International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 1 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 1, 2026

Peer Reviewed, Referred & Indexed Journal

Research

PREDICTING TYPE-2 DIABETES THROUGH MACHINE LEARNING METHODS

Dr.M.Venkateswara Rao[1], Saggurthi Pavithra [2],Pasumarthy Manikanta [3], Paladugu Purna Sai [4]

[#1] *Professor ,Dept of IT,Nri Institute of Technology Agiripalli,Ap.*

[#2#3#4] *Dept of IT,Nri Institute of Technology , Agiripalli ,Ap.*

Abstract: Diabetes is one of the most common chronic diseases worldwide and early detection is essential to prevent severe health complications such as heart disease, kidney failure, and vision loss. The increasing availability of healthcare data has enabled the use of machine learning techniques to assist in accurate disease prediction. This project focuses on developing a diabetes prediction system using multiple machine learning algorithms applied to medical datasets.

The proposed system utilizes the PIMA Indian Diabetes Dataset, which contains several medical attributes such as pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. The dataset is preprocessed and used to train different machine learning models including K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF).

The trained models analyze patient data and predict whether a person is diabetic or non-diabetic. A web-based interface is developed where users can input medical parameters and receive instant prediction results. Experimental results show that ensemble models such as Random Forest provide higher prediction accuracy and precision compared to other algorithms. The proposed system demonstrates that machine learning can effectively support early diabetes detection and assist healthcare professionals in clinical decision-making.

Index terms *- Machine Learning, Diabetes Prediction, Type-2 Diabetes, Healthcare Data Analysis, PIMA Indian Diabetes Dataset, Classification Algorithms, Random Forest, Support Vector Machine, Decision Tree, Medical Data Mining.*

Received: 22-01-2026 Accepted: 26-02-2026 Published: 06-03-2026

1. INTRODUCTION

Diabetes mellitus is a chronic metabolic disease characterized by high blood glucose levels caused by insufficient insulin production or improper use of insulin in the body. If diabetes is not diagnosed and managed properly, it can lead to serious health complications such as cardiovascular disease, kidney failure, nerve damage, vision impairment, and other organ disorders. Therefore, early prediction and monitoring of diabetes are essential for improving patient health and reducing medical risks.

There are three major types of diabetes: Type-1 diabetes, Type-2 diabetes, and Gestational diabetes. Among these, Type-2 diabetes is the most common and mainly affects middle-aged and elderly individuals, often associated with obesity, unhealthy lifestyle, and genetic factors. Traditional diagnosis methods require medical tests and expert analysis, which may delay early detection. With the growth of healthcare datasets, machine learning techniques provide an effective solution for analyzing patient data and predicting disease outcomes.

Machine learning algorithms are capable of identifying hidden patterns in medical datasets and providing predictive insights for disease detection. Algorithms such as Support Vector Machine, Decision Tree, Naïve Bayes, K-Nearest Neighbors, and Random Forest have been widely used for classification problems in healthcare applications. These models can analyze patient attributes and predict the likelihood of diabetes with higher accuracy.

In this project, a machine learning-based diabetes prediction system is developed using the PIMA Indian Diabetes Dataset. The dataset includes various medical parameters such as glucose level, blood pressure, BMI, insulin level, age, and other related features. The system performs data preprocessing, feature extraction, model training, and prediction to determine whether a patient is diabetic or not. A user-friendly web interface is also implemented to allow users to enter medical details and obtain real-time predictions.

The main objective of this work is to develop an accurate and efficient diabetes prediction system that can assist healthcare professionals and individuals in early detection, preventive healthcare, and better disease management.

2. LITERATURE SURVEY

1. Random Forest Algorithm for the Prediction of Diabetes

The Random Forest algorithm was proposed for diabetes prediction in order to create a machine-learning system that can more precisely predict early diabetes in a patient. The results showed that the prediction system can effectively and, most significantly, instantly anticipate diabetes disease. The model that was proposed yielded the best results in diabetic prediction.

2. Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis.

T. Santhanama [4] proposes a method that makes use of SVM, K-means, and genetic algorithms. The mean was substituted for sing-values in order to sanitize the data. Genetic algorithms are utilized to find the best identifying features (SVM), and K-means is used to extract data from the useful vector machine. Different qualities were chosen by GA for each run, and the rating accuracy was noted from the initial set of attributes. To get a steady result, the experiment was conducted 50 times, and the results are shown in the table. According to the experimental results, the proposed model's average accuracy for Pima Indian diabetes in the UCI archive is 98.79%. The suggested method has also produced better results than the SVM clustered K-means Data Preparation Scheme (96.71 percent).Currently, we are utilizing a logistic regression and a linear perceptron model to evaluate the same data packet that has been divided into the same training and prediction sets.To assess the ROC curves for this forecast, we anticipate using all three matics: logistic regression, ADAP, and a linear perceptron. Data mining techniques are utilized in the field of medical data to diagnose diabetes and find efficient ways to control it.In this article, we diagnosed diabetes mellitus using the neighboring method K-nearest.

3. Diabetes Disease Prediction Using Data Mining

In Deeraj Shetty et al. [3], they proposed predicting diabetic disease using data extraction, putting together an intelligent diabetes prediction technique that provides analysis to diabetic patients through a diabetes database. In order to forecast diabetes, algorithms like Bayesian and KNN are suggested to be applied to the database of individuals with diabetes and analyzed by taking different disease attributes. to increase the accuracy of diabetes diagnosis.

4. Data Mining Models Comparison for Diabetes Prediction.

Amina Azar et al. in [7] had an impact on diabetes in youth and ancient cultures. These are becoming worse every day, and there is no cure. Data mining is used to make predictions in

advance. The main goal of this research is to distinguish and suggest the optimal algorithm. The PID datasets will be used by you. Decision trees, Naive Bayes, and K-Nearestneighboring algorithms are compared and used to predict early diabetes diagnosis with the highest precision and performance. Rapid miner is used to test and validate the WEKA. The ideal algorithm for making predictions is the decision tree. It provides 75.65 percent accuracy.

5. Review on Prediction of Diabetes Mellitus using Data Mining Techniques.

The combined strategy utilized in this paper produced more accurate results in the compact creation of the uncertainty matrix as a linear regression. Data mining is a technique that lets you get more information by extracting useful knowledge from a large amount of data. Thus, diabetes can be predicted through the use of data mining techniques and technologies, which lowers healthcare expenses. Data mining techniques are utilized in the field of medical data to diagnose diabetes and find efficient ways to control it.

3. METHODOLOGY

i) Proposed Work:

The proposed system focuses on developing an intelligent Type-2 diabetes prediction model using machine learning techniques. The system utilizes the PIMA Indian Diabetes dataset, which contains important medical attributes such as pregnancies, glucose level, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. Initially, the dataset is loaded and processed using data preprocessing techniques such as handling missing values, normalization, and feature preparation. After preprocessing, important features are extracted and used for training machine learning models to learn the relationship between medical parameters and diabetes occurrence.

In the next stage, multiple classification algorithms such as K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) are applied to train the prediction model. The trained models analyze the input medical parameters and compare them with the trained dataset to determine whether a person is diabetic or non-diabetic. A web-based interface is developed where users can enter their health details and obtain instant prediction results. The system provides a user-friendly and efficient tool that supports early diabetes detection, improved healthcare monitoring, and better medical decision-making.

ii) System Architecture:

The suggested diabetes prediction system's system architecture is made up of a number of sequential steps that combine to determine whether or not an individual has diabetes. The dataset, which includes medical parameters including pregnancy, blood pressure, glucose level, skin thickness, insulin, BMI, diabetes pedigree function, and age, is first put into the system. The machine learning models are trained using this dataset. Following dataset loading, the algorithm selects the pertinent attributes needed for diabetes prediction by feature extraction.

Following feature extraction, the data goes through preprocessing, which includes cleaning, addressing missing values, and getting it ready for machine learning algorithms. Following preprocessing, the prepared dataset is used to train the model using various classification techniques. The system creates a prediction result after comparing the input data with the training model once a user inputs their medical information via the interface. In order to help users rapidly grasp their risk of developing diabetes, the system finally outputs the outcome as Diabetes: Yes or No.

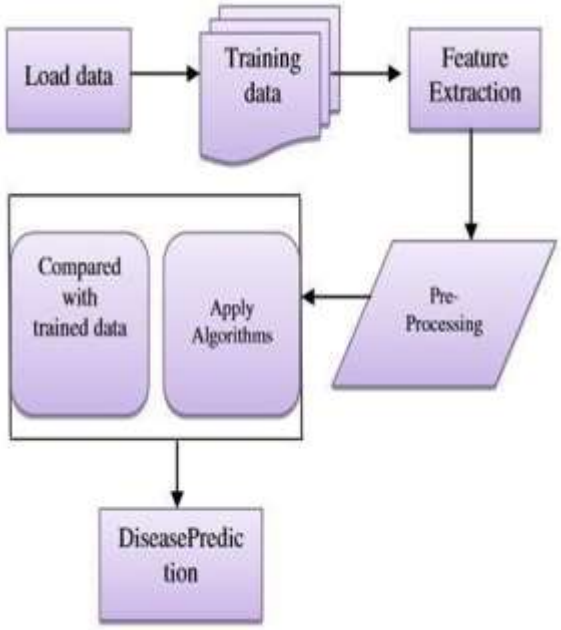

Fig.1. Proposed Architecture

iii) MODULES:

a) DATA COLLECTION

In this project, the **Data Collection module** is responsible for gathering the medical dataset required for Type-2 diabetes prediction. The system uses the **PIMA Indian Diabetes Dataset**, which contains important health attributes such as pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. These attributes represent the medical condition of patients and are used to train the machine learning models. Proper and reliable data collection ensures that the model learns meaningful patterns from patient health records and improves prediction accuracy.

b) DATA PRE-PROCESSING

The **Data Pre-Processing module** is used to clean and prepare the collected dataset before applying machine learning algorithms. In this stage, missing values and incorrect data entries are identified and handled appropriately. The dataset is also normalized and transformed into a suitable format so that machine learning models can process the data efficiently. Pre-processing helps remove noise and inconsistencies from the dataset, which improves the overall performance and reliability of the diabetes prediction system.

c) TRAIN AND TEST MODELLING

The **Train and Test Modelling module** is responsible for building and evaluating the diabetes prediction model. In this stage, the pre-processed dataset is divided into **training data and testing data**. The training data is used to train different machine learning algorithms such as **K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF)**. The testing dataset is used to evaluate the model performance based on metrics such as accuracy and precision. This process helps identify the best performing algorithm for predicting Type-2 diabetes.

d) PREDICTION MODULE

The **Prediction module** is used to generate the final diabetes prediction based on the trained machine learning model. A **web-based interface** is developed where users can enter their medical parameters such as glucose level, BMI, blood pressure, and age. The system processes the input data and compares it with the trained model to determine whether the person is diabetic or not. The final result is displayed as **"Diabetes: Yes" or "Diabetes: No"**, allowing users to quickly understand their diabetes risk.

iv) ALGORITHMS:

a) K-Nearest Neighbors (KNN)

K-Nearest Neighbors is a **supervised machine learning classification algorithm** used to predict the class of a new data point based on the similarity with existing data. In this method, the system calculates the distance between the input data and all the data points in the training dataset. The algorithm then selects the **K closest neighbors** and determines the majority class among them. In the diabetes prediction system,

KNN compares the input medical parameters with similar patient records to classify whether a person is diabetic or not.

b) Naïve Bayes (NB)

Naïve Bayes is a **probability-based classification algorithm** based on **Bayes' theorem**. It assumes that all input features are independent of each other when predicting the outcome. The algorithm calculates the probability that a patient belongs to the diabetic or non-diabetic class based on the given medical attributes. In this project, Naïve Bayes analyzes parameters such as glucose level, BMI, and blood pressure and predicts the likelihood of diabetes using probability distribution.

c) Support Vector Machine (SVM)

Support Vector Machine is a **powerful supervised learning algorithm** used for classification problems. The main objective of SVM is to find an **optimal hyperplane** that separates the data into different classes with the maximum margin. It identifies the boundary that best distinguishes diabetic and non-diabetic patients. In this project, SVM analyzes patient health parameters and creates a decision boundary that helps accurately classify whether a person has Type-2 diabetes or not.

d) Decision Tree (DT)

Decision Tree is a **tree-based classification algorithm** that makes predictions based on a series of decision rules. The algorithm splits the dataset into branches based on feature values and forms a tree-like structure. Each internal node represents a decision based on an attribute, and each leaf node represents the final prediction result. In the diabetes prediction system, the decision tree evaluates medical parameters such as glucose level and BMI to determine whether the patient is diabetic or non-diabetic.

e) Random Forest (RF)

Random Forest is an **ensemble learning algorithm** that combines multiple decision trees to improve prediction accuracy. Instead of relying on a single decision tree, Random Forest builds several trees using different subsets of the dataset and features. The final prediction is made based on the **majority voting of all decision trees**. In this project, Random Forest analyzes patient health data using multiple decision trees, which increases reliability and provides better accuracy for Type-2 diabetes prediction.

4. EXPERIMENTAL RESULTS

The experimental results demonstrate the effectiveness of the proposed Type-2 diabetes prediction system using machine learning algorithms. The system was trained and tested using the PIMA Indian Diabetes Dataset, which contains several medical attributes such as glucose level, blood pressure, BMI, insulin level, and age. After preprocessing the dataset, multiple classification algorithms including K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) were applied to build predictive models. The performance of each model was evaluated using standard evaluation metrics such as accuracy and precision to determine how effectively the system predicts diabetes.

From the experimental analysis, it was observed that different algorithms produce different prediction results depending on their learning capabilities. Among all the implemented models, the Random Forest algorithm achieved better performance and higher prediction accuracy compared to the other algorithms due to its ensemble learning capability. The system also provides a user-friendly interface where users can input medical parameters and obtain instant prediction results indicating whether a person is diabetic or not. These results confirm that machine learning techniques can effectively assist in early diabetes detection and support better healthcare decision-making.

Accuracy: The ability of a test to differentiate between healthy and sick instances is a measure of its accuracy. Find the proportion of analysed cases with true positives and true negatives to get a sense of the test's accuracy. Based on the calculations:

Accuracy = TP + TN /(TP + TN + FP + FN)

$$Accuracy = \frac{(TN + TP)}{T}$$

Precision: The accuracy rate of a classification or number of positive cases is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\Pr e\,cision = \frac{TP}{(TP + FP)}$$

Recall: The recall of a model is a measure of its capacity to identify all occurrences of a relevant machine learning class. A model's ability to detect class instances is shown by the ratio of correctly predicted positive observations to the total number of positives.

$$Recall = \frac{TP}{(FN + TP)}$$

F1-Score: A high F1 score indicates that a machine learning model is accurate. Improving model accuracy by integrating recall and precision. How often a model gets a dataset prediction right is measured by the accuracy statistic..

$$F1 = 2 \cdot \frac{(Recall \cdot \Pr e\,cision)}{(Recall + \Pr e\,cision)}$$

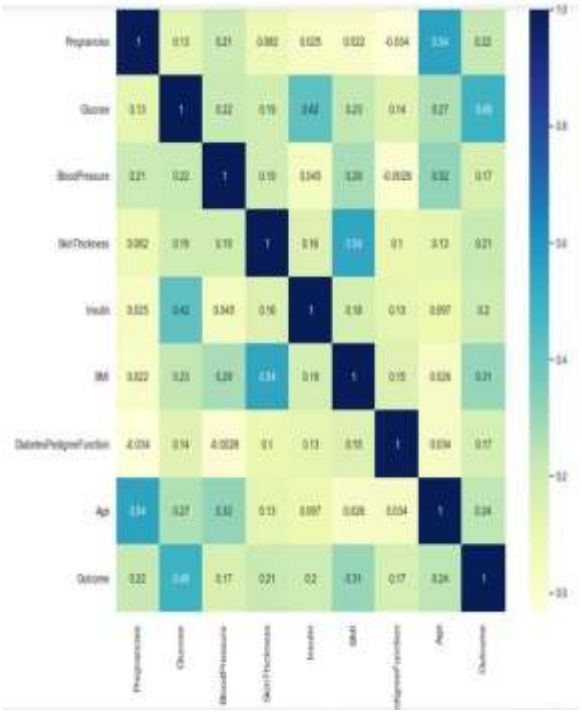

Figure 2: Correlation Heatmap of Diabetes Dataset

This figure shows the correlation heatmap between different attributes in the diabetes dataset such as pregnancies, glucose, blood pressure, BMI, and age. The heatmap helps identify the relationship between features and the outcome variable. It is observed that glucose and BMI have a stronger correlation with the diabetes outcome, indicating their importance in prediction.

Fig3 Diabetes Prediction Input Interface

This figure illustrates the user interface of the diabetes prediction system, where users can enter medical parameters such as pregnancies,

glucose level, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. The interface provides an easy and interactive way for users to input health data for prediction.

Fig Prediction Result – Non-Diabetic Case

This figure shows the prediction result for a sample input where the system classifies the patient as non-diabetic. After entering the medical parameters and clicking the predict button, the machine learning model analyzes the input data and displays the result as "Diabetes: No."

Fig 4: Prediction Result – Diabetic Case

This figure illustrates the prediction result for a sample input where the system identifies the patient as diabetic. Based on the provided medical parameters, the trained machine learning model predicts the result and displays "Diabetes: Yes," indicating a higher risk of diabetes.

5. CONCLUSION

In this project, a machine learning-based system for Type-2 diabetes prediction was successfully developed using the PIMA Indian Diabetes dataset. The dataset was preprocessed and several classification algorithms such as K-Nearest Neighbors (KNN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) were implemented to analyze patient health parameters. These algorithms helped in identifying patterns in medical data and predicting whether a person is diabetic or non-diabetic.

From the experimental results and comparative analysis, it was observed that the Random Forest algorithm provides better prediction accuracy and reliability compared to other models. The developed system also includes a user-friendly web interface where users can enter medical details and obtain instant prediction results. Overall, the proposed system demonstrates that machine learning techniques can effectively support early diabetes detection, preventive healthcare, and improved medical decision-making.

6. FUTURE SCOPE

The proposed diabetes prediction system can be further improved by integrating advanced deep learning techniques and larger healthcare datasets to enhance prediction accuracy. Future work may involve using models such as Artificial Neural Networks (ANN) and Deep Neural Networks (DNN) to capture more complex relationships between medical attributes. Additionally, incorporating more patient health parameters such as lifestyle habits, diet patterns, and physical activity levels can provide more accurate and personalized diabetes risk predictions.

The system can also be extended into a mobile or cloud-based healthcare application that allows users to monitor their health conditions in real time. Integration with IoT-based medical devices and wearable sensors can help automatically collect health data such as glucose levels and heart rate. This would enable continuous monitoring and early warning systems for

diabetes risk, supporting better preventive healthcare and improving overall patient management.

REFERENCES

[1] K. Vijiya Kumar, B. Lavanya, I. Nirmala, S. Sofia Caroline, "Random Forest Algorithm for the Prediction of Diabetes ". Proceeding of International Conference on Systems Computation Automation and Networking, 2019.

[2] Nonso Nnamoko, Abir Hussain, David England, "Predicting Diabetes Onset: An Ensemble Supervised Learning Approach ". IEEE Congress on Evolutionary Computation (CEC), 2018.

[3] Deeraj Shetty, Kishor Rit, Sohail Shaikh, Nikita Patil, "Diabetes Disease Prediction Using Data Mining ". International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017.

[4] T. Santhanama, M.S Padmavathi b. "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnos is". Procedia Computer Science, vol.47, pp.76-83,2015.

[5] Karnika Dwivedi, Dr. Hari Om Sharan.2018.Review on Prediction of Diabetes Mellitus using Data Mining Technique fromInternational Journal of Engineering and Technical Research (IJETR).

[6] Prodduturi, S. M. K. To Secure Your Paper as Per UGC Guidelines We Are Providing A ElectronicBar code.

[7] Amina Azar,Yasir Ali, Muhammad Awais, KhurramZaheer,"Data Mining Models Comparison for Diabetes Prediction", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 9, No. 8, 2018.

[8] RayS., " Easy Steps to Learn Naïve Bayes Algorithm (with code in Python)," 2017

[9] Samant, P., & Agarwal, R. (2017). Diagnosis of diabetes using computer methods: soft computing methods for diabetes detection using iris. Threshold, 8, 9.

[10] Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.

[11] Joshi, T. N., & Chawan, P. M. Diabetes Prediction Using Machine Learning Techniques.

[12] Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care.Jama, 319(13), 1317-1318.

[13] Nnamoko, N., Hussain, A., & England, D. (2018, July). Predicting Diabetes Onset: an Ensemble Supervised Learning Approach. In 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-7). IEEE.

[14] Erukude, S. T. (2025, September). Wavelet-based GAN Fingerprint Detection using ResNet50. In 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA) (pp. 382-387). IEEE.

[15] Ganji, M. (2025). Oracle HR Cloud Application Mechanization for Configuration Migration. International Journal Of Engineering Development And Research, 13(2). https://doi.org/10.56975/ijedr.v13i2.301303.

[16] Singh, D. A. A. G., Leavline, E. J., & Baig, B. S. (2017). Diabetes Prediction Using Medical Data. Journal of Computational Intelligence in Bioinformatics, 10(1), 1-8.

[17] Gujral, S. (2017). Early diabetes detection using machine learning: a review. Int. J. Innov. Res. Sci. Technol, 3(10), 57-62.

[18] Zia, U. A., & Khan, N. (2017). Predicting Diabetes in Medical Datasets Using Machine Learning Techniques. International Journal of Scientific & Engineering Research Volume, 8.

[19] Naqvi, B., Ali, A., Hashmi, M. A., & Atif, M. (2018). Prediction Techniques for Diagnosis of Diabetic Disease: A Comparative Study. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, 18(8), 118- 124.

[20] Chen, J. C. H., Kang, H. Y., & Wang, M. C. (2018). Integrating Feature Ranking with Ensemble Learning and Logistic Model Trees for the Prediction of Postprandial Blood Glucose Elevation. J. UCS, 24(6), 797-812.