BAUET JOURNAL
Published by
Bangladesh Army University of Engineering & Technology
Journal Homepage: http://journal.bauet.ac.bd
https://doi.org/10.59321/BAUETJ.V4I1.9

A Comparative Study of Machine Learning and Deep Learning Techniques for Diabetes Prediction

A. K. Soykot Azad Snigdho[1], Nazmul Hussain[1*], Md. Abdul Hamid2, Nasirul Mumenin[1], Tamanna Jannat[1], Anika Tahsin1, Md. Rajib Ali[1], Partha Pratim Debnath[1]

[1]Department of Information and Communication Engineering, Bangladesh Army University of Engineering & Technology, Qadirabad, Natore-6431, Bangladesh
[2]Department of Mechanical Engineering, Bangladesh Army University of Engineering & Technology, Qadirabad, Natore-6431, Bangladesh

Abstract: Diabetes mellitus is a cluster of conditions that impact the body's utilization of glucose, a vital energy source for the cells of muscles and tissues. International Diabetes Federations (IDF) shows that almost 382 million people are living with diabetes. The goal of this research is to predict diabetes by supervised machine learning algorithms. The result is then compared with the deep learning approaches. The conventional machine learning algorithms are used here i.e., logistic regression (LR), gradient boost (GB), decision tree (DT) and random forest (RF). Then deep learning (DL) method implement to predict and detect diabetes through neural network. The research is done with the Pima Indians Diabetes dataset which is publicly available. This dataset consists of eight input parameters with 768 samples where 268 samples for diabetic and rest of them are non-diabetic patients. The accuracy was obtained using the LR, GB, DT, RF and DL are 86%, 93%, 91.2%, 95% and 94.2% respectively. The accuracy shows that random forest gained better performance than the logistic regression, gradient boost, decision tree and deep learning.

Keywords: *Diabetes Prediction, Machine Learning, Logistic Regression, Deep learning, Gradient Boost.*

Introduction: Diabetes is a widespread health condition globally, often leaving individuals bewildered about its onset, progression, and symptomatology [1]. This metabolic disorder, referred to as diabetes, is marked by elevated glucose levels, which can inflict damage on vital organs, potentially resulting in additional health complications [2]. With the rise of wearable devices and sensor technologies capable of continuously monitoring a patient's health, the healthcare sector is witnessing a growing integration of machine learning. This technological advancement also aids healthcare professionals in analyzing data to identify patterns and early indicators, ultimately improving diagnostic accuracy and treatment strategies. World health organization declares that 422 million people have diabetes and majority of them are lower and middle-income countries [3]. Every year 1.5 million people die due to diabetes. There are generally two types of diabetics. Type 1 is an autoimmune disease. It is also called insulin dependent diabetes. It is generally seen in child and young adults. At the beginning, type 1 diabetes does not show any symptoms. When this type shows symptoms that means the insulin are produced by pancreas are destroyed. Type 2 is a common type of diabetes. Up to 95 % people with diabetes have type 2 [4]. Modern research says that if the diabetes can be predicted at the early stages, there is a

chance to recover it. In the continuous advancement of the technology machine learning and deep learning techniques are used to predict diabetes and various diseases more accurately.

Literature Review: Recently several researches have been found in literature that is predicting diabetes using the machine learning and deep learning techniques. For instance, ANN diabetes prediction model is created whose final accuracy is found 87% [5]. By the use of machine learning and deep learning approach the author show that the maximum accuracy is 90% only in XGBoost classifier [6]. The maximum accuracy is obtained 97% in deep learning model incorporating knowledge representation vectors [7]. The researcher displayed a maximum accuracy of 95% by employing quantum machine learning in combination with deep learning to predict diabetes [8]. For type 2 diabetes there are several algorithms such as KNN,SVM,DT,RF,GB,NN and NB are used and maximum accuracy is found in neural network which is 82.54% [9]. By the use of 768 samples in PIMA Indian datasets the maximum accuracy is obtained 88.315% in random forest algorithm [10]. Using random forest classifier the author shows that the maximum accuracy is obtained 83.67% [11]. Focusing on early detection, the author found that the support vector machine is well suited for obtaining a better accuracy and that is 77.73% [12].Three machine learning techniques i.e. naïve bayes, SVM, and decision tree, were used to diagnose diabetes at an earlier stage with an accuracy of 76.3%, 65.1 % and 73.82% respectively [13]. So, from the above literature it can be proved that machine learning and deep learning are well suited for diabetes prediction.

Article history:
Received 26 March 2023
Received in revised form 09 August 2023
Accepted 08 October 2023
Available online 15 November 2023

Corresponding author details: Nazmul Hussain
E-mail address: nazmul@bauet.ac.bd
Tel: +8801752061777

This research analysed PIMA Indian datasets to predict diabetes using machine learning and deep learning methods, in contrast to most prior works that did not concentrate on various machine learning techniques. The primary purpose of this study is to scrutinize the effectiveness of traditional machine learning and deep learning techniques in diabetes prognosis. Conventional machine learning techniques such as logistic regression, gradient boost, decision tree, and random forest are utilized accompanied with deep learning techniques. Random forest yields an accuracy rate of 95%, which is notably superior to other machine learning and deep learning techniques.

Materials and Methods: The overall methodology is described in Figure 1. For the diabetics prediction Pima Indian diabetes dataset (PIDD) are used. Data exploration is used to better understand the sources from which our data is gathered. The data must be preprocessed and normalize. It means that in the data set there should not be any missing, duplicate or unexpected value. For creating deep neural network architecture 64 neurons, 32

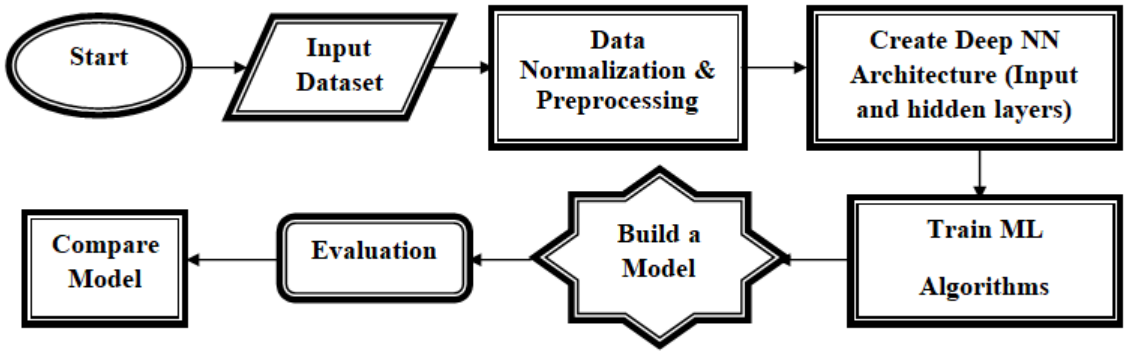

Fig. 1: Proposed framework.

neurons and 16 neurons are used. Then, conventional machine learning algorithm such as logistic regression, gradient boost, decision tree and random forest are used to train the model. After that performance is evaluated by using different indicator such as accuracy, f1-score, recall and auc curve. The proposed framework is illustrating the steps taken during implementation.

Data Description: The diabetes dataset is collected from PIDD. In the dataset there 768 instances with 8 different attributes. In the dataset '0' is used for negative diabetes and '1' is used for positive diabetes. The Table-1 shows the attribute description.

Table 1. The Datasets Variables.

Input	Mean	Standard Deviation
Pregnancies	0.226180	0.198210
Glucose	0.607510	0.16068
Blood Pressure	0.566438	0.158654
Skin Thickness	0.207439	0.161134
Insulin	0.094326	0.136222
BMI	0.476790	0.117499
Diabetes Pedigree Function	0.194990	0.136913
Age	0.410381	0.145188

Naïve Bayes Classifier : The Naive Bayes classifier is a machine learning method that is supervised and used for data categorization, such as text classification [14]. It falls under the group of generative learning algorithms and constructs a model of the input distribution of a specific category or class. Unlike discriminative classifiers such as logistic regression, Naive Bayes does not acquire knowledge about the crucial characteristics that aid in separating classes. Figure 2 show that true positive is 384, which tells how many positive classes are correctly classified true and false positives 19, which tells how many negative classes are incorrectly classified. On the other hand, false negative is 135 which shows how many positive classes are incorrectly classified. True negative is 76 which shows how many negative classes are correctly classified. In Figure 3 ROC curve (AUC) score is 0.82% summarized and accuracy of logistic regression which is 86%. The testing dataset is presented in Table-2.

Table 2. Classification report for Testing Dataset.

Classifier	Accuracy	Precision	Recall	F1- Score	Roc-auc score
Naive Bayes	76%	0.76	0.77	0.76	81%

Logistic Regression: The use of the logistic regression model has been widely used in many fields, including the biological sciences [15]. When dividing data objects into categories is the goal, the logistic regression approach is utilized. Figure 4 shows the amount of true positive are 384 and false positive are 19. Again, false negative 135 shows how many positive classes are incorrectly classified and true negative 76 shows how many negative classes are correctly classified. The area under the ROC curve (AUC) score is 0.82% which is shown in Figure 5. The accuracy of logistic regression is 86%. The logistic regression model shown in equation (1) below provides the foundation for the logistic regression procedure [15]. The testing dataset is presented in Table-3.

Fig. 2. The Confusion Matrix of Naive Bayes.

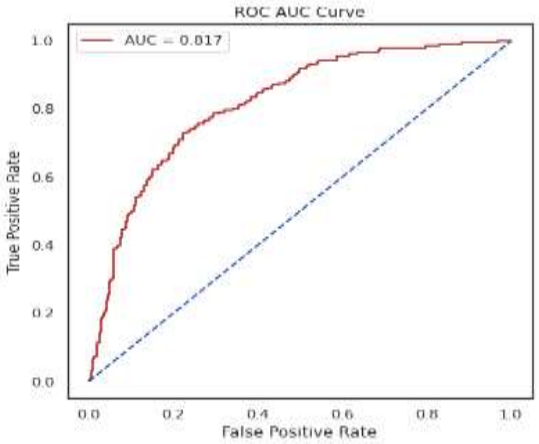

Fig. 3. ROC AUC Curve of Naive Bayes.

$$y = h_{\square}(x) = \square^{T} x \qquad \text{eq.1}$$

The function in equation (2) to predict the probability that a given patient (with given attributes) belongs to the "1" (positive) class versus the likelihood that it belongs to the "0" (negative) class given equation (1) will be incredibly inefficient at predicting our binary values (y (i) 0 and 1) [15].

$$p(y=1 \mid x) = h_{\theta}(x) = \frac{1}{1+\exp(-\theta^{T} x)} \equiv \sigma(\theta^{T} x) \qquad \text{eq.2}$$

$$p(y=0 \mid x) = 1 - p(y=1 \mid x) = 1 - h_{\theta}(x) \qquad \text{eq.3}$$

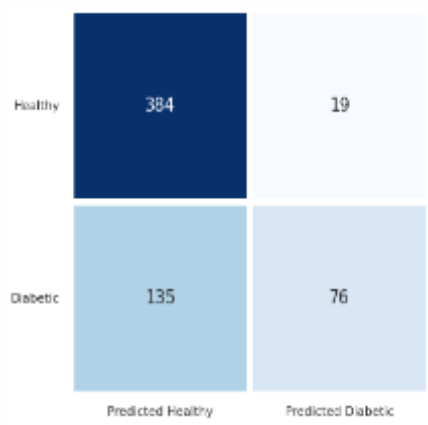

Fig. 4: The Confusion Matrix Logistic Regression.

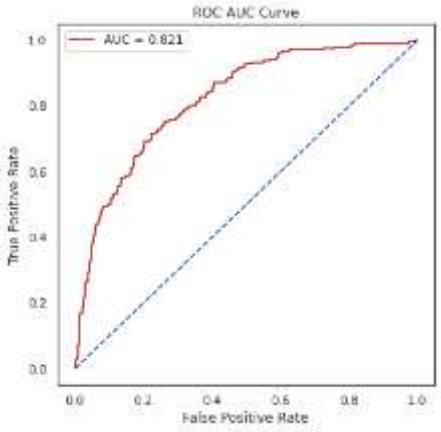

Fig. 5: ROC AUC Curve Logistic Regression.

Table 3. Classification report for Testing Dataset.

Classifier	Accuracy	Precision	Recall	F1- Score	Roc-auc score
Logistic Regression	86%	0.95	0.72	0.82	82%

Decision Tree : Under supervised machine learning, the decision tree method is included for classification and regression learning issues with regression [4]. A decision tree is a type of supervised machine learning algorithm, so that the data set needs to be labeled. The classification is completed using the decision tree algorithm upon a set of instructions. A node in a decision tree will stand fora feature, the branch for a rule and a leaf node will stand for result. It can be visualized as a tree this structure offers more accuracy and stability. Using decision trees, it will forecast the type of objects in the test class [16]. The number of false negatives reduces in the decision tree, which is more accurate than the logistic regression shown in Figure 6. The testing accuracy of decision tree, which is 91.2%, increased by about 5% compared to logistic regression. The ROC AUC curve gain is 0.91% which is depicted in Figure 7. The testing dataset is presented in Table-4.

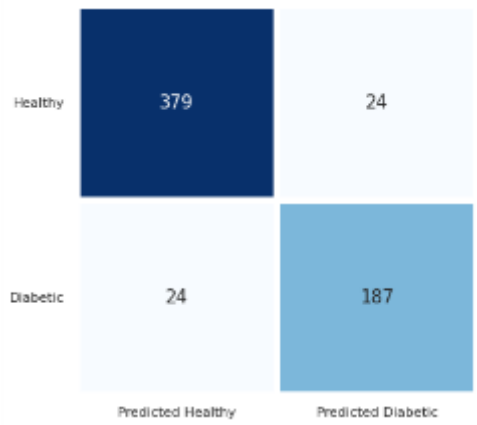

Fig. 6: The Confusion Matrix Decision Tree.

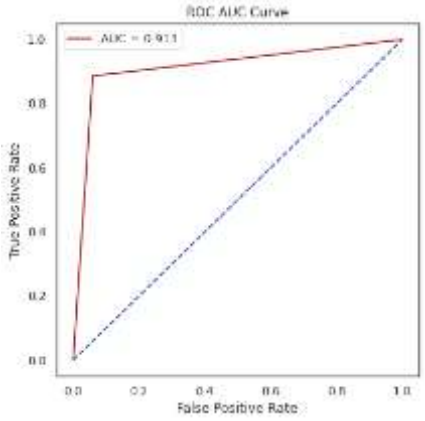

Fig. 7: ROC AUC Curve Decision Tree.

Table 4. Classification report for Testing Dataset

Classifier	Accuracy	Precision	Recall	F1- Score	Roc-auc score
Decision Tree	91%	0.89	0.89	0.89	92%

Gradient Boosting: Another ensemble learning technique is boosting. The weighting of the samples used to train each decision tree in order to reduce variance [17]. Figure 8 shows the amount of true positive are 382 and false positive are 21. Again, false negative 18 shows how many positive classes are incorrectly classified and true negative 193 shows how many negative classes are correctly classified. The area under the ROC curve (AUC) score is 0.96% which is shown in Figure 9. The accuracy of gradient boost is 93%. Compared to the decision tree, boosting accuracy improves by about 1.8%. The testing dataset is presented in Table-5.

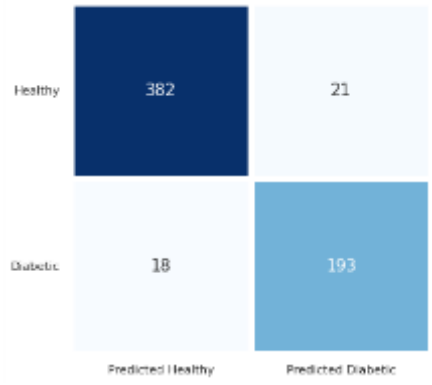

Fig. 8: The Confusion Matrix Gradient Boost.

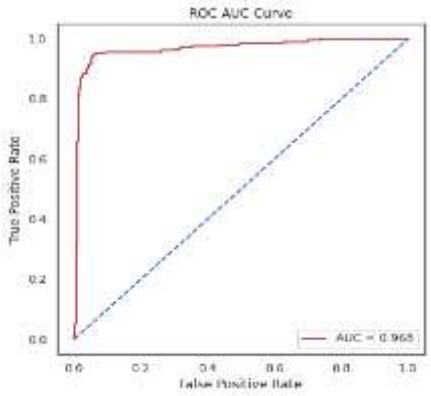

Fig. 9: ROC AUC Curve Gradient Boost.

Table 5: Classification report for Testing Dataset.

Classifier	Accuracy	Precision	Recall	F1- Score	Roc-auc score
Gradient Boosting	93%	0.90	0.91	0.91	96%

Random Forest : Random Forest is an integrated machine learning model. Both classification and regression issues are addressed by it. It is an ensemble model, which implies that it employs a combination of machine learning techniques to improve its performance in comparison to other techniques [18]. Random forest generates various decision trees by randomly selecting a component from the training data set. Confusion matrix show true positive are 391 and false positive 12 are incorrectly classified. Then, the false negative 19 people are incorrectly classified and true negative 192are correctly classified shown in Figure 10. The area under the ROC curve (AUC) score is 0.98% results summarized in Figure 11.The accuracy of random forest is 95%. The testing dataset is presented in Table-6.

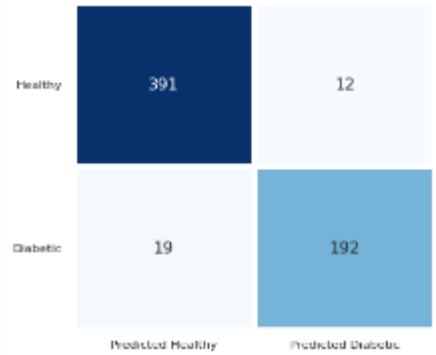

Fig. 10: The Confusion Matrix Random Forest.

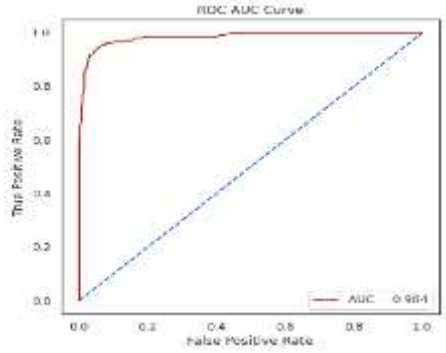

Fig. 11: ROC AUC Curve Random Forest.

Table 6. Classification report for Testing Dataset.

Classifier	Accuracy	Precision	Recall	F1- Score	Roc-auc score
Random Forest	95%	0.94	0.91	0.93	98.77%

Deep Neural Network : Neural networks are made up of layers of interconnected nodes, much like how the human brain is comprised of neurons. Every node in a layer is linked to other nodes situated in adjacent layers. The depth of the network is determined by the quantity of layers it encompasses [19]. We have selected a neural network with three hidden layers and 64, 32, and 16 neurons, respectively. For the diabetes prediction, we test various hidden layers and different neurons in various levels. We get the best outcome when the hidden layer is 3. The input layer is 8 and the output layer is 1. Rectified Linear Unit (ReLU) Activation Function used with 1000 epoch to get more accurate result. The train accuracy of deep neural network is around 98.5% and test accuracy 94.2% shown in Figure 12. The testing dataset is presented in Table-7.

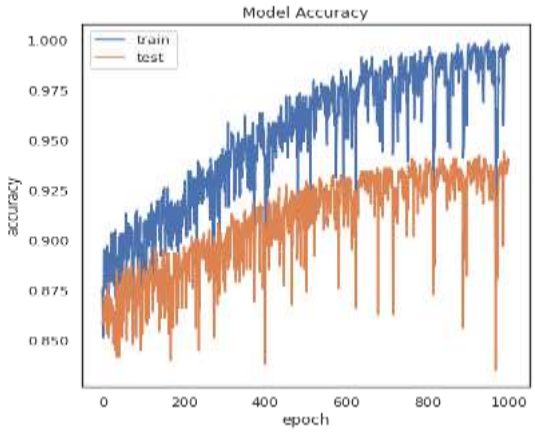

Fig. 12: The training and testing accuracy **.**

Table 7. Classification report for Testing Dataset.

Classifier	Accuracy
Deep Neural Network	94%

Result Analysis: Through the use of supervised machine learning techniques, several models have been created. Python is used to develop these models. The training and testing portions of the dataset are separated. We used 80% of the data for the model's training and the remaining 20% for the model's testing. As shown in Table-8 random forest algorithm shows 95% accuracy than the other algorithm. Here deep neural network shows 94% accuracy which is almost approximate to the random forest algorithm. Naïve bias algorithm shows the lowest accuracy like 76%. From the Table-9 it can be seen that this work shows a greater accuracy using random forest algorithm than the recent notable works.

Table 8. Result Analysis.

Classifier	Roc-auc score	Precision	Recall	F1- Score	Accuracy
Random Forest	98.77%	0.94	0.91	0.93	95%
Deep Neural Network					94%
Gradient Boosting	96%	0.90	0.91	0.91	93%
Decision Tree	92%	0.89	0.89	0.89	91%
Logistic Regression	82%	0.95	0.72	0.82	86%
Naive Bayes	81%	0.76	0.77	0.76	76%

Table 9. Comparison of result with the other works.

Ref.	Best Algorithm	Maximum Accuracy	Dataset
[20]	Logistics Regression	82.7 %	PIDD
[21]	Random Forest	93.75 %	PIDD
[22]	Extreme Gradient Boosting	79.22 %	PIDD
[23]	Extreme Gradient Boosting	77 %	PIDD
This Work	Random Forest	95 %	PIDD

Conclusion: Doctors will be able to diagnose patients correctly and provide them with prompt treatment with the aid of a reliable diabetes prediction model. To study the factors that affect diabetes, we execute expressive statistics on a dataset for diabetes risk prediction. Five machine learning models, including logistic regression, random forest, boosting, neural network and decision tree are used to create our diabetes prediction models. Five performance metrics—accuracy, recall, precision, f1-score, and ROC-AUC curve—are used. From the above algorithm it is found that random forest performs better than other. This algorithm achieved 95% accuracy, 0.94% precision, 0.93% f1-score, 0.91% recall and 98.77% ROC-AUC score. In future the accuracy of these algorithms can be experimented with models that have higher learning and adaptive capabilities and employ a large range of datasets.

Acknowledgement: We would like to express our sincere gratitude to all individuals and institutions who contributed to the success of this research.

References:

[1] D. A. Hasan, S. R. Zeebaree, M. A. Sadeeq, H. M. Shukur, R. R. Zebari, and A. H. Alkhayyat, 'Machine Learning-based Diabetic Retinopathy Early Detection and Classification Systems-A Survey', 1st Babylon International Conference on Information Technology and Science (BICITS), (2021), 16–21.
[2] V. Jaiswal, A. Negi, and T. Pal, 'A review on current advances in machine learning based diabetes prediction', Prim. Care Diabetes, 15 (2021) 435–443.
[3] A. Choudhury and D. Gupta, 'A survey on medical diagnosis of diabetes using machine learning techniques', Recent Developments in Machine Learning and Data Analytics: IC3(2018), Springer, (2019), 67–78.
[4] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, 'Early detection of type 2 diabetes mellitus using machine learning-based prediction models', Sci. Rep., 10 (2020) 11981–11991.
[5] N. El_Jerjawi and S. Abu-Naser, 'Diabetes Prediction Using Artificial Neural Network', J. Adv. Sci., 124 (2018) 1–10.
[6] Md. A. R. Refat, M. A. Amin, C. Kaushal, Mst. N. Yeasmin, and M. K. Islam, 'A Comparative analysis of Early Stage Diabetes Prediction using Machine Learning and Deep learning Approach', International Conference on Signal Processing, Computing and Control (ISPCC), (2021) 1–7.
[7] H. Xu et al., 'A Deep Learning Model Incorporating Knowledge Representation Vectors and Its Application in Diabetes Prediction', Dis. Markers, 2022 (20222) 1–17.

[8] H. Gupta, H. Varshney, T. K. Sharma, N. Pachauri, and O. P. Verma, 'Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction', Complex Intell. Syst., 8 (2022) 3073–3087.
[9] R. Akula, N. Nguyen, and I. Garibay, 'Supervised Machine Learning based Ensemble Model for Accurate Prediction of Type 2 Diabetes', SoutheastCon, Huntsville, AL, USA, (2019) 1–8.
[10] M. K. Dharani, R. Thamilselvan, D. Komarasamy, U. V, S. G, and S. M, 'Diabetes Prediction using Machine Learning Classification Algorithms', International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India: IEEE, (2022) 264–269.
[11] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, 'A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques', 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey: IEEE, (2019) 1–4.
[12] N. Sneha and T. Gangil, 'Analysis of diabetes mellitus for early prediction using optimal features selection', J. Big Data, 6 (2019) 1–13.
[13] D. Sisodia and D. S. Sisodia, 'Prediction of Diabetes using Classification Algorithms', Procedia Comput. Sci., 132 (2018) 1578–1585.
[14] K. Shah, R. Punjabi, P. Shah, and M. Rao, 'Real Time Diabetes Prediction using Naïve Bayes Classifier on Big Data of Healthcare', Int. Res. Jounral Eng. Technol., 7 (2020) 102–107.
[15] C. Zhu, C. U. Idemudia, and W. Feng, 'Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques', Inform. Med. Unlocked, 17 (2019) 100179–100189.
[16] H. Elaidi, Y. Elhaddar, Z. Benabbou, and H. Abbar, 'An idea of a clustering algorithm using support vector machines based on binary decision tree', International Conference on Intelligent Systems and Computer Vision (ISCV), (2018) 1–5.
[17] A.-Z. Sultan Bin Habib, T. Tasnim, and Md. M. Billah, 'A Study on Coronary Disease Prediction Using Boosting-based Ensemble Machine Learning Approaches', 2nd International Conference on Innovation in Engineering and Technology (ICIET), Dhaka, Bangladesh: IEEE, (2019) 1–6.
[18] Md. T. Islam, M. Raihan, F. Farzana, N. Aktar, P. Ghosh, and S. Kabiraj, 'Typical and Non-Typical Diabetes Disease Prediction using Random Forest Algorithm', 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India: IEEE ,(2020) 1–6.
[19] P. Pathak and A. Elchouemi, 'Predicting Early Phase of Type 2 Diabetic by Deep Learning', 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), Sydney, Australia: IEEE, (2020) 1–10.
[20] Chilupuri Anusha, 'A Machine Learning Approach for Prediction of Diabetes Mellitus', Int. J. Emerg. Trends Eng. Res., 11 (2023) 207–213.
[21] S. M. Ganie and M. B. Malik, 'Comparative analysis of various supervised machine learning algorithms for the early prediction of type-II diabetes mellitus', Int. J. Med. Eng. Inform., 1 (2021) 1–10.
[22] S. Kartal, 'Prediction of Diabetes Mellitus Using Tree-Based Machine Learning Algorithms: A Comparative Analysis' 3rd Global Conference on Engineering Research, (2023) 1–6.
[23] N. Sulayman, 'Predicting Type 2 Diabetes Mellitus using Machine Learning Algorithms', Int. J. Environ. Res. Public Health. 18 (2022) 3317–3327.