Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Early Detection of Diabetics using Machine Learning

[1]A. Meena Kabilan, [2]Divya Dharshini K V
[1] Professor, Dept. of AI & DS, MNM Jain Engineering College, Chennai
[2] Final Year, Dept. of AI & DS, MNM Jain Engineering College, Chennai

Abstract -Diabetes stands out as a long-term metabolic issue affecting countless folks around the world. Early spotting of it really matters for keeping preventive care on track. This work lays out a machine learning setup for predicting diabetes. It relies on just three key body measures. Those include BMI, glucose readings, and age. The goal is to sort people into diabetic or not diabetic groups. We cleaned up the data set first. Then we looked it over closely. After that, we fed it into various supervised learning methods. In the end, the chosen model hit 69 percent accuracy. This shows how even a small set of strong features can lead to useful health predictions. Overall, the tool works as a simple aid for checking diabetes risks early on.

Keywords: Diabetes Prediction, Machine Learning, BMI, Glucose, Classification Model, Healthcare Analytics

1. INTRODUCTION:

Diabetes and the Need for Early Prediction:

Diabetes mellitus counts as a big health problem on the global stage these days. It brings on lasting issues like heart problems, kidney trouble, and nerve harm over time. Catching the risk early helps cut down on how bad these get. That happens through prompt doctor help and changes in daily habits. Using machine learning for predictions gives a smart way to spot danger signs. It does this before any obvious symptoms show up.

Machine Learning for Health Diagnostics:

Machine learning algorithms pick up on patterns from past patient info. They classify or forecast health results with better and better precision. Here, we put ML methods to work on diabetes forecasting. The main traits we use are age, BMI, and glucose amount. Experts in medicine often point to these for their tight links to diabetes chances.

2. AIM OF THE STUDY

The aim here is to build a basic model that makes sense and can roll out easily. It takes in few details yet gives solid forecasts. Healthcare teams get a helper for fast choices in places short on resources.

3. METHODOLOGY:

3.1 Dataset Description

The data collection covers patient files with these traits.Table illustrates a sample subset of the dataset used for training and testing the proposed model.

Table1. Sample Datasets

ID	Pregnancies	Glucose (mg/dL)	BP	BMI (kg/m²)	Age (Years)	Outcome
S1	2	138	62	33.6	47	1
S2	0	95	64	26.6	31	0
S3	3	168	66	38.0	34	1

3.2 Data Preprocessing

We prepared the data to make training reliable. Steps covered dealing with gaps or zero entries. We also scaled the BMI and glucose numbers. Then we divided the set into train and test parts. That split went eighty to twenty.

Table 2. Pre processed Dataset Used for Model Training

Sample ID	Glucose (mg/dL)	BMI (kg/m²)	Age (Years)	Outcome
S1	138	33.6	47	1
S2	95	26.6	31	0
S3	168	38.0	34	1

Outcome:
1 – Diabetic, 0 – Non-Diabetic

3.3 Model Development

The Random Forest Classifier is an ensemble machine learning technique that combines multiple decision trees to improve prediction accuracy and stability. In this work, the Random Forest model is trained using a scrutinised dataset consisting of glucose level, body mass index (BMI), and age as input features.

During training, multiple decision trees are generated using random subsets of the dataset and features. The final prediction

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

is obtained through majority voting across all trees. This approach reduces overfitting and enhances generalization, making the model suitable for medical prediction tasks.

The trained model classifies individuals as diabetic or non-diabetic based on the learned patterns from the selected clinical attributes and achieved an accuracy of 69%.

3.4 System Workflow

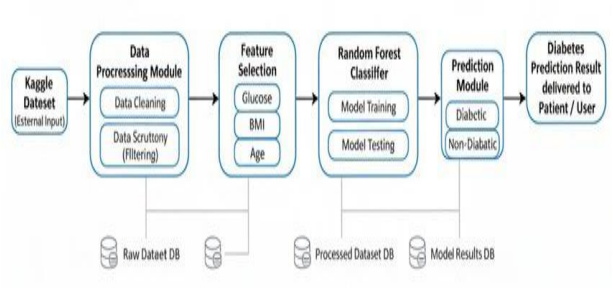

Fig.1.Early Detection of Diabetics Using Machine Learning

3.4.1 System Workflow of Proposed Diabetes Prediction Model

The system proposed here will follow a structured workflow for the early detection of diabetes using techniques from machine learning. To do this, the workflow is designed in such a manner that it provides seamless accuracy in data handling, uses the extracted features effectively, and offers reliable prediction outcomes. The overall process includes sequential modules, ranging from data acquisition to the delivery of prediction results to the end user.

I. Kaggle Dataset (External Input)

The workflow starts by acquiring data from a publicly available Kaggle dataset for diabetes. The dataset comprises anonymized patient health records and is one of the widely researched datasets within medical machine learning. The dataset acts as the external input to the system, providing the raw data that will be used in training and testing.

All patient records in the dataset are anonymized, ensuring privacy and ethical clearing. In the architecture, the raw dataset is stored initially in the Raw Dataset Database.

II. DATA PROCESSING MODULE

The Data Processing Module plays a very significant role in enhancing data quality before model training. This module consists of two major stages:

Data Cleaning

The dataset would be reviewed for any missing or zero/invalid value records in this stage. Glucose and BMI are medical attributes that may not realistically have zero values; thus, such anomalous records would be filtered out. This step will ensure that only clinically relevant data is retained for further processing.

Data Scrutiny (Filtering)

The data, after being cleaned, are then scrutinized to filter out irrelevant or noisy records. The age values are restricted to the clinically valid adult range to remove extreme outliers. Needless attributes not required in the final model are excluded at this stage. The processed and refined data is stored in the Processed Dataset Database.

Pre-processing increases model accuracy by smoothing out the noise and thereby making the entire system more reliable.

III. FEATURE SELECTION MODULE

The Feature Selection Module identifies the most relevant attributes required for diabetes prediction. By considering clinical importance and reliability in terms of data, three important features are selected:

- Glucose-a primary index of the level of sugar in the blood
- Body Mass Index (BMI) - reflects risk associated with obesity
- Age represents physiological changes associated with diabetes risk.

By limiting the model to these essential features, the system achieves better interpretability, reduced computational complexity, and suitability for real-world healthcare environments.

IV. RANDOM FOREST CLASSIFIER

The selected feature set is fed into a Random Forest Classifier serving as the core predictive component in the system.

Model Training

The classifier is then trained using the pre-processed dataset, in which multiple decision trees will be constructed using random subsets of data and features. This allows the model to learn complex patterns and relationships between the selected clinical attributes and diabetes outcomes in this ensemble.

Model Testing

In the testing phase, after the training is done, unseen data is used in order to assess the predictive performance of the model. The results of both the training and testing phases are stored in the Model Results Database for analysis and validation.

Random Forest was chosen for their robustness, resistance to overfitting, and capturing nonlinearities in medical data quite well.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

V. PREDICTION MODULE

The Prediction Module then uses the trained Random Forest model to classify new incoming patient data. The system, based on learned patterns, will predict one of the following outcomes:

- **Diabetic**
- **Non-Diabetic**

It is the class or classification that the system makes on an instance with regard to whether or not diabetes exists in the patient.

VI. DIABETES PREDICTION RESULT DELIVERY

In the final stage of the workflow, the prediction result will be delivered to the Patient/User. The output gives the indication related to the risk of diabetes in an interpretable way to let the person know for early awareness, in case timely medical consultations are required. The system will be able to provide support for early screening and decision-making rather than substituting for clinical diagnosis, thus working as an assistive tool for healthcare applications.

VII. DASHBOARD INTERFACE AND USER INTERACTION

The proposed system incorporates a lightweight dashboard to facilitate user interaction and real-time diabetes prediction. The dashboard allows users to manually input clinically relevant parameters, namely glucose level, body mass index (BMI), and age.

These inputs are preprocessed using the same scaling mechanism applied during model training to ensure consistency. The trained Random Forest classifier then processes the input data and generates a probability-based prediction indicating diabetes risk.

The dashboard presents the final classification outcome, along with a risk probability score, enabling clear interpretation of results. This interface enhances the practical usability of the system by bridging the gap between the predictive model and end users.

4. RESULTS AND DISCUSSION:

The performance of the proposed Random Forest classifier was evaluated using a test dataset comprising 150 samples. The model achieved an overall accuracy of 69.33%, demonstrating effective classification capability using glucose level, BMI, and age as input features.

The confusion matrix analysis indicates that the model correctly classified 77 non-diabetic and 27 diabetic cases. While some misclassifications were observed, the model maintained a balanced performance suitable for early-stage diabetes screening.

Further evaluation using precision, recall, and F1-score revealed stronger performance in identifying non-diabetic cases, with a precision of 0.79 and recall of 0.75. The diabetic class achieved a recall of 0.56, indicating the model's ability to identify a significant portion of diabetic cases despite the limited feature set.

Overall, the results validate the feasibility of the proposed approach as an assistive tool for preliminary diabetes risk assessment rather than a definitive diagnostic system.

5. APPLICATIONS OF PREDICTIVE ANALYTICS IN HEALTHCARE:

Predictive analytics boosts healthcare in several ways.

- It spots high-risk people soon.
- That lightens the load on clinics.
- It aids custom treatment paths.
- It cuts down on return hospital stays.
- It betters watch and follow-up routines.

Our diabetes setup slots into the big picture of AI in health. That world focuses on stopping problems early. It leans on data for smart calls.

6. TECHNOLOGIES AND TOOLS:

- **Python** was used as the primary programming language for implementing the overall system due to its simplicity and strong support for machine learning applications.
- **Scikit-learn** was utilized to develop and train the Random Forest classification model for diabetes prediction.
- **Pandas** and **NumPy** were employed for data preprocessing, cleaning, and efficient numerical computation.
- **Matplotlib** was used to visualize data patterns and model-related results during analysis.
- **Streamlit** was used to build an interactive web-based dashboard that enables users to manually input clinical parameters and obtain real-time prediction results.

7. FUTURE SCOPE

We could grow the model with extras like these.

- Blood pressure readings.
- Insulin amounts.
- Family background info.
- Activity tracking numbers.

Later work might try deep learning setups. It could add ongoing checks via IoT gear. Mobile apps for on-the-go risk checks seem promising too.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

8. Conclusion

This paper developed a machine learning–based system for early diabetes prediction using a Random Forest classifier and key clinical features such as glucose level, BMI, and age. The model achieved an accuracy of approximately 69% and demonstrated reasonable performance in distinguishing diabetic and non-diabetic cases. A lightweight dashboard was implemented to enable real-time user input and prediction. Overall, the system serves as an assistive tool for preliminary diabetes risk assessment and can be further improved by incorporating additional features and advanced optimization techniques.

9.REFERENCES

[1] World Health Organization, *Global Report on Diabetes*, WHO Press, Geneva, Switzerland, 2016.
[2] Kaggle, "Pima Indians Diabetes Database," Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
[3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[4] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[5] S. S. Kumar and M. V. Reddy, "Diabetes Prediction Using Machine Learning Algorithms," *International Journal of Computer Applications*, vol. 174, no. 8, pp. 12–18, 2017.
[6] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.