significant issue in the world health systems. The pancreas releases the insulin hormone, which permits glucose to transit from food to the bloodstream [1]. Diabetes is caused by the lack of insulin and can result in serious damage such as coma, heart attack, weight loss, cardiovascular dysfunction, blindness, ulcer, and damage of the nerve system [2]. In 2020, 10% of the United States population has diabetes according to centers of disease control and prevention (CDC) [3]. Studies show that the number of people having diabetes will reach 25% in 2030 [4]. The availability of accurate early prediction can significantly reduce the risk and severity of diabetes.

Over the last few years, there was a growth in the number of applications that use artificial intelligence in diagnosing diseases such as breast cancer [5], [6], Alzheimer [7], parkinson's disease [8], coronary artery disease [9], skin disease [10], anxiety and depression disorders [11]. The use of classification algorithms in computer-aided diagnoses systems has helped in the detection of diseases at an early stage. The main goal of this study is to find which classification algorithm achieves minimum error rate in predicting diabetic patients without human involvement. The early detection of diabetes can enhance the treatment process and save the patient's life.

In this research, we will evaluate the performance of different classification algorithms like an artificial neural network, decision tree, support vector machine, random forest, and naïve Bayes using a diabetes dataset. In recent years, several data mining techniques were used to predict diabetes. The use of such methods has helped in early detection, which resulted in minimizing the complication of diabetes. The summary of literature review is illustrated in Table 1.

Table 1. Summary of literature review

Ref.	Dataset Description	Implemented Method	Accuracy
Paul <i>et al.</i> [12]	consists of 392 instances and 8 features	evaluated the performance of 3 classifiers (random forest, support vector machine, and logistic regression). Random forest attains the highest accuracy	84%
Dey <i>et al.</i> [13]	consists of 768 instances and 8 features	used artificial neural network with min-max scaler	82%
Swapna <i>et al.</i> [14]	consists of Electrocardiograms (ECG) of 20 people	extracted complex features from heart rate variability data using long short-term memory, conventional neural network (CNN) and fed them to support vector machine	95%
Zhu <i>et al.</i> [15]	consists of 768 instances and 8 features	used PCA, k-mean and logistic regression	97%
Wu <i>et al.</i> [16]	consists of 768 instances and 8 features	used k-mean and logistic regression	95%
Tripathi and Kumar [17]	consists of 768 instances and 8 features	implemented different predictive analysis algorithms. Random forest achieved the highest accuracy rate	87%
Li <i>et al.</i> [18]	consists of 768 instances and 8 features	used metaheuristic algorithm with k-mean	91%
Husain and Khan [19]	not specified	implemented an ensemble model by combining the propabilities of different machine learning algorithms.	96%

Paul [12] analyzed the performance of three classification algorithms; random forest, support vector machine, and logistic regression in predicting diabetes. The results show that random forest algorithm achieves a high accuracy rate of 84% than support vector machine and logistic regression. Dey *et al.* [13] implemented a web application for diabetes prediction. The implemented model used an artificial neural network among different classification algorithms like support vector machine, k-nearest neighbor, and naïve Bayes. The proposed model of artificial neural network and min-max scalar achieves a high accuracy rate of 82% compared to other classification algorithms.

Swapna *et al.* [14] implemented a methodology to extract complex features from heart rate variability data using long short-term memory, conventional neural network (CNN). These features are fed into support vector machine for diabetes prediction. The proposed system achieved a 95% accuracy rate.

Zhu *et al.* [15] proposed a data mining framework to diagnose diabetes. The framework used principal component analysis to address the correlation issue in the feature set, a k-mean clustering algorithm to clean the data from outliers, and a logistic regression algorithm to classify the data. The proposed framework was able to successfully classify the data with a 97% accuracy rate.

Wu *et al.* [16] suggested a two-stage framework to diagnose diabetes. In the first stage, the incorrectly clustered data are eliminated using an improved k-mean algorithm; the resulting dataset is then used in the following stage as input to the logistic regression algorithm to classify the data. The model attains a high accuracy rate of 95%.

Tripathi and Kumar [17] used several predictive analysis algorithms namely support vector machine, linear discriminant analysis, random forrest, and k-nearest neighbor. The study shows that Random Forrest algorithm gives a high accuracy rate of 87% compared to other algorithms. Li *et al.* [18] implemented a diabetes prediction model using a k-nearest neighbor algorithm. The proposed system used metaheuristic algorithms with k-mean to select features. The model achieves a good accuracy rate of 91%. Husain and Khan [19] proposed a diabetes prediction model by merging the probabilities of different machine learning algorithms into an ensemble model.

In our research, the performance of different machine learning algorithms in diagnosing diabetes is evaluated. The research aims to find the algorithm that achieves the maximum accuracy in identifying diabetes at an early stage. The remainder of this study is organized as follows; Section 2 discusses the algorithms applied in this study and the structure of the dataset. Section 3 shows the results of different classification algorithms, and this study is concluded in section 4.

## 2. RESEARCH METHOD

### 2.1. Data set description

In this paper, we used diabetes dataset that can be found in Ukani [20] to assess the performance of different classification algorithms. The dataset contains medical information of female patients. The dataset consists of 8 numeric features and one target class to determine if a patient has diabetes or not (0 or 1). The diabetic dataset contains the information of 2000 female patients. In data preprocessing, we removed noise and outliers. The dataset is divided into train and test datasets with a test size equal to 20%.

## 2.2. Model design and implementation

In this paper, we analyzed the behavior of different classification algorithms to predict diabetes. The procedure used in this study is illustrated in Figure 1. Different machine learning algorithms namely artificial neural network, random forest, support vector machine, decision tree, and naïve Bayes have been implemented to find the algorithm that achieves the minimum accuracy rate in predicting diabetes.

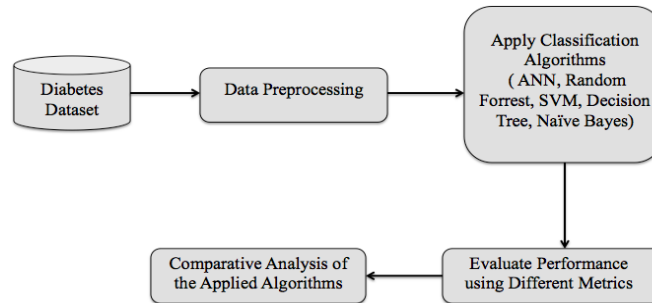


Figure 1. Diagram of the proposed model

### 2.2.1. Artificial neural network

One of the main algorithms in machine learning is an artificial neural network. They are designed to mimic the behavior of the human brain; it consists of input, hidden and output layers. The hidden layer is responsible of finding hidden information in the input data and translates it into a form that the output layer can use [21]. Figure 2 illustrates the architecture of the applied artificial neural network in this study.

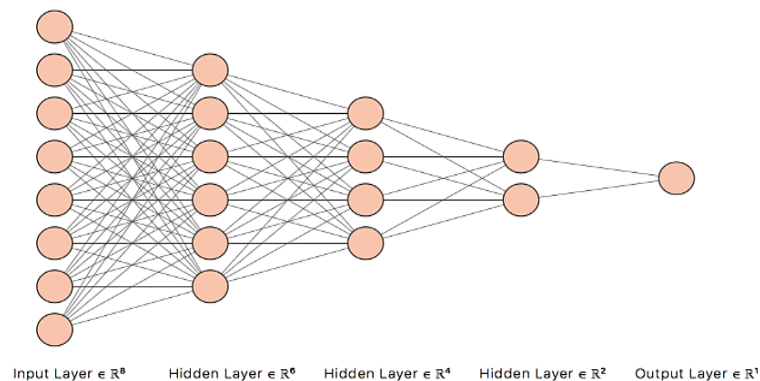


Figure 2. The architecture of the artificial neural network

### 2.2.2. Random forest

Random forest is a well-known classification algorithm that is utilized in solving regression and classification problems. It gets the prediction from multiple decision tree classifiers that work on parts of the dataset instead of the entire dataset. The performance of the model increases, when the number of classifiers increases [22]. We implemented a random forest model using Python libraries.

### 2.2.3. Support vector machine

Support vector machine is a powerful classification algorithm that is used in regression and classification. Support vector machine works by finding the best hyperplane that separates between different classes of the data. The data are plot in m-dimensional space where m represents the number of features [23]. We implemented support vector machine using Python libraries.

### 2.2.4. Decision tree

A decision tree is a supervised learning algorithm. It focuses on generating rules from the training data that are used to predict the class. Decision tree consists of a root, branches, and leaf nodes. One of the

most challenging points in decision tree is selecting the feature that best split the data [24]. We implement a decision tree algorithm using Python libraries.

### 2.2.5. Naïve bayes

It is a probabilistic machine-learning algorithm that depends on Bayes theorem. All features are independent and irrelevant from each other is assumed in naïve Bayes algorithm. One of the main advantages of naïve Bayes is its ability to work with data problems such as imbalanced datasets and missing values [25]. We implement it using Python libraries.

## 3. RESULTS AND DISCUSSION

Classification algorithms performance is assessed using a variety of metrics like precision, receiver operating characteristic (ROC) curve, recall, F-measure, accuracy, and area under the curve (AUC) [26], [27]. Five classification algorithms were evaluated in our study namely artificial neural network, naïve Bayes, random forest, decision tree, and support vector machine. The result of different classification algorithms in terms of true positive, true negative, false positive, and false negative are illustrated in Table 2.

Table 2. Confusion matrix of different classification algorithms

Model		Actual	
		Non-diabetic	Diabetic
Artificial Neural Network	Non-diabetic	235	28
	Diabetic	58	79
Decision Tree	Non-diabetic	239	24
	Diabetic	72	65
Gaussian NB	Non-diabetic	216	47
	Diabetic	54	83
Random Forrest	Non-diabetic	250	13
	Diabetic	22	115
Support vector machine	Non-diabetic	230	33
	Diabetic	73	64

Table 3 illustrates that random forest performs better than any other classifiers with regard to precision, recall, accuracy, and F-measure. On the other hand, support vector machine obtains the minimum accuracy rate of 73.50%. For more qualitative demonstration, Figure 3 depicts the ROC curve of the classification algorithms. Random forest ROC curve is closest to the top left corner than any other classifiers. We also measure the performance of the classifiers using AUC to show which classifier has high separation ability between different classes. Figure 4 shows that random forest obtains a high AUC value of 0.88.

Table 3. Performance measures for various classification algorithms

Model	Accuracy (%)	F-measure (%)	Precision (%)	Recall (%)
Artificial Neural Network	78.50	74.64	77.01	73.50
Decision Tree	76.00	70.40	74.94	69.16
Gaussian NB	74.75	71.61	71.92	71.36
Random Forrest	90.75	89.48	90.63	88.59
Support vector machine	73.50	67.99	70.94	67.08

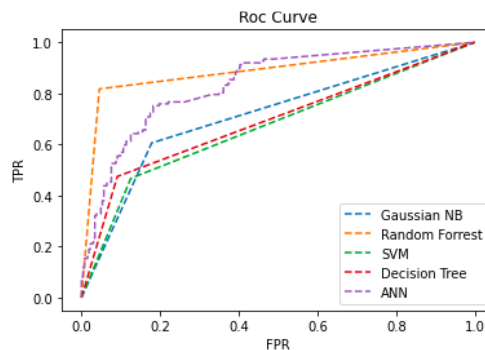


Figure 3. ROC curve for various classification algorithms

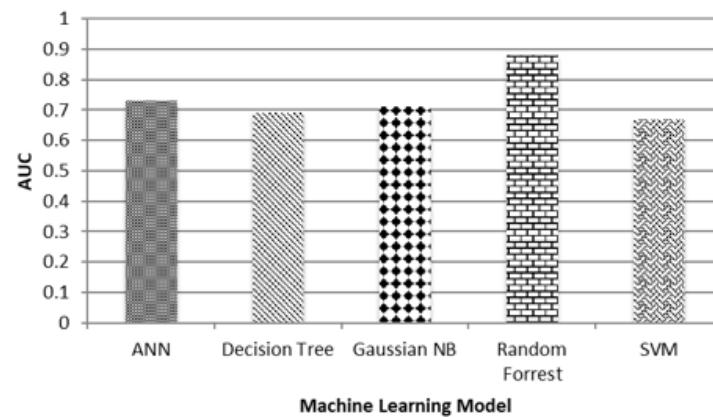


Figure 4. AUC scores of various classifiers

#### 4. CONCLUSION

Diabetes is an incurable disease that affects 9% of the world's population. Diabetes can lead to serious complications if it is not probably treated. As a result, diabetes prediction at an early stage is critical in the treatment process. We have analyzed the performance of multiple machine learning models namely artificial neural network, support vector machine, random forest, Gaussian naïve Bayes, and decision tree using diabetes dataset for 2000 female patients. We conclude that random forest surpasses other machine learning models in diagnosing diabetes with a 90.75% accuracy rate and 0.88 AUC score.




#### REFERENCES

- [1] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on Pima Indians diabetes dataset," in *2017 International Conference on Computing Networking and Informatics (ICCNi)*, 2017, pp. 1–5, doi: 10.1109/ICCNi.2017.8123815.
- [2] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Procedia Computer Science*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.
- [3] "National diabetes statistics report," Atlanta, GA, 2020.
- [4] H. Sung *et al.*, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.
- [5] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–17, 2019.
- [6] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset," in *Proceedings of the 2nd international conference on machine learning and soft computing*, 2018, pp. 5–9.
- [7] X. Liu, K. Chen, T. Wu, D. Weidman, F. Lure, and J. Li, "Use of multimodality imaging and artificial intelligence for diagnosis and prognosis of early stages of Alzheimer's disease," *Translational Research*, vol. 194, pp. 56–67, Apr. 2018, doi: 10.1016/j.trsl.2018.01.001.
- [8] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, and V. Venkatraman, "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease," *Future Generation Computer Systems*, vol. 83, pp. 366–373, 2018.
- [9] M. Abdar, W. Książek, U. R. Acharya, R.-S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 179, p. 104992, Oct. 2019, doi: 10.1016/j.cmpb.2019.104992.
- [10] N. S. A. AlEnezi, "A method of skin disease detection using image processing and machine learning," *Procedia Computer Science*, vol. 163, pp. 85–92, 2019, doi: 10.1016/j.procs.2019.12.090.
- [11] T. Richter, B. Fishbain, E. Fruchter, G. Richter-Levin, and H. Okon-Singer, "Machine learning-based diagnosis support system for differentiating between clinical anxiety and depression disorders," *Journal of Psychiatric Research*, vol. 141, pp. 199–205, Sep. 2021, doi: 10.1016/j.jpsychires.2021.06.044.
- [12] D. Paul, "Analysing feature importances for diabetes prediction using machine learning," *2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 924–928, 2018.
- [13] S. K. Dey, A. Hossain, and M. M. Rahman, "Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm," *2018 21st International Conference of Computer and Information Technology, ICCIT 2018*, pp. 1–5, 2019, doi: 10.1109/ICCITECHN.2018.8631968.
- [14] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes detection using deep learning algorithms," *ICT Express*, vol. 4, no. 4, pp. 243–246, 2018, doi: 10.1016/j.ict.2018.10.005.
- [15] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics in Medicine Unlocked*, vol. 17, no. January, p. 100179, 2019, doi: 10.1016/j.imu.2019.100179.
- [16] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics in Medicine Unlocked*, vol. 10, no. August 2017, pp. 100–107, 2018, doi: 10.1016/j.imu.2017.12.006.
- [17] G. Tripathi and R. Kumar, "Early prediction of diabetes mellitus using machine learning," *ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)*, pp. 1009–1014, 2020, doi: 10.1109/ICRITO48877.2020.9197832.




- [18] X. Li, J. Zhang, and F. Safara, "Improving the accuracy of diabetes diagnosis applications through a hybrid feature selection algorithm," *Neural Processing Letters*, no. 0123456789, 2021, doi: 10.1007/s11063-021-10491-0.
- [19] A. Husain and M. H. Khan, "Early diabetes prediction using voting based ensemble learning BT-advances in computing and data sciences," 2018, pp. 95–103.
- [20] V. Ukani, "Diabetes dataset."
- [21] D.-K. Bui, T. N. Nguyen, T. D. Ngo, and H. Nguyen-Xuan, "An artificial neural network (ANN) expert system enhanced with the electromagnetism-based firefly algorithm (EFA) for predicting the energy consumption in buildings," *Energy*, vol. 190, p. 116370, Jan. 2020, doi: 10.1016/j.energy.2019.116370.
- [22] P. Probst, M. N. Wright, and A.-L. Boulesteix, "Hyperparameters and tuning strategies for random forest," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 3, p. e1301, May 2019, doi: 10.1002/widm.1301.
- [23] I. Ahmad, M. Basher, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE Access*, vol. 6, pp. 33789–33795, 2018, doi: 10.1109/ACCESS.2018.2841987.
- [24] P. Sathiyarayanan, S. Pavithra., M. S. A. I. SARANYA., and M. Makeswari., "Identification of breast cancer using the decision tree algorithm," in *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2019, pp. 1–6, doi: 10.1109/ICSCAN.2019.8878757.
- [25] F. Razaque *et al.*, "Using naïve bayes algorithm to students' bachelor academic performances analysis," in *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, 2017, pp. 1–5, doi: 10.1109/ICETAS.2017.8277884.
- [26] H. Sahli, "An introduction to machine learning," in *TORUS 1-Toward an Open Resource Using Services*, Wiley, 2020, pp. 61–74.
- [27] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, Jul. 2019, doi: 10.1016/j.patcog.2019.02.023.

## BIOGRAPHIES OF AUTHORS



**Israa Nafea Mahmood**    receives a PhD degree in computer science from the University of Technology, Baghdad, Iraq with the Dissertation "Wisdom Model Building from Big Data based on Collective Intelligence Approach". She also received her B.Sc. and M.Sc. (Computer Science) from University of Baghdad, Iraq in 2012 and 2016, respectively. Her research intresets are in Security, Machine Learning, Data Mining, Artificial Intelligence and extracting wisdom from data. She has published several papers in international journals. She can be contacted at email: asraa\_nafaa@yahoo.com or cs.19.78@grad.uotechnology.edu.iq.



**Hasanen S. Abdullah**    Assist Professor qualified to direct research at University of Technology, Iraq, and other Iraqi Institutions. He got his B.Sc., M.Sc. and Ph.D. in computer science from University of Technology, Iraq in 2000, 2004 and 2008 respectively. His area of interests is Artificial Intelligence, Machine Learning, Swarm Intelligence, Pattern Recognition, Data Mining & warehouse, and Business Intelligence. He can be contacted at email: 110014@uotechnology.edu.iq.