Advanced International Journal for Research (AIJFR)
E-ISSN: 3048-7641 ● Website: www.aijfr.com ● Email: editor@aijfr.com

Prediction of Diabetes Using Machine Learning Algorithms

Premashree R[1], Dr. Manjunath Kumar B H[2]

Department of Computer Science and Engineering, S.J.C. Institute of Technology.
premagowdacse@gmail.com, manjunathabh@sjcit.ac.in

Abstract

Diabetes mellitus is one of the most prevalent chronic diseases worldwide, posing serious health risks and increasing the burden on healthcare system. Early and accurate prediction of diabetes can significantly improve patient outcomes by enabling timely medical interventions and lifestyle modifications. In this study, we developed a machine learning–based system to predict the likelihood of diabetes using clinical data. The dataset employed consisted of diagnostic attributes such as glucose level, blood pressure, insulin, body mass index (BMI), age, and other relevant parameters. To achieve reliable prediction, multiple supervised learning algorithms including Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost and Light GBM were implemented and evaluated. The models were trained and tested using the Pima Indians Diabetes Database, a widely used benchmark dataset in healthcare analytics. Performance metrics such as accuracy, precision, recall, and F1-score were used to compare the models. Among the tested algorithms, ensemble methods such as Random Forest demonstrated comparatively higher predictive performance, highlighting their ability to handle nonlinear relationships and feature interactions. This study emphasizes the importance of preprocessing techniques, including handling missing values, scaling numerical attributes, and balancing class distributions, which are critical for improving the model robustness. The results suggest that machine learning algorithms can serve as effective decision-support tools for medical practitioners when optimally optimized. This study contributes to the growing field of predictive healthcare analytics by demonstrating the potential of data-driven approaches to assist in the early diabetes diagnosis and risk assessment.

Keywords: Computational Diagnostics, Predictive Healthcare Analytics, Ensemble Learning Models, LightGBM

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from defects in insulin secretion, insulin action, or both. It is one of the most pressing global health challenges, affecting millions of people worldwide and contributing to severe complications such as cardiovascular disease, kidney failure, nerve damage, and vision loss. According to the World Health Organization, the prevalence of diabetes has been increasing steadily, making its early detection and effective management a critical necessity. Traditional diagnostic methods, while accurate, are often invasive, time-consuming, and dependent on clinical expertise, which limits their applicability in large-

scale screening. In recent years, computational intelligence and data-driven approaches have emerged as powerful alternatives for disease prediction and risk assessment.

Machine learning (ML), a subset of artificial intelligence, has shown remarkable potential in healthcare analytics due to its ability to learn complex patterns from data and generate predictive insights. By leveraging patient medical records and diagnostic attributes, ML algorithms can assist in the early identification of individuals at risk of developing diabetes. The integration of such approaches into healthcare systems not only supports clinicians in decision-making but also reduces costs and improves preventive care strategies. The present study focuses on applying supervised machine learning algorithms to predict the likelihood of diabetes using structured clinical data.

The dataset employed in this project is the widely used Pima Indians Diabetes Database, which contains diagnostic features including glucose level, blood pressure, insulin, body mass index (BMI), skin thickness, number of pregnancies, diabetes pedigree function, and age. These parameters are known to have strong correlations with the onset of diabetes, making them suitable predictors for machine learning–based classification models. Various algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and K-Nearest Neighbor (KNN) were implemented and evaluated in this work. To enhance the reliability of predictions, preprocessing steps like data normalization, handling missing values, and balancing class distribution were performed. The objective of this project is to compare the performance of different machine learning algorithms in terms of accuracy, precision, recall, and F1- score, thereby identifying the most suitable approach for diabetes prediction. By analyzing algorithmic performance on real-world medical data, this research highlights the applicability of machine learning as a diagnostic support tool. Ultimately, the study demonstrates how computational methods can aid in early diabetes detection, improve healthcare efficiency, and pave the way for personalized medical interventions.

2. Related Works

The assistance of Machine Learning in bio-informatics is not a new idea, over time, the interest in these domains together has increased gradually, increasing the number of people researching on the same.
Anisa and Kurniawan design a Flask web app with a Decision Tree trained on symptom features; cross-validation scores were 0.75 and 0.50, mean 0.62. Insight: lightweight stack with SQLite logging makes prototyping simple and transparent. Advantages: interpretability, clean integration of model saving and routing, and low barrier to deployment. Drawbacks: two-fold CV is weak, the symptom-only dataset is small and self-reported, and performance is modest, so clinical utility is limited without richer features and stronger validation.[1]

Raju et al. build a Flask app around multiple ML models and an ensemble, positioning the web layer as the bridge from training to

bedside. Useful takeaways: end-to-end pipeline description, deployment choices, and emphasis on usability. Strengths include open access and a concrete web implementation. Weaknesses: unclear external validation, limited discussion of class imbalance, privacy and security left thin, and dataset provenance not explicit, which limits generalizability and clinical readiness.[2]

Ahmed and Gupta present a Flask web application that wraps several classifiers KNN, LR, DT, SVM, RF trained on the Kaggle diabetes dataset. Insight: a modular Flask structure routes inputs to trained models for real-time risk scoring and clear separation of templates, forms, and inference. Pros: comparative modeling, practical web integration, and simple UI for rapid feedback. Cons: reliance on the Pima-style dataset, no external cohort or calibration checks, and limited treatment of fairness, privacy, and model monitoring in production. [3].

Sai Priya R and Sarin Priya D implement a Flask app using models such as Random Forest on Pima features; reported test accuracy ranges roughly 77–82 percent. Notable insight: explicit attention to security, logging, and an API route design home and predict that many researchers can reuse. Advantages: pragmatic architecture, clear preprocessing steps, and actionable blueprint for student or lab deployments. Drawbacks: binary outputs without calibrated probabilities, no AUC or external validation, and the authors position it as educational, so clinical translation remains a future step.[4]

Fidelis Obukohwo Aghware study shows XGBoost, when paired with balancing methods (SMOTE-Tomek/SMOTEEN), reached roughly 81–82% accuracy on the PIMA dataset, demonstrating strong sensitivity gains versus unbalanced training. Advantages: clear improvement from data balancing, robust tree-based handling of nonlinear interactions, and a practical Flask API deployment for real-time use. Insights: performance hinges more on preprocessing than model choice; ensemble/tree learners respond well to SMOTE variants. Drawbacks: small, imbalanced public data limit generalizability, feature selection was not applied, and high accuracy risks overfitting without external validation. Overall, useful for applied prototypes but needs larger, heterogeneous cohorts for clinical claims.[5]

3. Background

Supervised Machine Learning (SuML) entails developing algorithms capable of recognizing typical patterns and conditions by using given examples to predict results. In contrast to UnSuML techniques, SuML always uses labelled data inputs. Classification is an aspect of the SuML method, concentrating on organizing data based on existing information, which is split into testing and training datasets as required. This method is employed to categorize data elements into pertinent groups that exhibit similar characteristics. Within this framework, two classification algorithms are utilized to assess whether a person has diabetes. The algorithms are:

3.1 GRADIENT BOOSTING CLASSIFIER

XGBoost is a robust gradient boosting algorithm that enhances predictions by integrating multiple weak learners, typically decision trees, into a formidable model. It effectively handles missing values, manages imbalanced datasets, and mitigates overfitting through regularization. In the context of diabetes prediction, XGBoost adeptly captures complex patterns among medical attributes such as glucose levels and BMI, offering greater accuracy and stability compared to traditional classifiers, making it highly suitable for clinical decision support systems.

3.2 LIGHT GRADIENT BOOSTING MACHINE

LightGBM is a gradient boosting framework optimized for speed and efficiency, making it particularly suitable for processing extensive medical datasets. It utilizes a leaf-wise growth approach, which allows for deeper splits and improves the accuracy of capturing non-linear relationships among patient characteristics like glucose, insulin, and BMI. Its ability to handle categorical values, reduce memory consumption, and prevent overfitting through built-in regularization enhances prediction quality. In the realm of diabetes detection, LightGBM provides quicker training and reliable results for clinical use.

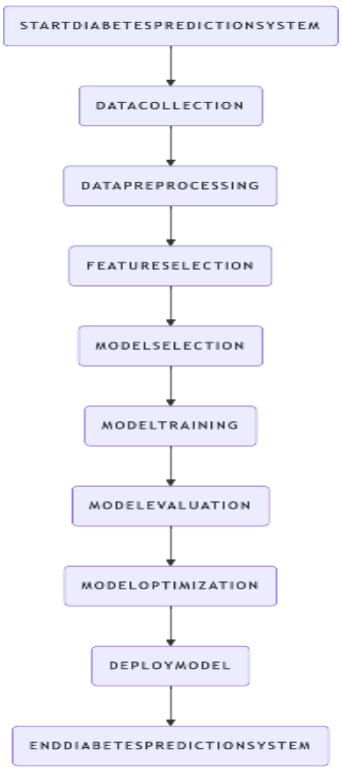

Fig.1. System Architecture

4. Result and Discussion

This paper gives the outcome of Diabetes prediction in order to get the training set and testing set from dataset for the preparation and testing of the models which will be used for the prediction test of diabetes.

The dataset which is used for the testing of the model is taken from the online source (www.kaggle.com), and the data consist of 10,000 cases and each case has 8 attributes for each individual, the attributes are:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- BMI (Body Mass Index)

- Diabetes Pedigree Function
- Age

Accuracy: It is the basic criteria for the evaluation of any model or algorithm. It is defined as the number of accurate responses given by the model concerning the total number of data cases given for the prediction.

Accuracy = Number of correct predictions/Total number of Predictions (1)

Table.1. Accuracy Score of the models

Classification Algorithm	Accuracy
XGBoost Classifier	0.958
LightGBM	0.964

From Table.1 we can deduce that the accuracy value of XGBoost Classifier and LightGBM is almost same. From the Table.1, we cannot conclude which model is better as the value of each model is almost equal. So, there is a need to evaluate other factors too.

Table.2. Average Precision, Precision and Recall of the models

Classification Algorithm	Average Precision	Precision	Recall
XGBoost classifier	0.9652	0.9652	0.9652
LightGBM	0.9664	0.9664	0.9664

In the Table.2, factors corresponding to models are:

Precision: It checks all the True positive values with respect to the total positive cases. Where total positive cases includes both true positive and false positive.

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

Recall: It checks the True Positive cases with total case predicted by the model of the classifier.

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

Average Precision: It is the area under the curves of Precision and Recall. The value of average precision is always between 0 and 1 and the model which has the higher average precision is better model for prediction.

$$AP = \int_0^1 r dr \quad (4)$$

From the Fig 2 we can say that XGBoost classifier and LightGBM has no much variations in accuracy of predicting the model.

So let us deliberate upon more factors that can be considered for the selection of the classifier. The factors which are shown in the Table.3, are F1-measure, log loss, ROC AUC, build time (s).

Table.3. Representation of F1, ROC AUC of the model

Classification Algorithm	F1	ROC AUC
XGBoost Classifier	0.9652	0.994
LightGBM	0.9664	0.9941

$$F1 = 2 * \frac{\text{Precision} \times \text{Recal}}{\text{Precision} + \text{Recal}} \quad (5)$$

Log Loss: It is the most essential factor for classification based on the probabilities.

ROC AUC: It is a curve which shows the performance of the classification problems at different threshold settings. ROC AUC are two distinguished parts, ROC is the probability curve and AUC is used to measure the separability or it represent the degree of separability.

From all the Table.3, it is clear that LightGBM is better than XGBoost classifier. As it clearly evident in Fig.2, the graph of LightGBM is peak compared to that of XGBoost classifier.

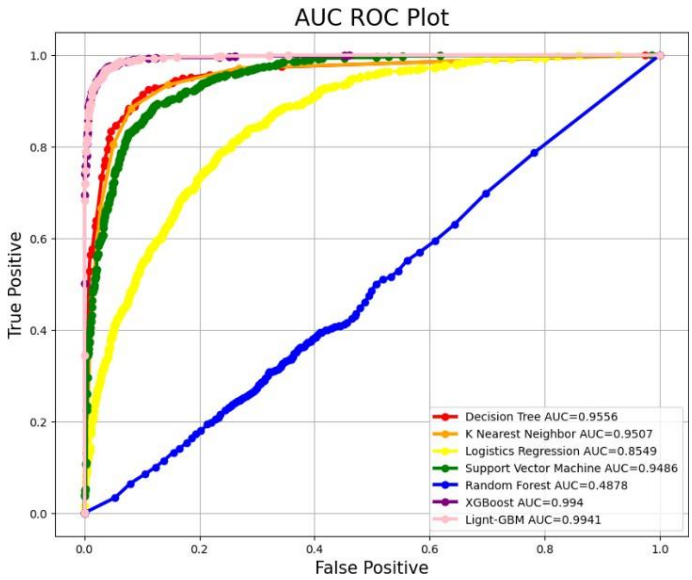

Fig.2. AUC ROC Plot

5. Conclusion

This project demonstrates how machine learning models can effectively predict diabetes using patient health data. By integrating the trained model with a Flask web application, the system provides quick and accessible predictions for users. The approach highlights the potential of algorithms in supporting early detection and preventive care. While results are promising, future work should involve larger and more diverse datasets, enhanced feature engineering, and external validation to improve reliability for real-world healthcare deployment.

References

1. "Diabetes Prediction Using Flask and Decision Tree Classifier with Cross Validation." By the author Nor Anisa, Anggara Kurniawan L. 2024.
2. "Diabetes Prediction Using Machine Learning and Flask." by the author N Kushal Kumar Raju, Keshav Krishnamurthy, Bhuvanagiri Prahal Bhagavath, Nathan Shankar, M. Janani, N Avinash.
3. "A Web Application for Predicting Diabetes Using Machine Learning Methods." By the author Riyaz Ahmed.
4. "Diabetics Detection Using Python." By the author Sai Priya.R, Sarin Priya.D
5. "Effects of Data Balancing in Diabetes Mellitus Detection: A Comparative XGBoost and Random Forest Learning Approach" by the author Fidelis Obukohwo Aghware, Maureen Ifeanyi Akazue.
6. "Risk Identification and Mitigation for Diabetes Prediction and Control Systems in Healthcare Software." By the author Sana Rizwan, Hassan Jamal, Asadullah, Rehan Rabbani Baig.
7. "Diabetes Prediction." By the author Sana Rizwan, Hassan Jamal, Asadullah, Rehan Rabbani Baig. 2025
8. "Flask app for diabetic prediction by improvising SVM." By the author Sankalp Dwivedi.
9. "Diabetes prediction and visualization platform based on machine learning." By the author Yvjie Li, Chenzhong Yang, Junyi Yang, Chongyv Zhang, Wang Gao.
10. "Web Application for Diabetes Prediction using Machine Learning Techniques." By the authors