IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

Diabetes Prediction Using Machine Learning

Sparsh Tyagi[1], Tushar Kumar[2], Varun Tyagi[3], Prince Tyagi[4], Santosh Kumar[5]
Student, Department of Information Technology[1–4]
Assistant Professor, Department of Information Technology[5]
Raj Kumar Goel Institute of Technology (RKGIT), Ghaziabad, Uttar Pradesh, India
26itvijsh@rkgit.edu.in[1], 26itpawar@rkgit.edu.in[2],
26itvisun@rkgit.edu.in[3], 26itdevce@rkgit.edu.in[4],
skwebfit@gmail.edu.in[5]

Abstract: *Diabetes is a chronic disease with the potential to cause a worldwide healthcare crisis. According to the International Diabetes Federation, 382 million people are currently living with diabetes across the world, and this number is expected to rise to 592 million by 2035. Diabetes occurs due to an increase in blood glucose levels, leading to symptoms such as frequent urination, increased thirst, and increased hunger. It is one of the leading causes of serious health complications, including blindness, kidney failure, amputations, heart failure, and stroke. When we eat, our body converts food into glucose, and the pancreas releases insulin to help glucose enter the cells for energy. However, in people with diabetes, this process does not function properly. The most common forms of the disease are Type 1 and Type 2 diabetes, along with other types such as gestational diabetes, which occurs during pregnancy.*

Machine learning is an emerging field in data science that focuses on enabling machines to learn from experience. The aim of this project is to develop a system capable of early prediction of diabetes in patients with higher accuracy by combining the results of different machine learning techniques. Various algorithms such as K-Nearest Neighbour, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree are used in this approach. The accuracy of each model is calculated, and the algorithm with the best performance is selected as the final model for predicting diabetes.

Keywords: cardiac risk analysis, learning-based classifiers, tree model, regression approach, neighbour method, clinical report processing

I. INTRODUCTION

Diabetes is a chronic metabolic disorder that poses a significant threat to global public health. According to the International Diabetes Federation, approximately 382 million individuals are currently affected worldwide, and this number is expected to increase to 592 million by 2035. The disease is primarily characterized by elevated blood glucose levels, which lead to symptoms such as frequent urination, excessive thirst, and increased hunger. If not properly managed, diabetes can result in severe complications, including blindness, kidney failure, cardiovascular diseases, amputations, and stroke. Under normal physiological conditions, the human body converts consumed food into glucose, and the pancreas secretes insulin to facilitate the absorption of glucose into cells for energy production. However, in diabetic patients, this mechanism is either impaired or insufficient. The most prevalent forms of the disease include Type 1 diabetes Type 2 diabetes, along with gestational diabetes, which occurs during pregnancy.

In recent years, advancements in data science have introduced machine learning (ML) as a powerful tool for medical diagnosis and prediction. Machine learning techniques enable systems to learn patterns from historical data and make accurate predictions without explicit programming. The primary objective of this study is to develop an efficient system for the early prediction of diabetes by leveraging multiple machine learning algorithms.

DOI: 10.48175/IJARSCT-34252

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 17, April 2026 Impact Factor: 8.2

Diabetes is a chronic metabolic disorder that poses a significant threat to global public health. According to the International Diabetes Federation, approximately 382 million individuals are currently affected worldwide, and this number is expected to increase to 592 million by 2035. The disease is primarily characterized by elevated blood glucose levels, which lead to symptoms such as frequent urination, excessive thirst, and increased hunger. If not properly managed, diabetes can result in severe complications, including blindness, kidney failure, cardiovascular diseases, amputations, and stroke. Under normal physiological conditions, the human body converts consumed food into glucose, and the pancreas secretes insulin to facilitate the absorption of glucose into cells for energy production. However, in diabetic patients, this mechanism is either impaired or insufficient. The most prevalent forms of the disease include Type 1 diabetes Type 2 diabetes, along with gestational diabetes, which occurs during pregnancy.

In recent years, advancements in data science have introduced machine learning (ML) as a powerful tool for medical diagnosis and prediction. Machine learning techniques enable systems to learn patterns from historical data and make accurate predictions without explicit programming. The primary objective of this study is to develop an efficient system for the early prediction of diabetes by leveraging multiple machine learning algorithms.

II. LITERATURE REVIEW

[1] conducted a study on the classification of diabetes using diverse datasets to determine whether a person is diabetic or not. The dataset was collected from a hospital data warehouse and consisted of 200 instances with nine attributes, categorized into two groups: blood tests and urine tests. The implementation was carried out using WEKA, and evaluation was performed using a 10-fold cross-validation approach, which is effective for small datasets. Various classification algorithms, including Naïve Bayes, J48, REP Tree, and Random Tree, were applied. The study concluded that the J48 algorithm achieved the highest accuracy of 60.2% among the compared models.

[2] aimed to develop an efficient method for early detection of diabetes by analysing patterns in medical data using classification techniques. The study utilized Decision Tree and Naïve Bayes algorithms on the PIMA dataset, with performance evaluated through cross-validation and a 70:30 data split. The results indicated that the J48 algorithm achieved an accuracy of 74.8%, while Naïve Bayes outperformed it with an accuracy of 79.5%, suggesting its effectiveness in diabetes prediction tasks.

[3] focused on evaluating and comparing the performance of multiple classification algorithms based on accuracy, sensitivity, and specificity. The study implemented algorithms such as JRIP, JGraft, and BayesNet using WEKA and compared the results with other tools like RapidMiner and MATLAB. The findings revealed that the JGraft algorithm achieved the highest accuracy of 81.3%, with a sensitivity of 59.7% and specificity of 81.4%. Additionally, the study concluded that WEKA performed better compared to MATLAB and RapidMiner for this task.

[4] investigated the application of the CART (Classification and Regression Tree) algorithm on a diabetes dataset. The study emphasized the issue of class imbalance, which is common in datasets with binary outcomes. To address this, a resampling technique was applied during the preprocessing stage to balance the dataset. The authors highlighted that handling class imbalance prior to model training significantly improves prediction accuracy and enhances the performance of the classification model.

III. PROPOSED METHODOLOGY

• In this section, we discuss the various classification techniques used in machine learning for the prediction of diabetes. The proposed methodology focuses on improving prediction accuracy by applying multiple machine learning algorithms and evaluating their performance. Five different classification methods are utilized in this study, and each method is analysed based on its effectiveness. The performance of these models is measured using accuracy metrics, and the most suitable model is selected for diabetes prediction.

• The dataset used in this study is obtained from the publicly available Kaggle repository:
https://www.kaggle.com/johndasilva/diabetes

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

• The diabetes dataset consists of 2000 patient records, each containing several medical attributes related to the diagnosis of diabetes. The primary objective of this dataset is to predict whether a patient is diabetic or non-diabetic based on these input features. The dataset includes important parameters such as glucose level, blood pressure, insulin, body mass index (BMI), age, and other relevant factors that contribute to accurate prediction.

➔ The dataset contains multiple instances with 9 input features.

➔ The target variable is "Outcome", which indicates:

• 0 – Non-diabetic patient

• 1 – Diabetic patient

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	2	138	62	35	0	33.6	0.127	47	1
1	0	84	82	31	125	38.2	0.233	23	0
2	0	145	0	0	0	44.2	0.630	31	1
3	0	135	68	42	250	42.3	0.365	24	1
4	1	139	62	41	480	40.7	0.536	21	0

➔ The diabetes data set consists of 2000 data points, with 9 features each.

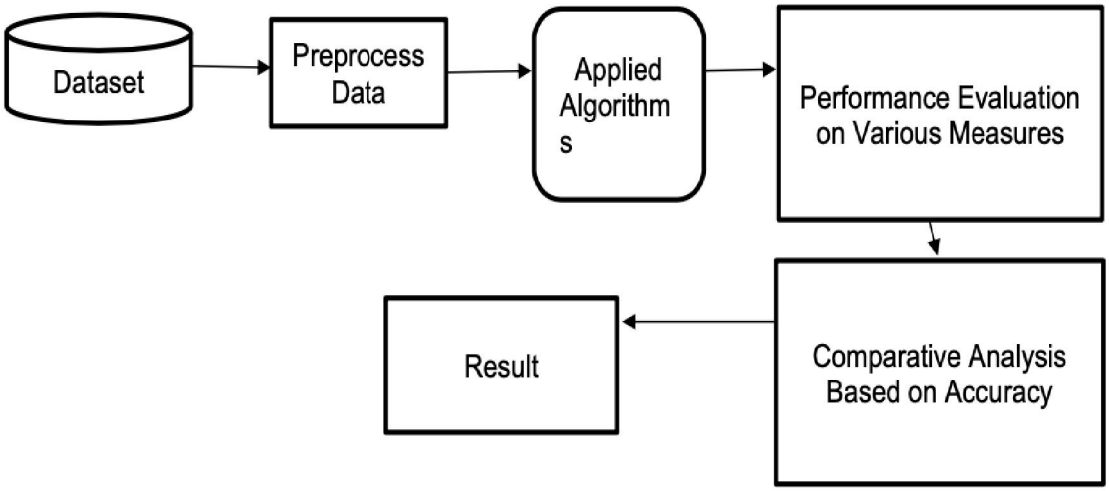
➔ "Outcome" is the feature we are going to predict, 0 means No diabetes, 1 means diabetes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               2000 non-null   int64
 1   Glucose                   2000 non-null   int64
 2   BloodPressure             2000 non-null   int64
 3   SkinThickness             2000 non-null   int64
 4   Insulin                   2000 non-null   int64
 5   BMI                       2000 non-null   float64
 6   DiabetesPedigreeFunction  2000 non-null   float64
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

➔ There is no null values in dataset.

DOI: 10.48175/IJARSCT-34252

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2



Proposed Model Diagram

IV. RESULTS AND DISCUSSION

Histogram:

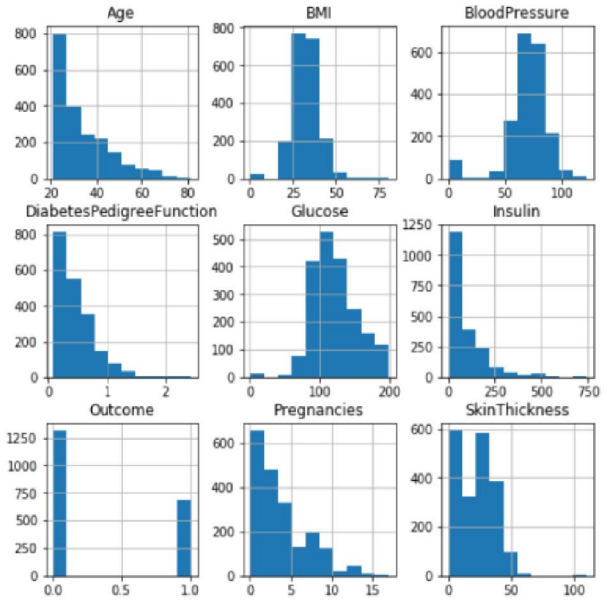


Fig. 3. Confusion Matrix

DOI: 10.48175/IJARSCT-34252

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

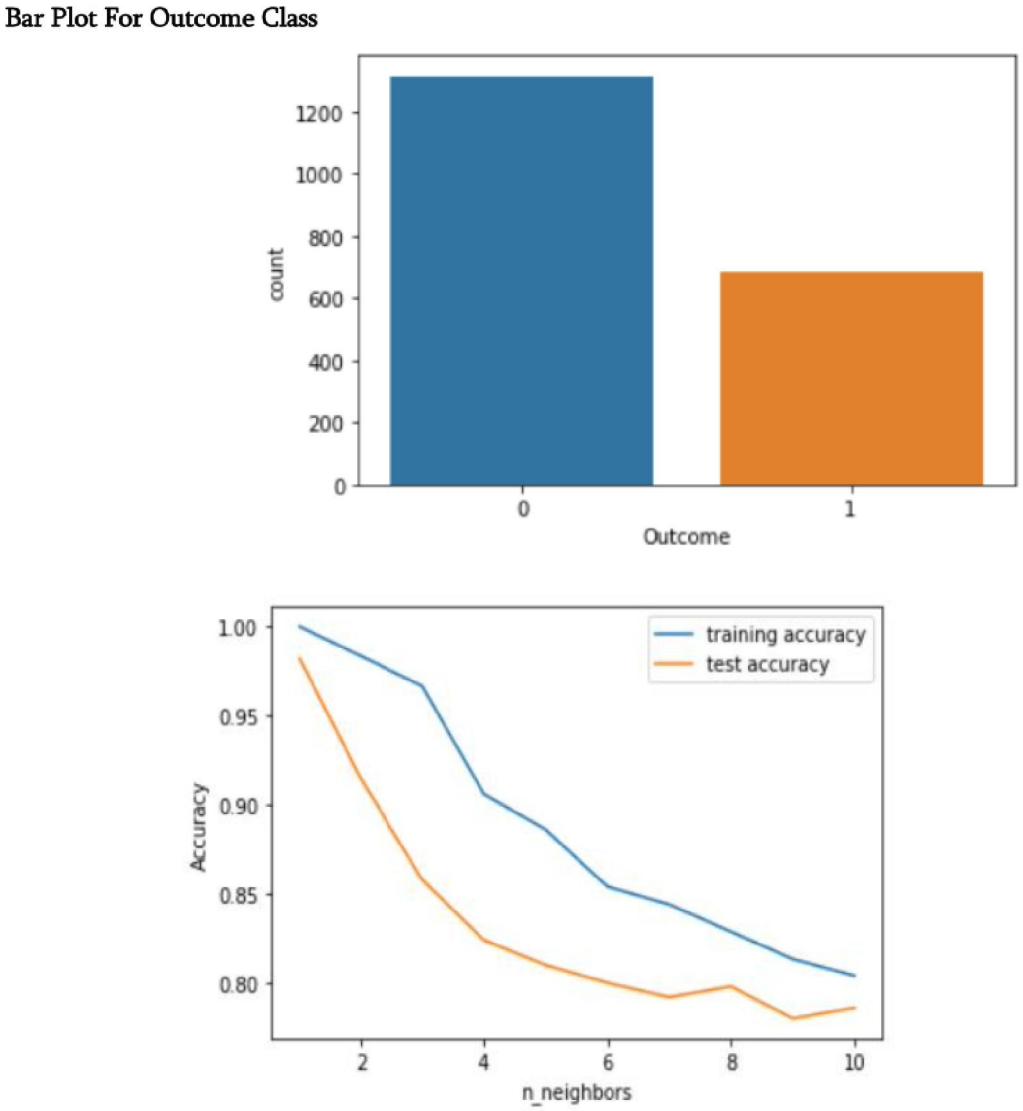


Fig. 4. ROC Curve

Let us examine the graphical plots representing the distribution of features and the target variable. These visualizations illustrate how each feature is spread across different ranges, which highlights the importance of feature scaling for improving model performance and ensuring uniform contribution of all attributes. The presence of discrete bars in the plots indicates that certain features behave as categorical variables. Such variables need to be properly handled and encoded before applying machine learning algorithms to ensure accurate predictions.

The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients.

k-Nearest Neighbours:

The k-NN algorithm is arguably the simplest machine learning algorithm. Building the model consists only of

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

storing the training data set. To make a prediction for a new data point, the algorithm finds the closest data points in the training data set, its "nearest neighbours."
First, let's investigate whether we can confirm the connection between model complexity and accuracy:

Training Accuracy	0.81
Testing Accuracy	0.78

The above plot shows the training and test set accuracy on the y-axis against the setting of neighbours on the x-axis. Considering if we choose one single nearest neighbour, the prediction on the training set is perfect. But when more neighbours are considered, the training accuracy drops, indicating that using the single nearest

Decision Tree:
Decision Tree is a supervised machine learning algorithm used for classification. It creates a tree-like structure where each node represents a feature, each branch represents a decision, and each leaf node represents a class label. The model splits the data based on important features to classify the data points. The complexity of the model can be controlled by adjusting parameters such as the maximum depth of the tree.

Training Accuracy	1.00
Testing Accuracy	0.99

The Decision Tree algorithm is easy to understand and interpret, making it highly useful for decision-making problems. It can handle both numerical and categorical data efficiently. However, the model may lead to overfitting if the tree becomes too deep. Therefore, techniques such as pruning are used to improve the model's generalization and performance.
The training accuracy of the model is 100%, while the testing accuracy is also high, indicating good performance and generalization capability of the model.
Decision Tree Performance Summary: The model shows strong performance with only a small gap between training and testing accuracy, indicating effective learning without significant overfitting. This suggests the Decision Tree has captured the underlying patterns in the dataset. Overall, the results indicate a reliable and efficient classification model.

Accuracy Comparison:

Algorithms	Training Accuracy	Testing Accuracy
k-Nearest Neighbors	81%	78%
Logistic Regression	78%	78%
Decision Tree	98%	99%
Random Forest	94%	97%
SVM	76%	77%

The above table shows the accuracy values for all five machine learning algorithms.

The table shows that Decision Tree algorithm gives the best accuracy with 98% training accuracy and 99% testing accuracy.

DOI: 10.48175/IJARSCT-34252

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 17, April 2026
Impact Factor: 8.2

V. CONCLUSION AND FUTURE WORK

One of the important real-world medical problems is the detection of diabetes at its early stage. In this study, systematic efforts are made in designing a system which results in the prediction of diabetes. During this work, five machine learning classification algorithms are studied and evaluated on various measures. Experiments are performed on john Diabetes Database. Experimental results determine the adequacy of the designed system with an achieved accuracy of 99% using Decision Tree algorithm. In future, the designed system with the used machine learning classification algorithms can be used to predict or diagnose other diseases. The work can be extended and improved for the automation of diabetes analysis including some other machine learning algorithms.

Furthermore, the system can be enhanced by incorporating larger and more diverse datasets to improve its accuracy and reliability. Integration of real-time data and advanced techniques such as deep learning can also be explored to achieve better performance. Additionally, developing a user-friendly interface can make the system more accessible for practical healthcare applications and assist medical professionals in decision-making.

REFERENCES

[1]. Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003.

[2]. Arora, R., Suman, 2012. Comparative Analysis of Classification Algorithms on Different Datasets using WEKA. International Journal of Computer Applications 54, 21–25. doi:10.5120/8626-2492.

[3]. Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[4]. Choubey, D.K., Paul, S., Kumar, S., Kumar, S., 2017. Classification of Pima Indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection, in: Communication and Computing Systems: Proceedings of the International Conference on Communication and Computing System (ICCCS 2016), pp. 451–455.

[5]. Dhomse Kanchan B., M.K.M., 2016. Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis, in: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication, IEEE. pp. 5–10.

[6]. Sharief, A.A., Sheta, A., 2014. Developing a Mathematical Model to Detect Diabetes Using Multigene Genetic Programming. International Journal of Advanced Research in Artificial Intelligence (IJARAI) 3, 54–59. doi:doi:10.14569/IJARAI.2014.031007.

[7]. Sisodia, D., Shrivastava, S.K., Jain, R.C., 2010. ISVM for face recognition. Proceedings - 2010 International Conference on Computational Intelligence and Communication Networks, CICN 2010, 554–559 doi:10.1109/CICN.2010.109.

[8]. Sisodia, D., Singh, L., Sisodia, S., 2014. Fast and Accurate Face Recognition Using SVM and DCT, in: Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, Springer. pp. 1027–1038.

[9]. https://www.kaggle.com/johndasilva/diabetes

[10]. Rani, A. S., & Jyothi, S. (2016, March). Performance analysis of classification algorithms under different datasets. In Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on (pp. 1584-1589). IEEE.