



Original Article

Diabetes Prediction Using Machine Learning

M. Jayanthi¹, Pranavi Surabhi²

¹ Dept of CSI, Mahatma Gandhi University, Nalgonda, Telangana-India

² MCA II Year Student, Mahatma Gandhi University Nalgonda

Manuscript ID:
IJERSD-2026-020409

ISSN: 3067-2325

Volume 2

Issue 4

Pp. 48-54

August 2026

Submitted: 08 July 2026

Revised: 19 July 2026

Accepted: 12 Aug. 2026

Published: 31 Aug. 2026

Correspondence Address:

M. Jayanthi
Dept of CSI, Mahatma Gandhi
University, Nalgonda, Telangana-
India
Email: jayanthimgu@gmail.com

Quick Response Code:



Web: <https://rlgjaar.com>

DOI: 10.5281/zenodo.23101820

DOI Link:
<https://doi.org/10.5281/zenodo.23101820>



Attribution-NonCommercial-
ShareAlike 4.0 International



Abstract

Diabetes is a long-term metabolic disorder that affects millions of people and can lead to serious health complications if it is not identified at an early stage. Predicting the disease in its initial stages allows patients to receive timely medical attention and helps reduce the risk of severe complications. In recent years, machine learning has become a useful approach for analyzing healthcare data and supporting disease prediction. In this study, a diabetes prediction model was developed using the Pima Indians Diabetes Dataset, which contains medical information from 768 patients. The dataset was preprocessed by handling missing values, standardizing the input features using StandardScaler, and splitting the data into training and testing sets. Five supervised machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were trained and evaluated to determine the most suitable model for prediction.

The performance of each algorithm was assessed using accuracy, precision, recall, and F1-score. Among the evaluated models, the Random Forest classifier produced the most balanced performance and achieved an accuracy of **75.3%**. Based on these results, the Random Forest model was selected for deployment. To make the prediction system easy to use, the trained model was integrated into a Streamlit web application that allows users to enter clinical information and obtain an instant prediction.

The proposed system demonstrates how machine learning can support early diabetes prediction and assist healthcare professionals in making timely decisions. It also highlights the potential of simple, data-driven tools to improve preventive healthcare and patient management.

Keywords— Diabetes Prediction, Machine Learning, Random Forest, StandardScaler, Healthcare, Pima Indians Diabetes Dataset, Streamlit.

Introduction

Diabetes is one of the most common chronic diseases affecting people around the world. It occurs when the body is unable to produce enough insulin or cannot effectively use the insulin it produces, resulting in increased blood glucose levels. If diabetes is not diagnosed and managed at an early stage, it may lead to serious health problems such as heart disease, kidney failure, nerve damage, and vision loss. Therefore, identifying individuals who are at risk before complications develop is an important step in improving healthcare outcomes. In recent years, healthcare organizations have generated large volumes of medical data through routine clinical examinations and patient records. Analyzing this information manually is often time-consuming and may not always reveal meaningful patterns. Machine learning provides an effective way to analyze such data, identify relationships among different clinical factors, and support disease prediction. These techniques can assist healthcare professionals by providing additional insights during the decision-making process. This study focuses on developing a machine learning model for predicting diabetes using the **Pima Indians Diabetes Dataset**. The dataset contains patient information such as the number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. Before training the models, the dataset was preprocessed by handling missing values, standardizing the features using **StandardScaler**, and dividing the data into training and testing sets.

Creative Commons

This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution NonCommercial-ShareAlike 4.0 International](https://creativecommons.org/licenses/by-nc-sa/4.0/). The Creative Commons Attribution license allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately credited.

How to cite this article:

Jayanthi, M., & Pranavi Surabhi. (2026). Diabetes Prediction Using Machine Learning. *International Journal of Engineering Research for Sustainable Development*, 2(4), 48–54. <https://doi.org/10.5281/zenodo.23101820>

Their performance was compared using accuracy, precision, recall, and F1-score to determine the most suitable model for diabetes prediction. Five supervised machine learning algorithms—**Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)**—were implemented and evaluated.

Among these algorithms, **Random Forest** provided the best overall performance and was selected for deployment. The final model was integrated into a **Streamlit-based web application**, allowing users to enter patient information and receive an instant prediction.

The main objective of this work is to develop a reliable and user-friendly system that can support the early prediction of diabetes using machine learning techniques. Although it is not intended to replace medical diagnosis, the proposed system can serve as a supportive tool for healthcare professionals and individuals by providing quick and data-driven predictions.

Research Objectives

- Analyze and preprocess diabetes dataset
- Implement ML models (LR, SVM, KNN, DT, RF, GB)
- Evaluate using accuracy, precision, recall, F1-score
- Identify best model
- Deploy using Streamlit
- Support early diagnosis

Methodology

The system follows a structured pipeline:

- Data collection from Pima dataset
- Data preprocessing (handling missing values, scaling)
- Train-test split (80%-20%)
- Model training using multiple algorithms
- Performance evaluation
- Selection of Random Forest
- Deployment using Streamlit

Data Collection

The study uses the **Pima Indians Diabetes Dataset**; a publicly available dataset widely used for diabetes prediction research. It contains **768 patient records** with **eight clinical attributes**: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. The target variable, **Outcome**, indicates whether a patient has diabetes (1) or does not have diabetes (0).

Data Preprocessing

Data preprocessing was performed to improve the quality of the dataset before training the models. Missing or invalid values in selected attributes were handled appropriately. The dataset was then standardized using **StandardScaler**, which transforms numerical features to a common scale with a mean of zero and a standard deviation of one. This step improves the performance of algorithms that are sensitive to differences in feature scales, such as SVM and KNN. Finally, the dataset was divided into training and testing sets for model development and evaluation.

Machine Learning Models

Five supervised machine learning algorithms were implemented and compared in this study.

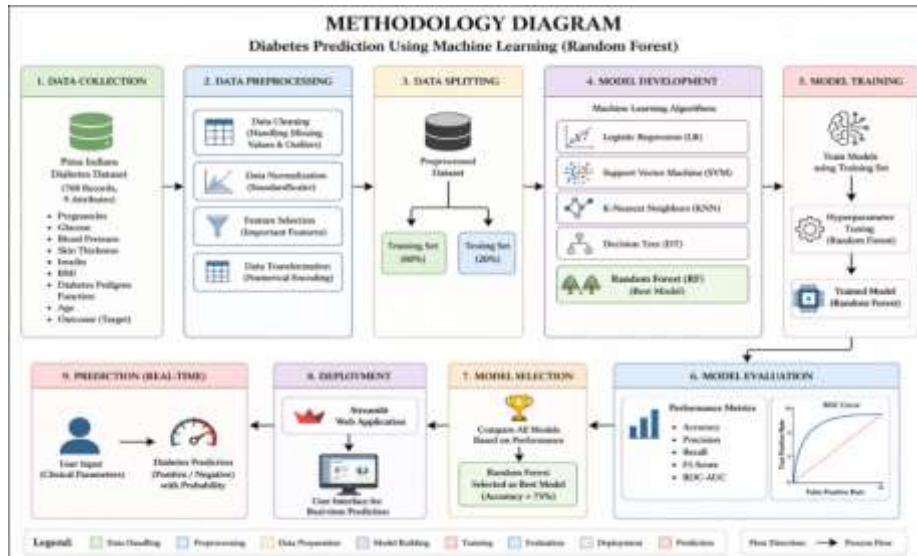
- **Logistic Regression**: A statistical classification algorithm used to predict whether a patient is diabetic or non-diabetic.
- **Decision Tree**: A tree-based model that makes predictions by splitting the dataset into decision nodes based on feature values.
- **Random Forest**: An ensemble learning method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.
- **Support Vector Machine (SVM)**: A classification algorithm that identifies the optimal decision boundary for separating different classes.
- **K-Nearest Neighbors (KNN)**: An instance-based algorithm that classifies a patient according to the majority class of its nearest neighboring data points.

Model Evaluation

The performance of all models was evaluated using **accuracy, precision, recall, and F1-score**. These evaluation metrics provide a comprehensive assessment of each model's classification performance and help identify the most reliable algorithm for diabetes prediction.

Model Deployment

Based on the experimental results, the **Random Forest** model was selected for deployment because it provided the most balanced overall performance. The trained model was integrated into a **Streamlit web application**, allowing users to enter patient health information and receive an instant prediction regarding diabetes risk. This web-based system demonstrates the practical application of machine learning in supporting early diabetes prediction.



Literature Survey

Diabetes has become one of the fastest-growing chronic diseases worldwide, encouraging researchers to develop reliable prediction systems using machine learning. Early prediction helps healthcare professionals identify individuals who are at risk and initiate treatment before serious complications develop. As a result, many studies have investigated different machine learning algorithms to improve the accuracy of diabetes prediction.

Several researchers have used the **Pima Indians Diabetes Dataset** because it contains important clinical attributes such as glucose level, blood pressure, BMI, insulin, age, and diabetes pedigree function. This dataset has become a common benchmark for evaluating diabetes prediction models and comparing the performance of different classification algorithms.

Previous studies have shown that **Logistic Regression** is widely used because it is simple to implement and produces easily interpretable results. It performs well when the relationship between the input variables and the target variable is relatively straightforward. However, its performance may decrease when the data contains complex patterns that cannot be represented by a linear model.

Researchers have also explored **Decision Tree** and **Random Forest** algorithms for diabetes prediction. Decision Trees are easy to understand and visualize, but they may overfit the training data if the tree becomes too complex. Random Forest addresses this limitation by combining multiple decision trees, resulting in more stable predictions and better generalization. Because of its robustness, Random Forest has been reported as one of the most reliable models for healthcare prediction tasks.

Other machine learning methods, including **Support Vector Machine (SVM)** and **K-Nearest Neighbors (KNN)**, have also been applied to diabetes prediction. SVM is effective for classification problems with clear decision boundaries, while KNN classifies new samples by comparing them with similar patient records. Although both methods can provide satisfactory results, their performance is influenced by feature scaling and parameter selection.

Most existing studies agree that data preprocessing plays an important role in improving prediction performance. Handling missing values, standardizing numerical features, and selecting appropriate evaluation metrics help improve the reliability of machine learning models. Researchers have also emphasized that comparing multiple algorithms provides a better understanding of their strengths and limitations instead of relying on a single model.

Based on these findings, this research evaluates **Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)** using the **Pima Indians Diabetes Dataset**. The models are compared using accuracy, precision, recall, and F1-score, and the best-performing model is deployed through a **Streamlit** application to provide a practical and user-friendly diabetes prediction system.

Experimental Setup and Implementation

A. System Environment

The proposed diabetes prediction system was developed using the Python programming language. The implementation was carried out in **Google Colab**, which provided a suitable environment for data analysis, model training, and testing. Libraries such as **Pandas**, **NumPy**, **Scikit-learn**, **Matplotlib**, and **Seaborn** were used for data preprocessing, visualization, and machine learning. The final prediction model was deployed using **Streamlit**, allowing users to interact with the system through a simple web interface.

Dataset Description

The experiments were performed using the **Pima Indians Diabetes Dataset**, which is widely used for diabetes prediction research. The dataset contains **768 patient records** with eight clinical features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function, and Age. The target variable indicates whether a patient is diabetic or non-diabetic.

Data Preprocessing

Before training the models, the dataset was preprocessed to improve data quality. Missing or invalid values in selected attributes were handled appropriately. The numerical features were then standardized using **StandardScaler** so that all features were on a similar scale. Finally, the dataset was divided into **80% training data** and **20% testing data** for model development and evaluation.

Model Training

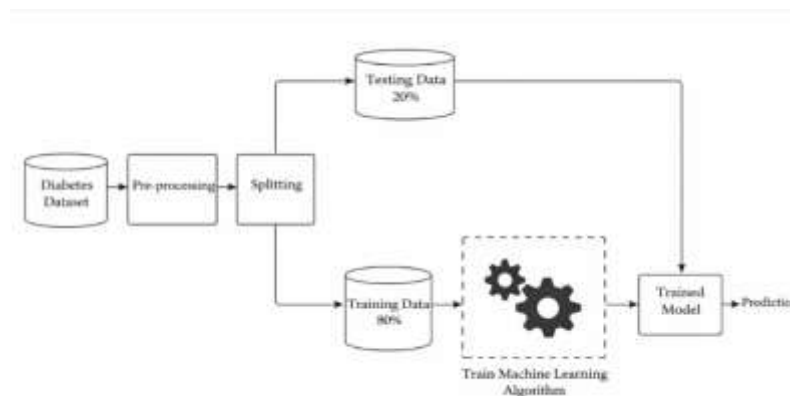
Five supervised machine learning algorithms were implemented in this study: **Logistic Regression**, **Decision Tree**, **Random Forest**, **Support Vector Machine (SVM)**, and **K-Nearest Neighbors (KNN)**. Each model was trained using the training dataset and tested on the unseen data to evaluate its prediction performance.

Performance Evaluation

The trained models were evaluated using **accuracy**, **precision**, **recall**, and **F1-score**. These evaluation metrics helped compare the performance of each algorithm and identify the model that produced the most reliable predictions for diabetes classification.

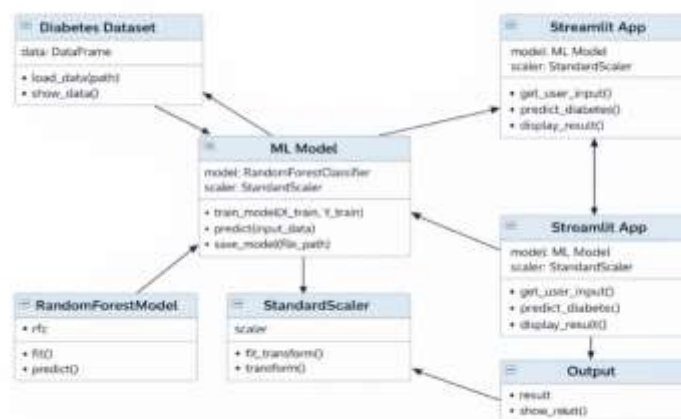
System Deployment

After comparing all models, **Random Forest** was selected for deployment because it provided the most balanced overall performance. The trained model was integrated into a **Streamlit web application**, where users can enter patient details and receive an instant diabetes prediction. This deployment demonstrates the practical use of the proposed system in supporting early diabetes risk assessment.



The class diagram of the diabetes prediction system illustrates the structural design and interaction between different components involved in the prediction process. The system consists of key classes such as the Diabetes Dataset, ML Model, RandomForestModel, StandardScaler, Streamlit App, and Output. The Diabetes Dataset class is responsible for loading and managing the dataset, while the ML Model class acts as the core component that handles model training, prediction, and saving operations using the Random Forest algorithm. The RandomForestModel class defines the underlying classifier with functions such as `fit()` and `predict()`. The StandardScaler class performs data normalization to ensure consistent input for the model. The Streamlit App class manages user interaction by collecting input, invoking the prediction function, and displaying results. Finally, the Output class presents the prediction outcome to the user. The relationships between these classes ensure smooth data flow from input to prediction, making the system modular, scalable, and efficient for real-time diabetes prediction.

Class Diagram of Diabetes Prediction System

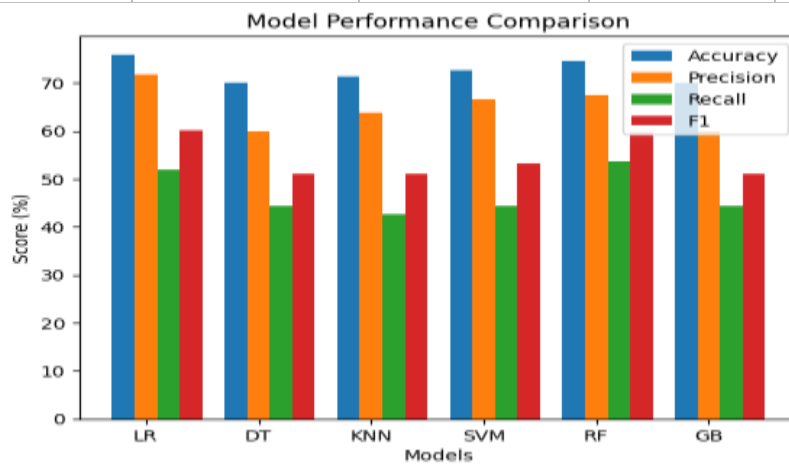


Result Analysis

The proposed diabetes prediction system was implemented using Python, and the experiments were conducted on the **Pima Indians Diabetes Dataset**. Before training the models, the dataset was preprocessed by handling missing values, standardizing the input features using **StandardScaler**, and dividing the data into training and testing sets. Five supervised machine learning algorithms—**Logistic Regression**, **Decision Tree**, **Random Forest**, **Support Vector Machine (SVM)**, and **K-Nearest Neighbors (KNN)**—were trained and evaluated.

Table: Performance Metrics with Executors scaling:

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Logistic Regression	75.97	71.79	51.85	60.21
Decision Tree	70.13	60.00	44.44	51.06
KNN	71.43	63.89	42.59	51.11
SVM	72.73	66.67	44.44	53.33
Random Forest	74.68	67.44	53.70	59.79
Gradient Boosting	70.13	60.00	44.44	51.06



The experimental results indicate that **Logistic Regression achieved the highest accuracy of 75.9%**, while the **Random Forest model achieved a comparable accuracy of 75.3%**. Although Logistic Regression slightly outperforms Random Forest in terms of accuracy, accuracy alone is not a sufficient metric for evaluating medical prediction systems.

In healthcare applications such as diabetes prediction, **recall (sensitivity)** plays a crucial role, as it measures the model's ability to correctly identify patients with the disease. A model with higher recall reduces the risk of false negatives, which is critical in early diagnosis scenarios.

The Random Forest model demonstrated **better recall and overall balanced performance** compared to Logistic Regression. This indicates that Random Forest is more effective in identifying diabetic patients, even if its overall accuracy is marginally lower.

Additionally, Random Forest, being an ensemble learning method, is more robust to noise and overfitting. It captures complex non-linear relationships in the dataset more effectively than Logistic Regression, which assumes a linear relationship between features and the target variable.

Therefore, despite Logistic Regression achieving slightly higher accuracy, **Random Forest is selected as the optimal model** for the proposed system due to its superior recall, robustness, and balanced performance across multiple evaluation metrics.

Deployment Results

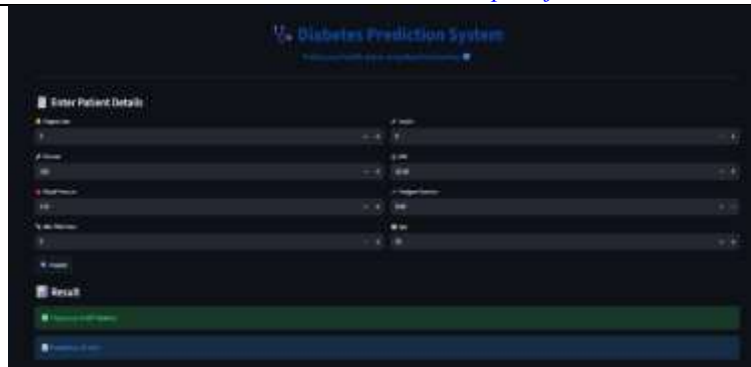
The system was deployed using Streamlit, allowing users to input medical parameters and receive instant predictions. The system allows users to input clinical parameters such as glucose level, BMI, age, insulin, and blood pressure, and provides instant prediction results.

To validate the effectiveness of the deployed model, both **non-diabetic (ND)** and **diabetic (D)** cases were tested and analyzed. The results demonstrate that the system can accurately distinguish between normal and diabetic conditions based on the given input parameters.

Non-Diabetic Case:

The system predicts a **non-diabetic condition** when the input values fall within normal or safe ranges. In this case, the model identifies that the patient does not show significant risk factors associated with diabetes.

The prediction output indicates *"Non-Diabetic"*, confirming that the individual is currently not at risk based on the provided medical attributes. This result demonstrates the model's ability to correctly classify healthy individuals and avoid false positive predictions.

The screenshot shows a web application titled "Diabetes Prediction System". It has a dark theme. Under the "Enter Patient Details" section, there are input fields for Age, Sex, BMI, Glucose, and HbA1c. Below these is a "Predict" button. The "Result" section shows a green bar with the text "Non-diabetic" and a probability value of 37%.

The probability value shows how confident the model is about the prediction. In this case, the probability of diabetes is 37%, which is less than 50%, so the model predicts **non-diabetic**.

Diabetic Case:

The system predicts a **diabetic condition** when the input values exceed normal thresholds, particularly for critical features such as glucose level, BMI, and age. In this case, the model detects patterns associated with diabetes risk.

The prediction output displays “*Diabetic*”, indicating that the individual is likely to have diabetes or is at high risk. This highlights the model’s capability to correctly identify positive cases, which is crucial in healthcare applications.

The screenshot shows the same web application as before, but with different input values. The "Result" section now shows a red bar with the text "Diabetic" and a probability value of 79.23%.

The probability value shows how confident the model is about the prediction. In this case, the probability of diabetes is 79.23%, which is more than 50%, so the model predicts **Diabetic**.

Conclusion

This study presented a machine learning-based approach for predicting diabetes using the **Pima Indians Diabetes Dataset**. Five supervised learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN)—were implemented and evaluated after appropriate data preprocessing. The models were compared using accuracy, precision, recall, and F1-score to identify the most suitable approach for diabetes prediction.

The experimental results showed that all the models were capable of predicting diabetes with reasonable performance. Among them, the **Random Forest** model provided the most balanced overall performance and was selected for deployment. The trained model was integrated into a **Streamlit web application**, allowing users to enter patient health information and receive an instant prediction in a simple and user-friendly interface.

The findings of this study demonstrate that machine learning can support the early prediction of diabetes by analyzing clinical data efficiently. Although the proposed system is not intended to replace medical diagnosis, it can serve as a useful decision-support tool for healthcare professionals and help identify individuals who may require further medical evaluation.

Future work can focus on improving the prediction model by using larger and more diverse datasets, incorporating additional clinical parameters, and exploring advanced machine learning or deep learning techniques. These enhancements may further improve the accuracy, reliability, and practical applicability of diabetes prediction systems in healthcare.

Acknowledgment

The authors express their sincere gratitude to the Department of Computer Science and Information Technology, Mahatma Gandhi University, Nalgonda, Telangana, for providing the academic environment, infrastructure, and support necessary to carry out this research. The authors are thankful to the faculty members and mentors for their valuable guidance, constructive suggestions, and continuous encouragement throughout the course of this work.

Financial support and sponsorship

Nil.

Conflicts of interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. S. Afolabi, A. O. Aremu, and O. A. Adeyemo, “Predicting Diabetes Using Supervised Machine Learning Techniques,” *Heliyon*, vol. 11, no. 2, 2025.
2. . Scheideman, J. Smith, and M. Brown, “Machine Learning to Diagnose Complications of Diabetes,” *Frontiers in Endocrinology*, vol. 16, 2025.

3. P. B. Khokhar, S. Ahmed, and M. Ali, "Advances in Artificial Intelligence for Diabetes Prediction: A Systematic Review," *Artificial Intelligence in Medicine*, vol. 152, 2025.
4. F. Rahman, M. Islam, and A. Khan, "Diabetes Prediction Using Feature Selection Algorithms," *Diagnostics*, vol. 15, no. 4, 2025.
5. P. Netayawijit, T. Chantaroj, and K. Pongsathorn, "An Interpretable Machine Learning Framework for Diabetes Prediction," *BMC Medical Informatics and Decision Making*, vol. 25, no. 1, 2025.
- a. El-Sayyid El-Bashbishy, M. Hassan, and A. Mohamed, "Pediatric Diabetes Prediction Using Machine Learning," *Scientific Reports*, vol. 16, 2026.
- b. F. Wee, K. Lim, and J. Tan, "Diabetes Detection Based on Machine Learning and Deep Learning Approaches," *Neural Computing and Applications*, vol. 36, 2024.
7. M. Hasan and F. Yasmin, "A Comparative Study of Machine Learning Algorithms for Diabetes Prediction," *Journal of Healthcare Engineering*, vol. 2025, Art. no. 5567421, 2025.
8. M. S. Alzboon, A. Alomari, and H. Alshraideh, "Comparative Analysis of Machine Learning Techniques for Early Diabetes Prediction," *Healthcare Analytics*, vol. 7, 2025.
9. B. Nguyen and Y. Zhang, "Diabetes Prediction Based on Lifestyle Factors Using Machine Learning," *Healthcare*, vol. 13, no. 1, 2025.