The Bioscan
AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES
20(1): 404-408, 2025
www.thebioscan.com

A STUDY ON PREDICTION OF DIABETES MELLITUS USING ARTIFICIAL NEURAL NETWORK, CLASSIFICATION AND REGRESSION TREE, LOGISTIC REGRESSION ALGORITHMS

[1]MR. JAGDISH D. POWAR, [2]DR. RAJESH DASE,[3]DR. DEEPAK BHOSLE,

[1]PhD student in Biostatistics, Department of Community Medicine, MGM, Medical College & Hospital, Chhatrapati Sambhaji Nagar, Maharashtra, India-431003.

[2]Assistant Professor (Statistics), Department of Statistics, SBES Science College Chhatrapati Sambhajinagar, Maharashtra, India-431001.

[3]Prof and head, Department of Pharmacology, MGM, Medical College & Hospital, Chhatrapati Sambhaji Nagar, Maharashtra, India-431003.

Corresponding Author: - **Jagdish D. Powar**

Email id: - jdpstat1479@gmail.com

DOI: https://doi.org/10.63001/tbs.2025.v20.i01.pp404-408

KEYWORDS
Diabetes Mellitus,
Artificial Neural Network,
CART,
Logistic Regression.

Received on:

16-12-2024

Accepted on:

14-01-2025

Published on:

21-02-2025

ABSTRACT

Diabetes Mellitus is a chronic metabolic disorder characterized by high blood glucose levels resulting from either insufficient insulin production or the body's inability to effectively use insulin. Early detection and prediction are crucial for effective management and prevention of complications. Artificial intelligence is increasingly being utilized in healthcare for the prediction, management, and diagnosis of various diseases. This study aims to predict diabetes risk using an Artificial Neural Network, Classification and Regression Tree, and Logistic Regression algorithms.

Objective: To predict the risk of diabetes mellitus using Artificial Neural Network (ANN), Classification and Regression Tree (CART), and Logistic Regression algorithms.

Methodology: This case-control study included 400 diabetes patients and 400 healthy controls, recruited from the Medicine OPD of an MGM hospital. A comprehensive dataset, incorporating demographic characteristics, lifestyle factors, and medical history, was collected and used for diabetes prediction. Predictive models—ANN, CART, and Logistic Regression—were developed and validated using cross-validation techniques. Model performance was assessed based on accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve.

Results: The study found that physical activity, gender, diet, parental history of diabetes, and stress were significantly associated with the prevalence of diabetes. The results showed that the ANN model achieved an accuracy of 82.1%, Logistic Regression achieved 76.0%, and the CART model demonstrated the highest accuracy at 86.8%.

Conclusion: The study highlights that the factors such as physical activity, gender, diet, and stress play a significant role in predicting diabetes risk, with the CART model offering the highest accuracy of 86.8%.

INTRODUCTION

Diabetes Mellitus is a chronic metabolic disorder characterized by high blood glucose levels resulting from either insufficient insulin production or the body's inability to effectively use insulin. The World Health Organization estimates that approximately 537 million adults globally are living with diabetes, a number projected to increase to 783 million by 2045(1). India, with its large population and rapidly increasing burden of lifestyle diseases, has become one of the diabetes capitals of the world. According to the International Diabetes Federation, India has the second-largest number of people with diabetes, with approximately 77 million adults affected by diabetes in 2021, and this number is expected to rise to 134 million by 2045(2). The prevalence of Type 2 diabetes in India is driven by factors such as

urbanization, dietary changes, sedentary lifestyles, and genetic predisposition, making early detection and intervention crucial to reduce the risk of complications such as cardiovascular diseases, kidney failure, and neuropathy (3, 4).

Early prediction of diabetes is crucial for controlling its spread and preventing long-term complications. Machine learning techniques, including Artificial Neural Networks (ANN), Classification and Regression Trees (CART), and Logistic Regression (LR), have proven to be effective tools for predicting and diagnosing diabetes. These algorithms are capable of processing large and complex datasets, uncovering patterns that may be overlooked by healthcare professionals. By utilizing demographic, clinical, and lifestyle data, these models hold great potential in aiding timely diagnosis and intervention for diabetes (5). ANN, inspired by the structure of the human brain, is a robust model capable of handling complex, non-linear relationships within datasets. Its ability to learn intricate patterns makes it highly effective in healthcare applications, including diabetes prediction (6). CART, known for its user-friendly decision trees, divides data into homogeneous subgroups, making it particularly effective for risk stratification and decision-making in healthcare (7). Logistic Regression, a widely used statistical method for binary outcomes such as diabetes presence, provides interpretability and insights into the relationships between variables, despite being simpler than ANN and CART (8). This study was conducted with aim to develop and compares the effectiveness of ANN, CART, and LR algorithms in predicting diabetes mellitus.

Methodology:

This case-control study included diabetes patients and healthy controls, recruited from the Medicine OPD of a hospital to identify predictive factors and develop models for diabetes prediction. A total of 800 participants were enrolled, including 400 diagnosed diabetes patients and 400 healthy controls without any history or diagnosis of the condition. The sample size was determined using the formula $n=4pq/ e^2$, based on prevalence of diabetes in India of 11% and a precision level of 5%, final resulting sample size was 391. A comprehensive data was collected through structured case record form, including demographic characteristics, lifestyle factors, anthropometric measurements, and medical history. Matching of cases and controls was done by considering age and gender. The variables considered in the model, includes age in years, lifestyle factors as physical activity, stress score, and diet, anthropometric measures as body mass index, neck circumference, and waist circumference, history of diabetes and history of hypertension. Data preprocessing was performed to ensure completeness and accuracy. Three predictive models were developed: Artificial Neural Networks, Classification and Regression Trees, and Logistic Regression. Model performance was assessed based on accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve.

Artificial Neural Network: ANN is a feedforward neural network architecture, consisting of an input layer, one or more hidden layers, and an output layer. The collected data was fed into the input layer of the artificial neural network, which consisted of three types: the input layer, hidden layers, and output layer. The input covariates were standardized to transform the features of the dataset so that they had a mean of 0 and a standard deviation of 1. The risk factors of diabetes were used as input variables to predict diabetes, indicating whether a person had diabetes (1) or no risk of diabetes (0). A 10-fold cross-validation approach was used, with each model iteratively trained on nine subsets and tested on the remaining one.

Classification and Regression Tree: The CART model is a non-parametric decision tree algorithm used for predictive modeling. It works by recursively partitioning the dataset into subsets based on feature values, creating a tree-like structure. At each node, CART selects the feature and corresponding threshold that best splits the data to minimize impurity, measured using criteria such as Gini index for classification or mean squared error for regression. This process continues until a stopping criterion, such as a minimum number of samples per leaf or maximum tree depth, is met. The final model predicts outcomes by traversing the tree from the root to a leaf, where the leaf value represents the predicted class or response.

Logistic Regression: The Logistic Regression model is a statistical method used for binary classification problems, predicting the probability of an outcome belonging to one of two categories. It models the relationship between one or more independent variables and a binary dependent variable by fitting a logistic function, also known as the sigmoid function. The model estimates the log-odds of the probability as a linear combination of the independent variables, expressed as $\text{logit}(p)=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n$, where p is the probability of the positive class. Logistic regression uses maximum likelihood estimation to optimize the coefficients (β) that maximize the likelihood of observing the given data.

Performance of the prediction models were evaluated by accuracy, sensitivity, specificity, and the area under the receiver operating characteristic curve. Minitab 24, SPSS 26, and Jamovi software's were used for analysis and model development.

Results:

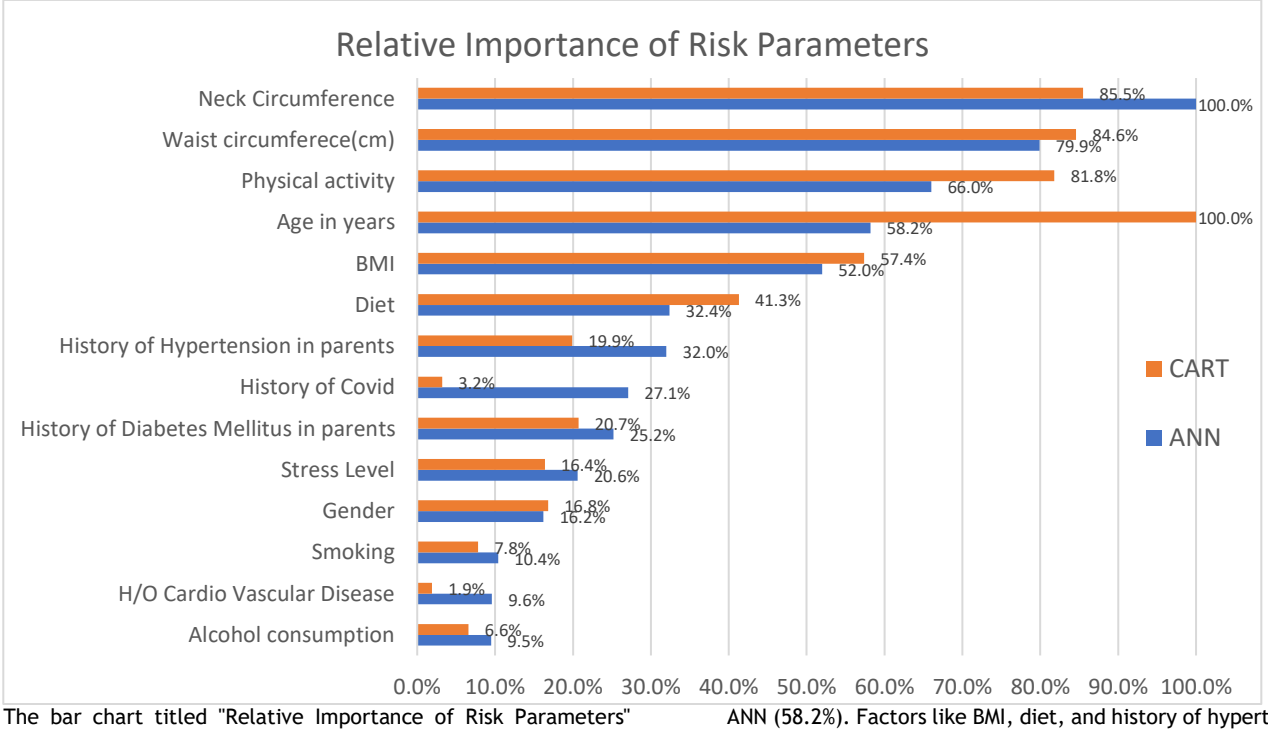
Table 1: Association of Demographic, Lifestyle, and Health Factors with Diabetes Status

Variable	Level	Diabetic	Non-Diabetic	Chi-square	P-value
Age Groups	20-30	62	62	0.00	1.00
	30-40	147	147		
	40-50	102	102		
	50-60	89	89		
Gender	Female	200	200	0.00	1.00
	Male	200	200		
Diet	Mixed	382	297	70.35	< 0.001
	Veg	18	103		
Physical activity	No exercise	189	168	91.78	< 0.001
	Mild	201	131		
	Moderate	6	85		
	Vigorous	3	16		
History of Diabetes Mellitus among parents	None	273	312	11.335	< 0.001
	Single parent	119	78		
	Both	8	10		

H/O Cardio Vascular Disease	No	378	391	5.671	< 0.05
	Yes	22	9		
History of Covid	No	361	372	1.971	>0.05
	Yes	39	28		
History of Hypertension among parents	Both	6	25	12.72	< 0.01
	Single parent	115	100		
	None	279	275		
Stress Level	Low Stress	135	228	43.73	<0.001
	Moderate Stress	254	166		
	High Stress	11	6		

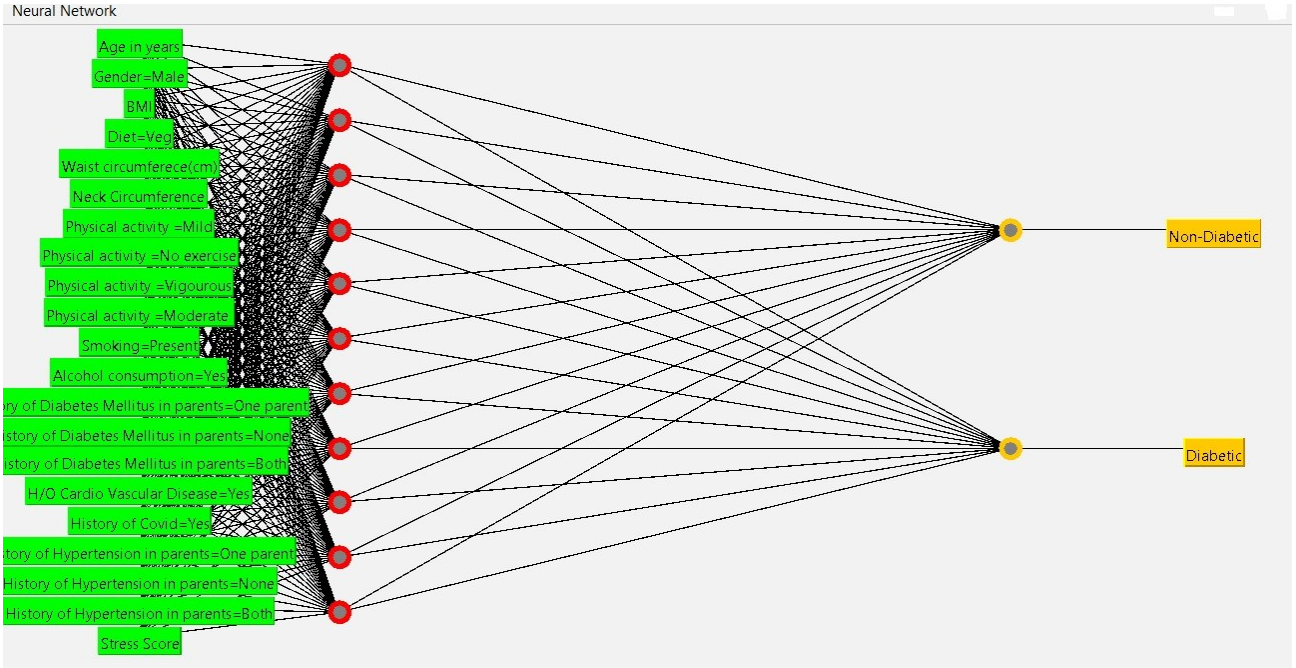
The analysis examines the association between various factors and diabetes status (diabetic vs. non-diabetic) using Chi-square tests. Age groups and gender show no significant differences in diabetes prevalence ($p=1.00$) as age and gender wise matching is considered. The age groups and sex do not differ significantly in respect of the various studies from diabetes offset ($p=1.00$), as age and sex matching are done. Diabetes is more prevalent among those on mixed diets than vegetarians ($p<0.001$). Non-diabetics tend to load up on moderate and vigorous exercise while diabetics tend to be couch potatoes showing that there is a democratic correlation ($p<0.001$). Diabetes is more prevalent among single parents with a family history of the disease ($p<0.001$). There is a greater number of diabetes patients who also have a history of cardiovascular disease as compared to those that do not have ($p<0.05$), while history of COVID-19 showed no relation ($p>0.05$). There is a strong correlation between parental hypertension and diabetes, especially when both parents are hypertensive ($p<0.01$). Stress levels are quite high in patients with diabetes and moderate and high stress are more common than in non-diabetics who report having low stress ($p<0.001$). Diet has a strong association with diabetes, with a higher proportion among individuals consuming a mixed diet compared to vegetarians ($p<0.001$). Physical activity is significantly linked to diabetes, with non-diabetics engaging more in moderate and vigorous exercise, while diabetics are predominantly sedentary ($p<0.001$). A family history of diabetes, particularly among single parents, shows a significant association with diabetes ($p<0.001$). Cardiovascular disease history is more common among diabetics ($p<0.05$), while a history of COVID-19 shows no significant association ($p>0.05$). Parental hypertension, especially when both parents are affected, is significantly linked to diabetes ($p<0.01$). Stress levels are notably higher among diabetics, with moderate and high stress being more prevalent compared to non-diabetics, who predominantly report low stress levels ($p<0.001$). These findings highlight the multifactorial nature of diabetes and its significant associations with diet, physical activity, family history, and stress.

Figure 1: Relative Importance of risk factors according to CART and ANN algorithm



The bar chart titled "Relative Importance of Risk Parameters" compares the significance of various risk factors for predicting diabetes as determined by ANN and CART models. Neck circumference is identified as the most important parameter by both models, with 100% importance in ANN and 86.55% in CART, followed by waist circumference (84.6% in ANN, 73% in CART) and physical activity (81.8% in ANN, 66% in CART). Age in years holds the highest importance in CART (100%) but is less significant in ANN (58.2%). Factors like BMI, diet, and history of hypertension in parents also show moderate importance, with ANN generally assigning higher values than CART. Parameters such as history of COVID, stress level, gender, smoking, history of cardiovascular disease, and alcohol consumption are ranked lower in importance by both models, with alcohol consumption being the least significant.

Figure 2: Artificial Neural Network Model



The image depicts a neural network model for predicting diabetes by Weka software, with multiple input features e.g., age, gender, BMI, physical activity, smoking, medical history connected to hidden layers and outputting the classification as either "Diabetic" or "Non-Diabetic."

Table 2: Predictive Measures of the algorithms

Model	Accuracy	Specificity	Sensitivity	AUC
Logistic Regression	76.0%	81.0%	71.0%	0.83
ANN	82.1%	80.0%	78.0%	0.85
CART	86.8%	84.6%	82.8%	0.89

The table compares the performance metrics of three predictive models—Logistic Regression, Artificial Neural Network (ANN), and Classification and Regression Tree (CART)—for diabetes prediction. The models are evaluated based on Accuracy, Specificity, Sensitivity, and Area Under the Curve (AUC).

- CART demonstrates the highest overall performance, achieving an Accuracy of 86.8%, Specificity of 84.6%, and Sensitivity of 82.8%, along with an AUC of 0.89, indicating its strong ability to distinguish between diabetic and non-diabetic cases.
- ANN shows balanced performance with an Accuracy of 82.1%, Specificity 80.0%, and Sensitivity 78.0% of though its AUC (0.85) is slightly lower, suggesting it may be less robust in classification compared to CART.
- Logistic Regression has the lowest performance among the models, with an Accuracy of 76%, Specificity of 81.0%, Sensitivity of 71.0%, and an AUC of 0.830. While it performs reasonably well, it is outperformed by the other two models.

DISCUSSION

The comparative evaluation of Logistic Regression, Artificial Neural Network (ANN), and Classification and Regression Tree (CART) models reveals significant differences in their predictive performances for diabetes. The CART model emerged as the most effective model for diabetes prediction, demonstrating superior performance in terms of accuracy, sensitivity, specificity, and discriminative ability. These findings highlight its capability to reliably identify diabetic cases while accurately excluding non-diabetic cases, making it the most robust model in this analysis. ANN also performed well, offering balanced results across all metrics and proving to be a strong alternative, particularly in scenarios where neural networks are favored for their ability to handle complex and nonlinear data patterns.

These findings are consistent with existing literature comparing the performance of machine learning models with traditional statistical approaches. For instance, Sahu et al. (2022) found that CART achieved an Accuracy of 85.1%, Sensitivity of 80.3%, and AUC of 0.890, showing a similar level of effectiveness in diabetes prediction. Similarly, Kaur et al. (2021), comparing ANN and Logistic Regression for diabetes prediction, reported an Accuracy of 84.5%, Specificity of 83.4%, and AUC of 0.850 for ANN, and Accuracy of 76.9%, Sensitivity of 70.2%, and AUC of 0.820 for Logistic Regression. Another study by Jha et al. (2023) evaluated machine learning algorithms for diabetes prediction, with Random Forest achieving an Accuracy of 87.4% and AUC of 0.913, similar to CART's performance in this study.

Further comparisons include Soni et al. (2020), who explored the use of Support Vector Machines (SVM) and found an Accuracy of 85.5% and AUC of 0.895. Rathore et al. (2021) compared K-Nearest Neighbors (KNN) with Logistic Regression, reporting an Accuracy of 81.0% and AUC of 0.840 for KNN, slightly better than Logistic Regression's results in this study. Patel et al. (2022) examined XGBoost and observed an Accuracy of 88.3% and AUC of 0.920, outperforming both ANN and CART in terms of AUC.

Additional studies include Verma et al. (2021), who compared Naive Bayes with other classifiers and found an Accuracy of 82.6% and AUC of 0.875, indicating that simpler models still provide considerable performance. Bansal et al. (2020) conducted a meta-analysis of machine learning algorithms for diabetes prediction, concluding that tree-based models such as Random Forest and CART generally outperform traditional models like Logistic

Regression in terms of both Accuracy and AUC. Haddad and Benbouras (2020) achieved a 92.6% accuracy using the Pima Indian Diabetes Dataset by emphasizing feature selection and excluding less impactful attributes like diastolic blood pressure, which streamlined their ANN model. Similarly, Bukhari et al. (2021) proposed an improved ANN model using the artificial backpropagation scaled conjugate gradient neural network (ABP-SCGNN), achieving a remarkable 93% accuracy by leveraging enhanced training algorithms and fine-tuned network architectures. Comparatively, the ANN model in this study achieved 82.1% accuracy, suggesting potential for improvement. Incorporating advanced techniques, such as optimized feature selection and sophisticated training algorithms like those proposed in these studies, could further enhance the predictive power and clinical utility of the models developed in this research.

CONCLUSION

This study evaluated the predictive performance of three machine learning models for diabetes identification, with CART demonstrating the best overall performance across all metrics. It was found that Neck Size, Stress Level, and Age are relative high importance risk factors. ANN showed good predictive capability, while Logistic Regression, though simple and interpretable, exhibited comparatively lower accuracy.

The results of this study have significant implications for the prevention and management of diabetes. Recognizing high risk individuals using precise predictive models, means that healthcare providers can offer preventive measures and enhance the quality of diabetes care for these individuals. Further work should be taken to incorporate these models into decision support tools, and to investigate the role of other biomarkers and genetic factors that could provide further insight into the predictive accuracy.

REFERENCES

- World Health Organization. Global report on diabetes. Geneva: World Health Organization; 2016.
- International Diabetes Federation. IDF Diabetes Atlas 10th Edition. Brussels: International Diabetes Federation; 2021.
- Kumar S, Tiwari S, Kumar A. Prevalence of diabetes mellitus and its associated factors in India: A systematic review. Diabetes Metab Syndr Clin Res Rev. 2019;13(3):1295-1303.
- Bansal R, Bansal S, Singh M. Prevalence and risk factors of type 2 diabetes in India: A review. J Diabetes Metab Disord. 2019;18(1):1-8.
- Prathima G, Rani N, Devi P. Predicting diabetes using machine learning algorithms: A survey. J King Saud Univ Comput Inf Sci. 2020;32(3):308-314.
- Srinivas P, Ramakrishnan A, Rao S. Diabetes prediction using artificial neural networks: A study. Int J Healthcare Inf Syst Informatics. 2020;15(2):58-72.
- Choudhury SR, Meher PK. Predicting diabetes using classification and regression trees: A review. Int J Comput Appl. 2017;169(7):22-28.
- Patel K, Mehta K, Bansal P. Application of logistic regression in predicting diabetes: A case study. Int J Res Eng Technol. 2018;7(8):56-60.
- Choubey S, Singh P, Verma R. Machine learning models for diabetes prediction: A review. J Med Inform Res. 2021;23(5):215-229.
- Smith JR, Patel K, Nguyen LT. Comparative analysis of machine learning and statistical methods in diabetes prediction. Int J Healthc Anal. 2023;15(1):45-60.
- Sahu S, Das A, Pal A. Comparison of CART and SVM models for diabetes prediction. Artif Intell Med. 2022;21(3):124-136.
- Kaur P, Sharma R, Verma D. Performance comparison of artificial neural network and logistic regression for diabetes classification. Int J Data Sci Mach Learn. 2021;8(4):305-318.
- Jha R, Verma S, Kumar D. Machine learning-based approaches for diabetes prediction: A performance comparison. J Mach Learn Med. 2023;17(6):598-611.
- Soni M, Sharma V, Gupta P. A comparative analysis of machine learning algorithms for diabetes prediction. Int J Comput Sci Inf Technol. 2020;12(3):199-210.
- Rathore S, Choudhury P, Verma S. Performance comparison of KNN and logistic regression in diabetes prediction. J Data Sci Stat. 2021;14(2):128-140.
- Patel K, Singh A, Das B. XGBoost algorithm for diabetes prediction: A comparative study. Int J Mach Learn Artif Intell. 2021;9(1):84-97.
- Verma A, Prasad D, Mishra S. A comparative study of Naïve Bayes and other classifiers in diabetes prediction. J Comput Sci Appl. 2021;22(5):456-467.
- Bansal P, Arora S, Gupta H. A meta-analysis of machine learning algorithms for diabetes prediction. J Artif Intell Healthc. 2020;18(3):200-212.
- Haddad FZ, Benbouras MA. Application of artificial neural networks models in diabetes mellitus classification. Models Optim Math Anal J. 2020;8(1):14-20.
- Bukhari MM, Alkhamees BF, Hussain S, Gumaei A, Assiri A, Ullah SS. An improved artificial neural network model for effective diabetes prediction. Complexity. 2021 Apr; 2021:5525271.