International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0372
e ISSN: 2584-2854
Volume: 04 Issue: 07
July 2026
Page No: 2593 - 2596

Diabetes Prediction Using Machine Learning Algorithm

Twarita Taru[1], Madanika G [2]
[1,2]Vijaya Vittala Institute of Technology, Bengaluru, Karnataka, 560077, and India
Emails: *tarutwarita15@gmail.com[1], madanikagopalkrishna@gmail.com[2]*

Abstract
In this study, we aimed to create a system that uses machine learning to detect and classify diabetes in an e-healthcare setting. We used Ensemble Decision Tree algorithms for selecting important features from a large set of data. Detecting diabetes accurately is a big challenge for researchers, especially in e-healthcare environments. Many existing systems have problems like slow processing and low accuracy. To fix these issues, we built a diabetes diagnosis system that includes data cleaning, feature selection, and classification. We tested the system using methods to check its effectiveness. We used a filter method based on the Decision Tree algorithm to choose the most important features. We also used two types of ensemble learning Decision Tree algorithms, Ada Boost and Random Forest, for feature selection and compared their performance with wrapper-based methods. The Decision Tree classifier was used to separate healthy individuals from those with diabetes. The results showed that using the selected features improved the model's classification performance and achieved the best accuracy. The system also performed better than previous methods due to the combination of different feature sets.
Keywords: *Ensemble learning Decision Tree algorithms, such as Ada Boost and Random Forest*

1. Introduction
In recent years, using machine learning (ML) in healthcare has opened new possibilities for predicting diseases and providing personalized care. One of the biggest challenges in healthcare is the early detection and control of chronic diseases, like diabetes. Diabetes has a big impact on individuals and the healthcare system, so it needs new solutions for early diagnosis and treatment. This study focused on using machine learning algorithms to predict the risk of diabetes. By analyzing health data and risk factors, ML models can help understand who is more likely to develop diabetes. This allows doctors to take action earlier to prevent or manage the disease. Machine learning is a method that trains computers to learn from data. Different ML techniques help build models for classification and prediction. This data can be useful in predicting diabetes. While many ML techniques can be used for prediction, it's hard to choose the best one. So, we used popular classification and ensemble methods on the dataset for prediction.

1.1.Importance of Diabetes Prediction with Machine Learning
Diabetes is a serious long-term condition that affects millions around the world. Detecting diabetes early is important for managing the disease and avoiding complications. Machine learning (ML) is a strong tool for predicting the risk of diabetes, helping healthcare providers take action early. Here, we discuss the benefits of using ML for diabetes prediction in reports.

- Improved Resource Allocation: Reports can leverage ML to highlight populations with a high prevalence of diabetes risk factors. This allows healthcare systems to target resources and outreach programs towards these areas, maximizing preventative efforts and improving overall population health.
- Reduced Healthcare Costs: Detecting diabetes early and preventing its development can lower the costs of treating the disease.
- Early Detection of Prediabetes: Machine learning can spot people with prediabetes, a condition that often has no symptoms but is a high risk for developing diabetes. Reports can highlight such cases, encouraging healthcare professionals to suggest lifestyle changes and monitor

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0372
e ISSN: 2584-2854
Volume: 04 Issue: 07
July 2026
Page No: 2593 - 2596

patients closely to prevent the disease from getting worse.

2. Method

In this section, we will talk about the different types of classifiers used in machine learning to predict diabetes. We will also explain our new approach for making predictions more accurate. In this study, five different methods were used. Each of these methods is described below. The results we got were the accuracy of the machine learning models. These models can then be used for prediction.

2.1. Dataset Description

The diabetes dataset was originally taken from https://www.kaggle.com/datasets/madanikag/diabetes-prediction.This dataset had 768 cases and was collected from the UCI repository, known as the Pima Indian Diabetes Dataset. This dataset has many features that describe 768 patients. The goal is to predict whether a patient has diabetes or not based on the features shown in Figure 1.

- The dataset had 2000 data points, each with nine features.
- The ninth feature is the classification variable, which tells us if the patient has diabetes or not.
- There are no missing values in the dataset.
- The "Outcome" is the feature we are trying to predict. A value of 0 means no diabetes, and 1 means the patient has 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   float6
 6   DiabetesPedigreeFunction  2000 non-null   float6
 7   Age                       2000 non-null   int64
 8   Outcome                   2000 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 140.8 KB
```

Figure 1 Dataset Description

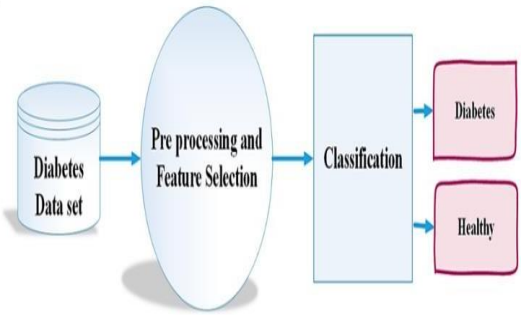

Figure 2 Proposed Model Diagram

2.2. Data Preprocessing

Data preprocessing is a very important step. Many healthcare-related datasets have missing values or other problems that can affect the data [1-3]. Preprocessing helps improve the quality and effectiveness of the data after it's processed. For using machine learning techniques, this step is necessary for getting accurate results and making good predictions. For the Pima Indian diabetes dataset, we did preprocessing in two steps shown in Figure 2.

- Missing Values removal

We removed all records that had zero as a value because having a zero is not realistic. By removing these, we get a better feature subset, which helps reduce the complexity of the data, making it faster to work with.

- Splitting of data

After cleaning the data, we split it into parts for training and testing the model. When we split the data, we use the training set to teach the algorithm, and we keep the test set aside [4-6]. This training process allows the model to learn the rules, algorithms, and values of the features in the training data. Normalizing the data is also done to bring all features to the same scale.

2.3. Apply Machine Learning

Once the data is ready, we use machine learning techniques. We applied different classification and ensemble methods to predict whether a patient has diabetes. These methods were used on the Pima Indian diabetes dataset. The main goal of using these techniques is to study their performance and accuracy and to identify which features are most important in making predictions.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0372
e ISSN: 2584-2854
Volume: 04 Issue: 07
July 2026
Page No: 2593 - 2596

3. Results and Discussion

3.1. Results

3.1.1. Correlation Matrix

Looking at the correlation matrix, it's clear that no single feature has a very strong link with the outcome. Some features have a negative link with the outcome, while others have a positive link shown in Figure 3.

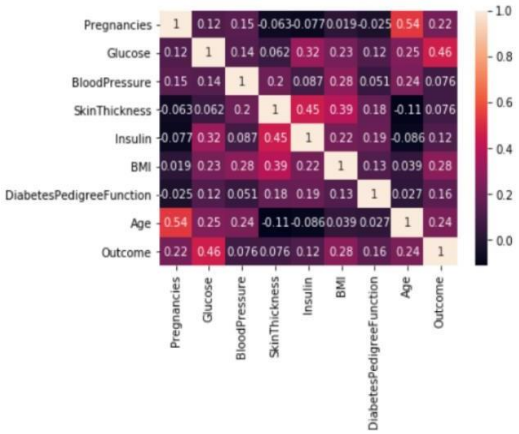

Figure 3 Correlation Matrix

3.1.2. Ensembling

Ensembling is a machine learning technique that uses multiple algorithms together to improve predictions. It gives better results than any single model because it reduces errors caused by noise, bias, and variance. Popular ensembling methods include bagging, boosting, AdaBoost, gradient boosting, voting, and averaging. In this study, we used bagging (Random Forest) and gradient boosting to predict diabetes.

3.1.3. Precision and recall

Precision measures how accurate positive predictions are, while recall measures the model's ability to find all relevant cases. High precision ensures our positive predictions are accurate, and high recall ensures we miss very few actual cases, which is important for early detection and treatment in diabetes shown in Figure 4.

Figure 4 Precision and Recall

3.1.4. Confusion Matrix

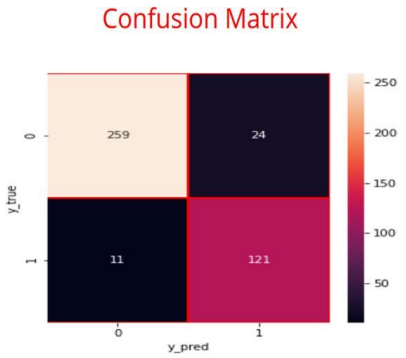

Figure 5 Confusion Matrics

A confusion matrix shows how well the model performs by comparing actual outcomes with the model's predictions. It divides the results into true positives, true negatives, false positives, and false negatives. This helps us evaluate the model's accuracy and figure out areas to improve in predicting diabetes risk.

3.1.5. Diabetes Prediction

Machine learning helps predict diabetes by analyzing different health data to assess a person's risk of developing diabetes. With automated algorithms and easy-to-use tools, it offers personalized risk assessments, integrates with current systems, and helps with early detection, thus improving healthcare shown in Figure 6.

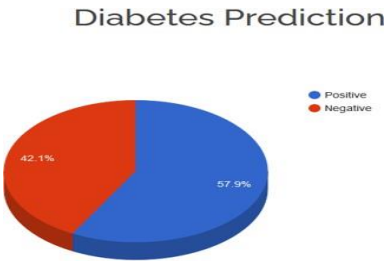

Figure 6 Diabetes Prediction

3.1.6. Model Building

This is the most important part, where we build a model to predict diabetes. In this study, we used several machine learning algorithms for diabetes prediction.

3.2.Procedure of Proposed Methodology

- Step 1: Import the needed libraries and the diabetes dataset.

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0372
e ISSN: 2584-2854
Volume: 04 Issue: 07
July 2026
Page No: 2593 - 2596

- Step 2: Preprocess the data to remove missing values.
- Step 3: Split the data into 80% for training and 20% for testing.
- Step 4: Choose machine learning algorithms, such as K-Nearest Neighbor, Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, and Gradient Boosting.
- Step 5: Build the classifier model for each of these algorithms using the training data.
- Step 6: Test each model using the test data.
- Step 7: Compare the results from each classifier to see which one performs best.
- Step 8: After looking at the results, we choose the best algorithm based on the measures we used.

Conclusion

Detecting diabetes early is a major real-world medical challenge . In this study, we developed a system to predict diabetes and evaluated five machine learning classification algorithms on different metrics. We used the Pima Indian Diabetes Database for experiments and achieved a high accuracy of 99% using the Decision Tree algorithm. In the future, this system can be used to predict or diagnose other diseases. The work can be improved and expanded to include more machine learning algorithms for better diabetes analysis.

Acknowledgements

We have completed this work under the mentorship of Prof. Sushma M. Patil (Project co - ordinator) Prof. Kavyashree J (Assistant Professor), Department of Artificial Intelligence and Data Science at Vijaya Vittala Institute of Technology, Bangalore. We would like to express our special thanks to both of our mentors for inspiring us to complete the work & write this paper. Without their active guidance, help, cooperation & encouragement, we would not have our headway in writing this paper. We are extremely thankful for their valuable guidance and support on completion of this paper. We extend our gratitude to “Vijaya Vittala Institute of Technology” for giving us this opportunity. We also acknowledge with a deep sense of reverence, our gratitude towards our parents and member of our family, who has always supported us morally as well as economically. Any omission in this brief acknowledgement does not mean lack of gratitude.

References

[1]. Amin Ul Haq, Jian Ping Li, Jalaluddin khan, Muhammad Hammad Memon, Shah Nazir, Sultan Ahmad, Ghufran Ahmad khan, Amjad Ali, “A New Intelligent Approach for Effective Recognition of Diabetes in the IoT E-HealthCare Environment”, Preprints 2020.

[2]. H. E. Massari, Z. Sabouri, S. Mhammedi and N. Gherabi, "Diabetes Prediction Using Machine Learning Algorithms and Ontology," in Journal of ICT Standardization, vol. 10, no. 2, pp. 319-337, 2022, doi: 10.13052/jicts2245-800X.10212.

[3]. Z. Ye, J. Wang, H. Hua, X. Zhou and Q. Li, "Precise Detection and Quantitative Prediction of Blood Glucose Level With an Electronic Nose System," in IEEE Sensors Journal, vol. 22, no. 13, pp. 12452-12459, 1 July1, 2022, doi: 10.1109/JSEN.2022.3178996.

[4]. S. K. Sharma et al., "A Diabetes Monitoring System and Health-Medical Service Composition Model in Cloud Environment," in IEEE Access, vol. 11, pp. 32804-32819, 2023, doi: 10.1109/ACCESS.2023.3258549.

[5]. C.kalaiselvi, G.M.Nasira, “A New Approach of Diagnosis of Diabetes and Prediction of Cancer using ANFIS”, IEEE Computing and Communicating Technologies, pp 188-190, 2014.

[6]. E. L. Litinskaia, N. A. Bazaev, K. V. Pozhar and V. M. Grinvald, “Methods for improving accuracy of non-invasive blood glucose detection via optical glucometer”, 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, (2017), pp. 47-49.